

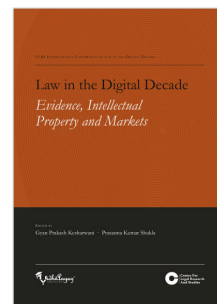
CHAPTER 10

A Comparative Legal Analysis of the Emerging Right to Know Concerning the Use of Copyrighted Works in AI Training Data

C. Sophia Jeyakar

in *Law in the Digital Decade: Evidence, Intellectual Property and Markets*, edited by Gyan Prakash Kesharwani and Prasanna Kumar Shukla (VidhiAagaz 2026), pp. 93–107

DOI <https://doi.org/10.63108/VAB.LDD.2.10>



CITATIONS

BLUEBOOK (21ST ED.)

C. Sophia Jeyakar, *A Comparative Legal Analysis of the Emerging Right to Know Concerning the Use of Copyrighted Works in AI Training Data*, in *LAW IN THE DIGITAL DECADE: EVIDENCE, INTELLECTUAL PROPERTY AND MARKETS* 93 (Gyan Prakash Kesharwani & Prasanna Kumar Shukla eds., VidhiAagaz 2026).

OSCOLA (4TH ED.)

C. Sophia Jeyakar, ‘A Comparative Legal Analysis of the Emerging Right to Know Concerning the Use of Copyrighted Works in AI Training Data’ in Gyan Prakash Kesharwani and Prasanna Kumar Shukla (eds), *Law in the Digital Decade: Evidence, Intellectual Property and Markets* (VidhiAagaz 2026) 93.

CHICAGO (17TH ED.)

Jeyakar, C. Sophia. “A Comparative Legal Analysis of the Emerging Right to Know Concerning the Use of Copyrighted Works in AI Training Data.” In *Law in the Digital Decade: Evidence, Intellectual Property and Markets*, edited by Gyan Prakash Kesharwani and Prasanna Kumar Shukla, 93–107. Maihar: VidhiAagaz, 2026. <https://doi.org/10.63108/VAB.LDD.2.10>.

APA (7TH ED.)

Jeyakar, C. S. (2026). A comparative legal analysis of the emerging right to know concerning the use of copyrighted works in AI training data. In G. P. Kesharwani & P. K. Shukla (Eds.), *Law in the digital decade: Evidence, intellectual property and markets* (pp. 93–107). VidhiAagaz. <https://doi.org/10.63108/VAB.LDD.2.10>

SOURCE

<https://doi.org/10.63108/VAB.LDD.2.10>

LICENCE

Open access under the Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0): it may be copied, shared and adapted for non-commercial purposes, with credit to the authors and the source.

Citations are given as a guide; check them against the style manual you follow. This page was added by VidhiAagaz to the downloaded file. The chapter follows, with the page numbers of the printed book.

CHAPTER 10

A Comparative Legal Analysis of the Emerging Right to Know Concerning the Use of Copyrighted Works in AI Training Data

C. SOPHIA JEYAKAR*

The rapid development of generative artificial intelligence (AI) has intensified concerns regarding the use of copyrighted works in training datasets, particularly where authors and right holders have limited or no knowledge of whether their works have been accessed, copied, processed, or incorporated into AI training systems. This raises an emerging legal question: whether authors and copyright holders should possess a legally enforceable right to know whether, and to what extent, their copyrighted works have been used as AI training data, and whether existing copyright law provides adequate mechanisms to secure such transparency. The research examines the tension between the commercial and technological interests of AI developers and the informational interests of copyright holders, with particular focus on dataset transparency, disclosure of training-data sources, and the evidentiary difficulties involved in establishing the use of particular copyrighted works.

The study adopts a comparative doctrinal legal methodology, analysing statutory provisions, judicial decisions, regulatory developments, policy documents, and scholarly literature concerning copyright and AI training data. It comparatively examines the emerging approaches in India, the European Union, the United States, and selected other jurisdictions, with particular attention to transparency obligations and copyright-based claims concerning AI training. The paper argues that although a general statutory “right to know” concerning AI training datasets is not yet uniformly recognised, existing copyright principles, transparency obligations, and emerging regulatory approaches provide a foundation for developing such a right. It proposes that meaningful dataset transparency should form an essential component of the contemporary copyright framework by enabling authors to identify potential unauthorised uses, exercise available rights, seek appropriate remedies, and participate more effectively in licensing and remuneration mechanisms. The study therefore advocates a balanced transparency framework that protects copyright interests without imposing disproportionate disclosure obligations that could undermine legitimate AI innovation, trade secrets, or technological development.

Keywords: Artificial Intelligence; Copyright Law; Right to Know; AI Training Data; Dataset Transparency

1 INTRODUCTION

1.1 Background: Generative AI and Copyrighted Training Data

The quick advancement of generative artificial intelligence (AI) has changed the way in which information, creative works and other kinds of digital content are gathered, processed and reproduced. Today’s generative AI systems, especially large foundation models, are created through training procedures that make use of large datasets consisting of text, images, audio, video, software code and other types of information. The size of these datasets has fundamentally affected the connection between copyright law and technological progress. Unlike traditional forms of digital use, the development of AI can involve

* Assistant Professor & Research Scholar at Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India.

the systematic handling of huge amounts of works, many of which come from publicly available websites, digital libraries, repositories, online platforms and other web sources. AI training datasets can include works that are the subject of copyright, such as books and academic publications, newspaper articles and material from the web, photographs and artistic works, musical and audiovisual material, computer programs and databases. Yet the fact that a copyrighted work is included in a dataset does not on its own mean that copyright infringement has taken place. The issue of legal importance is what actions are carried out in obtaining, processing and using the work and whether or not those actions fall within the exclusive rights of the copyright owner or an applicable statutory exception or limitation.¹

It is especially important to make this distinction since the technological process of training an AI involves a number of possibly different stages. Getting access to a work simply means obtaining or looking at the information it contains, while downloading or copying might involve reproducing the work or a substantial part of it. The act of processing or extracting data from a work may include computational analysis, transformation or extraction of the information, depending on the type of work and the technology used. Training an AI model then constitutes an additional stage in which the information that has been extracted or processed is used to establish statistical or computational relationships inside the model. Lastly, output generation takes place when the trained model produces new content in response to a user's prompt.² The various stages should not be regarded as legally equivalent just because they are part of the same technological process. The fact is important since copyright law usually controls specific acts in connection with protected works rather than merely the presence of information in a technological system. When it comes to the legal interpretation of AI training, it is necessary to consider whether a protected work has been reproduced, communicated, adapted, extracted or in some other way used in a manner which falls within the copyright owner's exclusive rights. At the same time, the analysis has to take into account statutory exceptions, limitations, permitted uses and the new approaches to computational analysis and text and data mining.

Since copyrighted material is being used more and more in the development of AI, a complicated balance arises between technological innovation, the need for copyright protection and the legal right to access data.³ The purpose of copyright is to give authors real control and economic advantages with respect to their creative works, while the development of AI may require access to large and varied amounts of information. It is possible that overly strict measures could hinder technological research and innovation, and on the other hand, there is a risk that the commercial use of copyrighted works without restriction could damage the economic and informational interests of authors and rights holders. In this context, the key question is not merely whether copyrighted works are included in the AI training datasets, but whether authors and rights holders should be able to find out that their works have been used, identify the way in which they have been used and be able to exercise meaningful legal rights in relation to such use. This distinction forms the basis for looking at the new right to know about copyrighted works that have been incorporated into the AI training process and the various approaches taken by India, the European Union and other regions.

¹ U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (2025).

² Regulation 2024/1689, of the European Parliament and of the Council of June 13, 2024, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689) 1, recital 107 (EU).

³ Mark A. Lemley & Bryan Casey, Fair Learning, 99 Tex. L. Rev. 743, 743–44 (2021).

1.2 The Problem of Information Asymmetry and of the Evidence

The incorporation of copyrighted materials into the training of AI systems presents a clear issue in the form of information asymmetry between copyright holders and the developers of the AI. While copyright owners generally have knowledge of the works themselves, they often have very little or no reliable information concerning what happens to those works after they have been made available online or otherwise come within the reach of data-collecting systems. It is therefore possible that an author will know that a certain book, article, photograph, work of art, musical piece or software program exists but will not know whether it has been collected, included in a training dataset, processed as part of the AI development process, or finally used in training a specific model.

This lack of information can occur at a number of different stages. A copyright holder might not know whether a work has simply been accessed, whether a copy of it was made during the process of data collection, which dataset or data repository the work was in, whether the work survived the various filtering and preprocessing steps, or whether it was in fact used in training. Even if a copyright owner suspects that a work has been used, it may be hard to identify the exact AI model that was trained on it. The difficulty is increased by the huge size and constantly changing nature of current training datasets, since these may include millions or billions of individual data items gathered from a number of different sources and then subjected to a series of collection, filtering, deduplication, classification and transformation processes. On the other hand, AI developers and the organisations taking part in model development may have considerably more information about the origins and the way in which the training data was handled. According to their internal practices, they might keep dataset documentation, records of data acquisition, logs relating to filtering and preprocessing, provenance information, licensing records and other technical documentation connected with model development. This kind of information could allow a developer to find out if a particular work got into a dataset and how it was then treated. Nevertheless, the existence, availability and evidentiary value of such records can differ greatly from one developer to another.⁴

The asymmetry produced presents a major problem when it comes to evidence. A copyright holder who wants to enforce their rights cannot readily prove a fact which is probably in the knowledge or control of the AI developer, namely, whether their work was actually gathered and used during the training process. Simply because an AI-generated work is similar to an existing copyrighted work does not mean that the work was included in the training data. Likewise, the fact that a work is available to the public does not necessarily show that it was downloaded, copied, processed or incorporated into a specific training dataset. In order to verify this independently, access to technical details might be needed, details which the copyright holder does not have. The implications of this problem go beyond legal proceedings. Since they do not have reliable information about how their works have been used, copyright owners may find it difficult to decide whether or not to take action, to negotiate licences, to seek payment or to challenge certain types of AI development. Information about actual use could also be important for collective licensing and for enabling authors and rights holders to make well-informed commercial decisions about their works. Therefore, the lack of transparency can have an effect not only on the enforcement of current rights but also on the ability to exercise those rights in a meaningful way.

⁴ Regulation 2024/1689, recital 107, 2024 O.J. (L 1689) 1 (EU).

The way in which evidence is handled therefore changes the current discussion about copyright and AI training. It is not just a matter of determining whether the use of copyrighted works for training amounts to infringement under the relevant law; an earlier and possibly decisive issue is how a copyright owner can prove that such use has taken place. Since important information regarding data acquisition, provenance and model development is held by the AI developers, the question of who should bear the burden of providing this information becomes a significant legal issue.

The debate now developing around a right to know should not be seen simply as a call for general corporate transparency, but rather as a possible means of dealing with the particular structural imbalance between the people who create copyrighted works and those who build AI systems by using large-scale datasets. The main argument of this paper is that the copyright issue is not just a matter of whether training AI constitutes infringement; it also involves determining who is responsible for proving that copyrighted works were indeed used. This difficulty in providing evidence forms the basis for investigating whether the current copyright systems offer adequate protection for authors and rights holders and whether other legal systems are tending towards imposing specific transparency or disclosure requirements regarding the training data used by AI systems.⁵

1.3 The “Right to Know” Concept

The growing idea of a “right to know” is the possible legal claim of a copyright holder to get enough information in order to assess whether or not their copyrighted work has been used in the training of AI systems, and where appropriate, how it has been used. This notion has become more important since copyright owners often have no knowledge of whether their works have been gathered, included in training datasets, or employed in the development of specific AI models. In the absence of such information, it becomes difficult to put their copyright rights into practice.

It is necessary to differentiate the right to know from other similar concepts. The right to access information relates to the possibility of getting hold of the relevant records or data, while the right to know refers to the fundamental interest in ascertaining whether a work was used. Transparency, on the other hand, broadly speaking, involves obligations on AI developers to make available information about their systems, datasets or training methods. Yet general transparency does not allow an individual copyright holder to check that a particular work was used. Attribution means giving recognition or credit to the creator, and remuneration refers to the payment of financial compensation in respect of the use of a copyrighted work.⁶ Thus, the right to know should not automatically be understood as a right to payment, attribution, consent, or a finding of infringement. Rather, it represents an emerging informational entitlement concerning the use of copyrighted works in AI training, which may enable copyright owners to assess and exercise their existing legal interests.

1.4 Research Questions

- a) Whether existing copyright laws provide copyright holders with an enforceable right to know whether their works have been used in AI training?

⁵ Regulation 2024/1689, art. 53(1)(d), 2024 O.J. (L 1689) 1 (EU).

⁶ Regulation 2024/1689, art. 53(1)(d), 2024 O.J. (L 1689) 1 (EU).

- b) Whether emerging transparency obligations in different jurisdictions effectively enable copyright holders to identify such use?
- c) Whether disclosure requirements should extend to individual copyrighted works or remain limited to categories/sources of training data?
- d) What legal framework can balance copyright holders' informational interests with AI innovation, trade secrets and technological development?

1.5 Scope and Methodology

The study adopts a doctrinal legal research methodology supported by a comparative approach. It examines primary sources including statutes, regulations, judicial decisions and official regulatory materials, along with secondary sources such as academic scholarship and policy literature. The principal jurisdictions selected for comparison are India, the European Union and the United States, with additional jurisdictions considered only where they provide significant legal developments relevant to the emerging right to know in AI training data.

2 CONCEPTUAL AND LEGAL FOUNDATIONS OF THE RIGHT TO KNOW

2.1 Meaning and Scope of the Right to Know

The right to know means that a copyright owner is entitled to get enough information in order to assess whether and in what way their copyrighted work has been used in the training of AI systems. In order for this right to be of any value, the information provided should specify whether the work was used, where it was obtained from, the nature and extent of its use, and, where appropriate, the AI system or the developer responsible. This kind of information allows copyright owners to understand how their works have been treated and to assess their legal rights. At the same time, however, the extent of the information that has to be disclosed should be proportionate and must not unduly interfere with legitimate AI development, trade secrets or confidential information.⁷

2.2 The Copyright Owner's Informational Interests

Copyright holders have exclusive rights which can be influenced by the way their works are used in the training of AI systems, such as the right to reproduce, communicate or distribute where relevant, the right to adapt and the right to license. Knowledge of the data being used allows copyright owners to decide whether or not these rights have been exercised and whether or not further legal steps are needed. Such knowledge can help copyright owners in negotiations over licensing, in keeping an eye on unauthorised use, in claiming remuneration where this is legally possible, and in taking enforcement action or pursuing appropriate remedies. The right to know therefore serves to enable the practical exercise of the copyright rights which already exist rather than in itself creating a new substantive right; its importance consists in reducing the information gap between copyright owners and AI developers.⁸

⁷ Regulation 2024/1689, recital 107, 2024 O.J. (L 1689) 1 (EU).

⁸ Copyright Act, 1957, § 14 (India).

2.3 The Problem of Hidden Use and Copyright Infringement

Normal cases of copyright infringement usually revolve around a specific act involving a work that is the subject of copyright protection, allowing the copyright holder to identify and oppose the supposed infringing use. With AI training, however, this situation becomes more complicated since copyrighted works can be gathered, copied, processed and included in large datasets without any obvious interaction taking place with the copyright holder. This results in an “invisible use” problem, meaning that the owner may find it difficult to tell whether a given work was part of the training data or later used in the development of the AI model. Even if infringement has taken place, the lack of available information about the data sources and the processing makes it much more difficult for copyright owners to detect the infringement, to gather evidence and to take enforcement action.

2.4 The Right to Know as Compared with the Right to a Licence or Payment

The right to know must be separated from the right to obtain a licence or to be paid. The fact that information about the use of a copyrighted work is available does not mean that the copyright owner is entitled to payment or that the AI developer was legally obliged to get prior permission. On the other hand, transparency can act as a necessary condition for exercising the copyright and commercial rights which already exist. Knowing that a work has been used may allow the owner to decide whether or not licensing negotiations, claims for payment or actions for infringement are suitable in accordance with the relevant law. Hence, the right to know is mainly of an informational nature, whereas licensing and remuneration rely on the specific legal framework relating to the way in which the work has been used.

2.5 The Right to Know Takes the Form of Dataset Transparency

Dataset transparency offers the practical means by which the right to know can be put into practice in the context of AI. Since a copyright owner cannot properly ascertain whether their work has been used without having access to adequate information about the sources, categories or origin of the training data, transparency requirements may therefore necessitate that AI developers make available appropriate details regarding the datasets, while taking into account confidentiality, security and legitimate commercial interests. The main argument of this paper is that the right to know only becomes practically meaningful in so far as legal mechanisms require a certain level of transparency regarding the training data used in AI. This kind of transparency does not have to involve the disclosure of each individual work but should instead provide sufficient information to enable the meaningful verification and exercise of copyright interests.

3 COPYRIGHT, AI TRAINING AND THE EMERGING TRANSPARENCY FRAMEWORK: A COMPARATIVE ANALYSIS

3.1 Copyright Protection and AI Training Data

The process of training AI usually involves gathering huge amounts of digital content by means such as web scraping, downloading and extracting data. In cases where the material obtained consists of works protected by copyright, it becomes necessary to consider whether the acts of copying, storing, processing and including those works in the training datasets constitute a breach of the copyright holders’ exclusive

rights. When carrying out a legal assessment it is essential to make a distinction between the first collection of a work, its reproduction or processing while preparing the dataset, and its later use in training an AI model. The mere fact that copyrighted material is present in a dataset does not mean that infringement has taken place; instead, the legality of each particular action must be determined in accordance with the relevant copyright rules and any statutory exceptions. This provides the solid basis for looking at transparency: before a copyright owner can judge whether their rights have been affected, they might need to be given information regarding the collection and handling of their work.

3.2 European Union

The European Union has one of the most advanced sets of regulations regarding copyright and transparency in relation to data used for training AI. The Directive on Copyright in the Digital Single Market includes specific rules about text and data mining (TDM). According to Article 3, eligible research organisations and cultural heritage institutions may carry out reproductions and extracts for scientific research purposes, provided that certain conditions are met. Article 4 introduces a wider exception for TDM, while at the same time enabling copyright holders to retain their rights in respect of works that have been made available online using suitable methods.

The EU framework is especially important in the light of the EU AI Act, which imposes on providers of general-purpose AI models duties relating to transparency. Such providers must keep and make available information about the content which has been used to train their models, in line with the applicable regulatory framework. The aim of these obligations is to enhance transparency regarding the training content while at the same time taking into account both copyright protection and the practical circumstances of large-scale AI development. The main question that this paper addresses is whether these obligations constitute an effective “right to know” for individual copyright holders. A general disclosure of the training content may contribute to greater transparency without necessarily enabling an individual author to determine whether a particular work was used. The EU framework thus offers a key opportunity to assess whether regulatory transparency can be turned into a real informational right.

3.3 United States

In the United States, the way in which copyright has addressed its relationship with AI training has mostly been carried out by means of copyright doctrine, fair-use analysis and litigation, not through a general legal right that would compel AI developers to make disclosure of their training data to copyright holders. When applying the fair-use framework, account must be taken of various factors such as the purpose and character of the use, the nature of the copyrighted work, the amount used and the effect on the potential market.⁹

The recent legal actions relating to AI training have led to doubts about whether copying copyrighted works for the purpose of machine learning can be considered fair use and whether such use is sufficiently transformative. Yet there are considerable difficulties in establishing the relevant facts. In order to determine whether specific works were actually used, copyright holders may need details about the training datasets, how the data was acquired, the filtering processes and the development of the model. This kind of information can be obtained through litigation and the discovery process, but this is quite different

⁹ 17 U.S.C. § 107 (2018).

from having a general right granted by statute to obtain the information prior to starting any legal action.¹⁰ The approach taken in the United States thus shows a clear difference between substantive copyright protection and the right to access information. The fair-use doctrine decides whether a given use may be legally allowed, while litigation and discovery offer the means of acquiring evidence. Neither of these, however, gives rise to a general and independent right to find out whether a particular copyrighted work was used in the training of AI systems.

3.4 International Developments

The United Kingdom, Japan and Singapore are examples of jurisdictions which offer helpful additional insights into AI training and copyright. These countries show that it is possible for different combinations of copyright exceptions, text-and-data-mining provisions, licensing arrangements and regulatory measures to be used in dealing with AI training. Such developments are relevant since they prove that transparency does not have to take only one legal form.

Nevertheless, these jurisdictions should still be treated as merely providing comparative examples rather than as the subject of separate country studies. The aim is to highlight various approaches which may shed light on certain aspects of the right to know, for instance those relating to the disclosure of data sources, the allowed use of computational methods, licensing arrangements or the limitations on copyright enforcement. Inclusion of such cases should therefore be restricted to those developments which make a direct contribution to answering the main comparative question – whether copyright owners can obtain meaningful information about the way their works are used in the training of AI systems.¹¹

3.5 Comparative Analysis

The analysis comparing these approaches shows three main kinds of regulation. The European Union mainly follows a regulatory-transparency model, linking the copyright rules relating to text and data mining with specific duties imposed on general-purpose AI models. While this method places the greatest emphasis in law on transparency regarding the training data, there are still uncertainties as to whether general disclosure enables individual copyright holders to check that particular works have been used. In contrast, the United States mainly depends on litigation, the fair-use doctrine and evidentiary procedures. This way of proceeding lets the courts decide on a case-by-case basis whether it is legal to use data for AI training, but it offers no general statutory right to copyright owners to get information about how their data has been used; access to the relevant information may therefore have to be obtained through litigation and the discovery process.¹²

Other legal systems show examples of emerging or sector-specific strategies, usually achieving a balance between promoting AI innovation and protecting copyright by means of exceptions, licensing schemes or limited transparency measures. The comparison thus indicates that transparency and the right to know are related yet conceptually different. The EU shows how regulatory transparency can act as an institutional means of providing disclosure, whereas the United States illustrates the shortcomings of depending mainly on litigation to reveal details about hidden AI training practices. The main issue under

¹⁰ *Authors Guild v. Google, Inc.*, 804 F.3d 202, 214–25 (2d Cir. 2015).

¹¹ Copyright, Designs and Patents Act 1988, c. 48, § 29A (UK); Chosakukenhō [Copyright Act], Law No. 48 of 1970, art. 30-4 (Japan); Copyright Act 2021, No. 22 of 2021, §§ 243–244 (Sing.).

¹² *Thomson Reuters Enter. Centre GmbH v. Ross Intelligence Inc.*, No. 1:20-cv-00613, 2025 WL 458520 (D. Del. Feb. 11, 2025).

comparison is therefore not merely which jurisdiction offers greater transparency, but whether the available mechanisms give copyright holders access to information that is specific enough, easy to obtain and practical enough to enable them to determine if their works have been used and to exercise their current legal rights.

4 THE INDIAN LEGAL POSITION AND THE PROPOSED RIGHT TO KNOW

4.1 Indian Copyright Act, 1957 and AI Training

The Indian Copyright Act, 1957 does not contain a provision specifically regulating the use of copyrighted works for training generative artificial intelligence systems. Nevertheless, the existing copyright framework provides the starting point for analysing whether activities associated with AI training may engage copyright protection. Section 14 confers upon copyright owners exclusive rights in relation to protected works, including rights concerning reproduction, adaptation and communication of the work to the public, subject to the limitations and exceptions contained in the Act. The relevance of these rights to AI training depends upon the particular stage of the technological process under consideration.¹³

AI training may involve several distinct activities, including collecting publicly available works, downloading or storing copies, preprocessing and extracting information, converting material into machine-readable formats, incorporating the processed material into training datasets and subsequently using that dataset to train an AI model. These activities should not automatically be treated as one legally indistinguishable act. In particular, the question whether the temporary or permanent storage of a copyrighted work constitutes reproduction must be examined separately from the question whether the subsequent computational use of that work falls within an applicable statutory exception.¹⁴

The issue has acquired particular significance in *ANI Media Pvt. Ltd. v. OpenAI OpCo LLC*, decided by the Delhi High Court in July 2026.¹⁵ The proceedings involved allegations that copyrighted ANI material had been scraped and stored for training OpenAI's large language models. The Court considered, among other matters, whether storage of copyrighted material for AI training could constitute infringement and whether the Indian Copyright Act could apply where training and storage occurred on servers outside India.

The Court, at the *prima facie* stage, rejected the proposition that the location of the training servers automatically removed the dispute from Indian jurisdiction. It considered the process holistically, including the access to and transmission of data from India and the connection between training and outputs generated within India. This decision demonstrates that AI training presents questions that cannot be resolved simply by identifying where the final model is hosted. The geographical distribution of data collection, transmission, storage, training and output generation may make conventional territorial assumptions increasingly difficult to apply. At the same time, the case demonstrates why training and output should remain analytically distinct. A model may be trained on copyrighted material without necessarily reproducing that material in its outputs. Conversely, an output may reproduce a protected expression even where the legal basis of the underlying training activity remains contested. The copyright analysis must therefore identify the particular act allegedly engaging the owner's rights. The Indian framework

¹³ *Copyright Act, 1957*, § 14 (India).

¹⁴ *Copyright Act, 1957*, §§ 14(a)(i), 52(1)(a) (India).

¹⁵ *ANI Media Pvt. Ltd. v. OpenAI OpCo LLC*, CS(COMM) 1028/2024 (Del. HC July 24, 2026) (India).

consequently provides substantive copyright protection but does not expressly answer the preliminary informational question central to this paper, “*how can a copyright owner determine whether a particular work was collected, stored or used in AI training?*”

4.2 Fair Dealing and AI Training

Section 52 of the Copyright Act provides statutory exceptions to infringement. Section 52(1)(a), in particular, recognises specified acts of fair dealing in relation to literary, dramatic, musical and artistic works, including fair dealing for purposes such as private or personal use, including research, criticism or review and reporting current events, subject to the statutory framework. The application of this provision to AI training is legally significant because AI developers may argue that training constitutes a form of research or computational analysis. However, the Indian concept of fair dealing should not simply be equated with the broader American doctrine of fair use. Section 52 operates through specifically identified statutory purposes, and Indian courts have developed contextual approaches to determining whether a particular dealing falls within the relevant exception.

The *ANI v. OpenAI* proceedings provide an important contemporary development. The Court considered whether the storage of ANI’s literary works for training the LLM underlying ChatGPT could fall within the expression “private or personal use, including research” in Section 52(1)(a)(i). At the *prima facie* stage, the Court accepted that AI-based model training could fall within the research component of the provision and concluded that the purpose requirement was satisfied.

Importantly, however, satisfying the purpose requirement did not automatically resolve the fair-dealing question. The Court separately considered whether the dealing itself was fair. It observed that there is no single uniform test equivalent to the American four-factor fair-use test and emphasised that fairness depends upon the facts, degree and overall circumstances of the particular use. Among the considerations identified was whether the training activity prejudiced the copyright owner’s interests, including actual or potential commercial harm. The Court ultimately held, at the interim stage, that OpenAI’s storage of ANI’s literary works for training fell within Section 52(1)(a) and therefore did not amount to infringement under Section 51 on the facts before it. It also found that ANI had not established, at that stage, substantial reproduction or memorisation of its works in the outputs relied upon. The Court expressly stated that its observations were for the purpose of deciding the interim application and would not determine the final outcome of the suit.

This is particularly important for the present research. The decision should not be interpreted as establishing that all AI training in India is automatically protected by fair dealing. Rather, it demonstrates that the legality of training may depend upon the purpose, nature of the use, commercial context, potential prejudice to the copyright owner and other factual circumstances. It also reinforces the importance of transparency. A copyright owner cannot effectively evaluate commercial prejudice, potential infringement or the fairness of a particular use without knowing whether and how the work was incorporated into the training process.

4.3 Absence of an AI-Specific Transparency Framework

Although Indian copyright law provides substantive rights and statutory exceptions, it does not presently establish a comprehensive, AI-specific transparency regime giving copyright owners a clear statutory mechanism to determine whether their individual works have been used in AI training.

There is no general copyright provision expressly requiring an AI developer to provide an individual copyright owner with:

- a) confirmation that a particular work was used for training;
- b) identification of the dataset in which the work appeared;
- c) information concerning the source from which the work was obtained;
- d) information regarding whether the work was actually used during model training;
- e) identification of the AI model trained using the work; or
- f) a dedicated procedure through which the copyright owner can verify or challenge such use.

This distinction is fundamental. Copyright protection and copyright transparency are not the same thing. Section 14 may establish the rights of the copyright owner, but it does not necessarily provide the information required to determine whether those rights have been engaged. The absence of a dedicated transparency mechanism becomes particularly problematic because AI training datasets can be extremely large and may be compiled from multiple sources. A copyright owner may discover that an AI system produces content resembling their work, but such similarity alone may not establish that the particular work was included in the training dataset. Conversely, the absence of visible reproduction in an output does not establish that the work was never used during training.¹⁶

The *ANI* proceedings illustrate this evidentiary complexity. The Court noted that although the storage of ANI's original literary works during training was admitted, the plaintiff had not provided sufficient factual material to establish the alleged memorisation and regurgitation of those works in the outputs relied upon. The Court treated the question of memorisation and regurgitation as requiring evidence. This reveals an important gap between knowing that AI training occurred and knowing precisely what copyrighted works were used in that training.

Accordingly, India's current framework may address the substantive question of infringement through existing copyright principles, but it does not provide an equally developed answer to the preceding informational question.

4.4 Evidentiary Challenges in Proving AI Training Data Use

The most significant difficulty for copyright owners may therefore be evidentiary rather than purely substantive. Traditional copyright disputes generally involve an identifiable work and an identifiable act of copying, communication, adaptation or other restricted use. AI training creates a different evidentiary environment. The allegedly relevant copying may occur during automated data collection and preprocessing, potentially involving enormous quantities of works. The copyright owner may have no direct observation of the process and may have no access to the underlying technical records. This produces a significant information asymmetry.

The copyright owner may possess:

- the original copyrighted work;
- evidence concerning authorship and ownership; and
- evidence that the work was publicly available.

The AI developer, however, may possess:

¹⁶ Dep't for Promotion of Indus. & Internal Trade, *Working Paper on the Interface Between Artificial Intelligence and Copyright Law, Part I* (Gov't of India 2025).

- data acquisition records;
- dataset documentation;
- source lists;
- filtering records;
- preprocessing information;
- provenance information;
- training logs; and
- information concerning the particular model and training stage.

The person seeking to enforce copyright may therefore need evidence that is primarily controlled by the alleged infringer. This raises questions concerning the burden of proof and access to evidence. A copyright owner cannot reasonably be expected to prove an invisible technical process solely through speculation. At the same time, requiring AI developers to disclose unrestricted datasets could create serious concerns relating to trade secrets, confidential business information, cybersecurity and model security. A proportionate solution must therefore balance these competing interests.¹⁷

Digital and electronic evidence may be particularly important. Dataset records, server logs, acquisition records, metadata, provenance information and technical documentation could potentially assist a court in determining whether a particular work was collected or processed. However, the technical complexity of AI systems may make such evidence difficult for a conventional copyright litigant to obtain, interpret and challenge. The *ANI* decision demonstrates this evidentiary problem. The Court distinguished between the existence of stored training material and the claim that the model subsequently memorised and reproduced particular works. It held that the latter issue could not be established merely from the outputs relied upon at the interim stage.

This distinction is central to the proposed right to know. If the copyright owner cannot obtain information about training data use, the owner may be unable even to formulate the evidentiary basis of a potential infringement claim. Thus, transparency is not merely a policy preference. It may have direct implications for access to justice and the practical enforceability of copyright.

4.5 Can an Emerging Right to Know Be Derived from Existing Indian Law?

The absence of an express statutory right does not necessarily mean that the concept has no place within Indian law. A potential right to know could develop indirectly through existing substantive rights, procedural mechanisms, judicial interpretation, contractual arrangements and future regulatory intervention.

- First, existing copyright rights provide the substantive foundation. If a copyright owner possesses exclusive rights over reproduction, adaptation and other protected acts, information concerning the exercise of those rights may become practically necessary for enforcement. The right to know could therefore be conceptualised as an informational mechanism supporting existing rights rather than as an entirely new copyright entitlement.
- Second, procedural and evidentiary mechanisms may provide a route for obtaining relevant information in litigation. Courts may, depending upon the applicable procedural framework and facts, require parties to produce relevant evidence. Such mechanisms could help address information

¹⁷ U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (2025).

asymmetry once litigation has commenced. Their limitation, however, is that they may provide information only after the dispute has reached a sufficiently advanced procedural stage.

- Third, judicial interpretation may gradually develop principles concerning disclosure where AI training and copyright disputes come before Indian courts. *ANI v. OpenAI* is particularly important in this respect because it demonstrates judicial engagement with training data, territorial jurisdiction, storage, fair dealing and the evidentiary distinction between training and output.
- Fourth, regulatory intervention could create a more systematic mechanism. Instead of requiring every copyright owner to initiate litigation to obtain information, legislation or subordinate regulation could impose proportionate disclosure obligations upon relevant AI developers.
- Fifth, contractual and licensing mechanisms could establish transparency obligations voluntarily. Where AI developers obtain training data through licensing arrangements, contracts could require reporting, attribution, provenance documentation or payment mechanisms.¹⁸

These possibilities demonstrate that the right to know could emerge incrementally. However, none presently establishes a comprehensive and universally applicable Indian statutory right enabling copyright owners to determine whether their works have been used in AI training. The more defensible position, therefore, is that Indian law presently provides the foundations for an informational interest but does not clearly recognise a standalone AI training “right to know.”

4.6 Need for Legislative Recognition

The preceding analysis indicates that relying entirely upon existing copyright law may leave a significant practical gap. Copyright owners may possess substantive rights but lack the information necessary to determine whether those rights have been affected.

A future Indian framework should therefore consider recognising a proportionate transparency obligation concerning AI training data. Such a framework need not require disclosure of every individual item contained within a massive dataset. Instead, it could establish graduated obligations depending upon the nature and scale of the AI system.

Possible requirements could include disclosure of:

- significant categories and sources of training data;
- information concerning publicly accessible datasets;
- mechanisms for recording data provenance;
- information sufficient to verify the use of a particular copyrighted work where a credible claim is raised;
- procedures for copyright owners to submit verification requests; and
- appropriate documentation for regulatory or judicial review.

Such obligations should be balanced against legitimate interests in trade secrets, confidential information, cybersecurity and protection against misuse of technical information. The objective should therefore not be unrestricted access to proprietary datasets but sufficient information to make copyright protection practically meaningful. India could consequently adopt a model in which the right to know operates as an informational and procedural entitlement rather than as an automatic right to prohibit AI training or demand remuneration. Once use is established, the copyright owner could then determine

¹⁸ Regulation 2024/1689, recital 107, 2024 O.J. (L 1689) 1 (EU).

whether existing copyright law provides grounds for licensing, remuneration, enforcement or other remedies. The Indian position therefore reveals the central argument of this paper: the effectiveness of copyright protection in the AI environment depends not only upon the substantive scope of copyright but also upon the availability of information concerning the use of copyrighted works. The absence of a dedicated transparency mechanism may leave copyright owners facing an evidentiary disadvantage that existing copyright provisions were not designed to address.¹⁹

5 CONCLUSION

The increasing use of copyrighted works in AI training has exposed an important limitation in traditional copyright frameworks: the effectiveness of a legal right may depend upon the ability of its holder to obtain information concerning its use. Copyright owners may possess exclusive rights over their works, yet remain unable to determine whether those works have been collected, incorporated into training datasets or used to develop particular AI systems. The resulting information asymmetry creates difficulties not only for infringement claims but also for licensing negotiations, remuneration and other legitimate exercises of copyright interests.

The comparative analysis demonstrates that jurisdictions are responding to this problem through different legal approaches. The European Union has moved towards a regulatory model that places greater emphasis on transparency concerning AI training content, while the United States continues to rely substantially upon fair-use doctrine, litigation and evidentiary mechanisms. Other jurisdictions have adopted more limited or sector-specific approaches. These differences demonstrate that transparency can be structured through different legal mechanisms and need not necessarily require unrestricted disclosure of proprietary datasets.

For India, the existing Copyright Act, 1957 provides a substantive framework through rights of reproduction, adaptation and communication and statutory exceptions under Section 52. However, it does not expressly establish a comprehensive mechanism through which copyright owners can determine whether their works have been used in AI training. Recent judicial engagement with AI-training disputes demonstrates that existing copyright principles can address particular questions of legality, but the evidentiary problem concerning hidden training-data use remains significant. This paper therefore proposes the recognition of a legally enforceable but proportionate right to know. Such a right should not automatically create an entitlement to licensing, remuneration or prohibition of AI training. Instead, it should provide copyright owners with sufficient information to determine whether their works have been used and to assess what legal or commercial action may be appropriate.

A three-level transparency model is proposed. General transparency would require disclosure concerning categories and major sources of training data and relevant copyright policies. Targeted disclosure would permit a copyright owner with a credible basis for suspicion to seek verification of whether a particular work was used. Evidentiary disclosure would operate during formal proceedings under judicial supervision, with appropriate safeguards for confidential information and trade secrets. Ultimately, the right to know should be understood as a precondition for meaningful enforcement rather than an independent right to control AI development. Where there is no reasonable means of knowing whether a copyrighted work has been used, the practical value of existing copyright protection may be substantially

¹⁹ Regulation 2024/1689, recital 107, 2024 O.J. (L 1689) 1 (EU).

weakened. A balanced Indian framework should therefore reconcile the legitimate interests of copyright holders with innovation, confidentiality and technological development. Meaningful AI regulation requires not unrestricted disclosure, but sufficient, proportionate and actionable transparency.

* * *