FROM SHADOWS TO SOURCES: LEGALIZING AI TRAINING DATA TRANSPARENCY IN INDIA

Annie Sharon Lloyd[*]

ABSTRACT

The fast development of generative artificial intelligence (AI) systems—able to generate text, images, music, and code—has come under urgent legal and ethical scrutiny regarding the data used to train them. In India, the Copyright Act, 1957, makes no provision for disclosure by the developers of AI of whether or not copyrighted material has been incorporated in training their models. This blind spot in the law is hazardous to the rights of authors, artists, and other creators of content whose works could be used for gain without permission, attribution, or remuneration.

This essay considers the implications of the regulatory gap and argues for the imposition of a statutory requirement to disclose training data sources. It discusses the merits of embracing aspects of the U.S. Generative AI Copyright Disclosure Act, requiring AI developers to register and disclose copyrighted content used in training large-scale models. Through comparative legal analysis and contextual assessment, the paper investigates how India can develop an equilibrium disclosure framework that safeguards creators without excessively encumbering innovation or breaching trade secrets. In addition, the research lays down enforcement institutional challenges such as a centralised registry being required, definitional precision of AI-generated works, and the possibility of safe-harbour provisions. Last but not least, the paper provides pragmatic legislative suggestions for infusing transparency and accountability into the application of generative AI, bringing Indian copyright law in sync with the changing world order.

Keywords: AI Training Data, Copyright Law, Transparency, Generative AI, Disclosure.

[*] BBA LLB, THIRD YEAR, KRISTU JAYANTI COLLEGE OF LAW.

METHODOLOGY

This article employs doctrinal legal research to examine how Indian copyright law treats generative AI. It also compares India's response with global models such as the U.S. Copyright Disclosure Act 2024. The goal is to identify gaps in the law and recommend reforms that find a balance between innovation and creators' rights.

INTRODUCTION

Generative AI technologies have moved very quickly to become a key part of the digital economy, producing content that varies from text and images to music and computer code. They learn by scanning massive amounts of existing material, a lot of which consists of copyrighted human-created works. In India, where AI research and use are in full swing, the legal system is confronted with a key challenge: regulating the utilisation of copyrighted content in training AI without discouraging innovation or violating creators' rights. The Indian Copyright Act, 1957, currently does not address the specific challenges presented by generative AI. Particularly, there is no obligation for AI developers or platforms to reveal the sources of copyrighted material utilised in the training of their models. That loophole exposes Indian creators to unauthorised use of their content, erodes transparency, and triggers questions over accountability. It also makes enforcement more challenging, as rights holders are deprived of the information needed to identify or oppose misuse. At the global level, policymakers are starting to address these problems. For example, the U.S. proposed the Generative AI Copyright Disclosure Act in 2024, requiring AI developers to disclose copyrighted content employed in training through an official registry. This bill is a groundbreaking move towards balancing content creators' rights with the needs of AI development, setting an example that India can assimilate into its legal and technological landscape.

This article discusses the necessity of having a similar legislative regime in India. This article critically analyses the shortcomings of the current copyright regime concerning AI training data and assesses how international regulatory models, specifically the U.S. model, can be applied. The research also outlines the enforcement issues of a practical nature, such as verification of data, compliance on a cross-border basis, and protection of trade secrets. Finally, this research seeks to outline a policy framework that ensures transparency of AI training data and, at the same time, encourages innovation and protects creators' rights. As India strives to emerge as an AI technology leader, filling these gaps in law is crucial for establishing a just and responsible AI ecosystem.

CURRENT LEGAL FRAMEWORK IN INDIA

Globally, countries are adapting. The European Union's Artificial Intelligence Act mandates transparency in training datasets and requires labelling of AI-generated content. In the US, lawsuits are testing the boundaries of 'fair use' in AI training. India, by contrast, continues to rely on legacy copyright laws and a reactive takedown model under the Information Technology Rules.[1] The 'Fair Use' doctrine tries to balance the valid rights of copyright holders to manage and profit from their work with the social interest in utilising such work for inter alia criticism, review, research, personal use, etc. Under copyright law, the doctrine relies on the purpose of use, nature, or quantity of work duplicated, and effect upon the potential market. In the issue of whether the utilisation of copyrighted materials to train AI models qualifies for the exception of Fair Use, AI firms tend to plead for transformative use. In the Indian context, the doctrine of 'Fair Use' envisaged under Section 52[2] of the Copyright Act's scope extends to private or personal use, including research, translations, criticism, etc. However, the use of copyrighted material as training data for AI models is not explicitly mentioned.[3]

The Copyright Act, 1957, regulates the copyright regime in India. The Act provides copyright protection on literary, dramatic, musical, and artistic works, cinematograph film, and sound recordings. AI tools are capable of typically generating literary, musical, and artistic work. However, the grant of copyright on generated work depends upon various criteria, inter alia, originality of the work and creativity on the part of an author.[4] Originality is one of the essential measures to check the copyrightability of a work. The fundamental requirement of originality is that the work should have originated from the Author and not be derived. The United States adheres to the "modicum of creativity" doctrine, which recognises a work to be original if it is "independently created" and possesses a "minimum degree of creativity". The Supreme Court of India in the case of *D.B. Modak and Anr. v. Eastern Book Company and Ors.,*[5] held that

[1] Srinath Sridharan, 'Generative AI puts India's outdated copyright and liability laws to test' (Policy Circle, 13 May 2025) <https://www.policycircle.org/opinion/generative-ai-india-copyright-laws/> accessed 4 July 2025

[2] The Copyright Act, 1957 (Act 14 of 1957), s 52

[3] Shivam Vikram and Vanshika Mittal, 'Training AI Models: Intersection Between AI And Copyright' (Mondaq 20 February 2025) https://www.mondaq.com/india/copyright/1587932/training-ai-models-intersection-between-ai-and-copyright accessed 4 July 2025

[4] Nikhil Mishra and Digvijay Singh, 'AI-Generated Work and its Implications on Copyright Law in India' [2025] (30) JIPR file:///C:/Users/Admin/Downloads/JIPR-209+corrected+proof.pdf accessed 4 July 2025

[5] 2008 (1) SCC 1

there must be some substantive variation in the work for it to be copyrightable. The second prerequisite to be met by an AI concerning the ownership of copyrighted works is that it should fall within the protection of an 'author' as specified under the Indian Copyright Law. This would be undesirable because an AI has typically been considered not to possess a legal personality. According to the Copyright Act, 1957, an author is the first owner of the copyright.[6] The AI's developer or programmer helps the AI function and produce work by coding it, training it on data, and then recoding and reconfiguring it in response to the training data's output to ensure optimal AI performance. One major problem in this context is the function of upstream contributors such as developers/ programmers, designers, trainers, and data providers, who may be material contributors for authorship purposes as opposed to the users. If it is presumed that the work created by generative AI tools, such as ChatGPT, Google's Bard, DeepAI, etc. are original, but still no rights accrue with the users under Indian law because the term used under "causes the work to be created" under Section 2(d)(vi)[7] is of tremendous importance in this context and refers to a substantial contribution of humans in the making of work.

LEGAL AND ETHICAL IMPLICATIONS OF OPAQUE AI TRAINING DATA

The problem of AI-generated works is not specifically addressed by India's Copyright Act. However, AI-generated works might not be eligible for copyright protection due to the Act's requirement of originality and human authorship. For AI developers and users, this presents serious issues because they might not have any legal options in the event of a violation. AI's legal ramifications include making sure AI systems abide by the laws and regulations established by governments. An AI-powered program may breach privacy rules, for example, if it gathers personal data without the required consent. Legal obligations can be met by enacting legislation that prioritises AI transparency. The existing legal systems find it difficult to hold artificial intelligence and other non-human entities responsible. This uncertainty may result in possible inconsistencies with accepted legal norms. The data needed to train generative AI models is one of the most urgent issues that arises from generative AI's profound challenges to the fundamental presumptions that underlie copyright law. The largest training datasets for these models include millions of text documents, photos, audio samples, and other types of material. These models require enormous amounts of data. Most of this material is protected

[6] The Copyright Act, 1957 (Act 14 of 1957), s 17
[7] The Copyright Act, 1957 (Act 14 of 1957), s 2(d) (vi)

by copyright, but AI developers have frequently made little or no effort to seek the permission of rightsholders for the use of their works, leading to infringement concerns.[8]

Another issue that arises is that because AI learns from past data, it can reflect biases and current social injustices. AI models can be trained using such biased datasets, then used again in the development of further systems to reinforce, or even worsen, the discrimination. That is, they could result in unfair outcomes in important domains such as criminal justice, lending, or employment.[9] Another crucial concern is data privacy, particularly when sensitive or personal data is utilised in training without user agreement or appropriate anonymisation, which is against data protection regulations. Furthermore, the issue of accountability gets complicated because it is not always clear who is responsible when AI systems produce biased or damaging results—the creators, users, or data producers. Additionally, this raises legal questions around who is responsible for harm brought on by deepfakes, misinformation, or AI-generated content. Clear legal frameworks are urgently needed as AI develops to handle these issues and make sure that creativity isn't unnecessarily constrained. Furthermore, generative AI techniques have occasionally even produced the original watermarks, reproducing identical copies of the content they were trained on. Whatever these instruments produce, then, regardless of the degree, clarity, and engagement of human prompting, is at best derivative, at worst a replica, and most definitely not transformative. For this reason, the claim that AI-generated outputs ought to be protected by copyright is unfounded.

INTERNATIONAL CASE STUDIES

Sarah Silverman v. Open AI:[10] The famous comedian and author Sarah Silverman, and the famous author Paul Tremblay, also representing other authors, had filed a lawsuit against tech giant OpenAI, alleging that the company used their copyrighted books and works without taking their consent or giving them compensation to train their AI text generative system, ChatGPT. This particular case garnered significant attention because it raised questions about

[8] Adam Buick, 'Copyright and AI training data—transparency to the rescue?' (2025) 20 JIPLP 182

[9] Riya Singh, 'Legal Implications of Artificial Intelligence and Machine Learning: Ethics and Regulations' (The Amicus Qriae) https://theamikusqriae.com/legal-implications-of-artificial-intelligence-and-machine-learning-ethics-and-regulations/#:~:text=Bias%20and%20Discrimination&text=AI%20models%20can%20be%20trained,making%20by%20the%20AI%20system. accessed 7 July 2025

[10] Silverman v. OpenAI, Inc., 3:23-cv-03416, (N.D. Cal.)

the legal boundaries of copyright law regarding the use of copyrighted works for training AI generative models.[11]

The lawsuit's scope was considerably reduced by the U.S. District Court's partial finding. Citing insufficient evidence that ChatGPT outputs were significantly similar to the plaintiffs' copyrighted works or that OpenAI had purposefully deleted copyright management information, the court rejected several claims, including vicarious copyright infringement and violations of the Digital Millennium Copyright Act (DMCA). Except for the "unfair" component, which the court permitted to advance, claims under California's Unfair Competition Law were also largely rejected. The plaintiffs must show that certain ChatGPT outputs illegally duplicated their protected expression to prove the sole remaining claim of direct copyright infringement.

Getty Images v Stability AI:[12] The claimants belong to a group of businesses that own and run websites like Getty Images, which contain millions of visual assets spanning a wide range of topics, including photos and videos. The defendant, Stability AI, introduced Stable Diffusion, a deep learning text-to-image AI model that uses user-inputted image suggestions and text commands to create realistic, detailed synthetic images. Approximately 12.3 million visual assets from the Getty Images websites, along with their corresponding captions, as well as publicly accessible third-party websites, were used to train Stable Diffusion. Stability AI is being sued by Getty Images for copyright, database rights, UK registered trademarks, and passing off violations.

By excluding important accusations of copyright infringement about the AI model's training procedure, Getty Images considerably reduced the scope of its claims. The court is currently considering whether the Stable Diffusion model, which incorporates Getty's copyrighted content, amounts to secondary copyright infringement when imported or utilised in the UK. Additionally, Getty has pursued charges of passing off and trademark infringement, claiming that certain AI-generated photos fraudulently bear the "Getty Images" watermark, possibly deceiving users and harming the company's reputation. Stability AI has retorted that the

[11] Savan Dhameliya, 'PARTIALLY DISMISSED: TREMBLAY, SARAH SILVERMAN v. OPEN AI' (IPRMENTLAW, 18 March 2024) https://iprmentlaw.com/2024/03/18/partially-dismissed-tremblay-sarah-silverman-v-open-ai/ accessed 10 July 2025

[12] Getty Images v Stability AI, [2025] EWCA Civ 749

watermark appearances are accidental and not intentional misrepresentations, and that the model was trained in the United States, which is outside the jurisdiction of the United Kingdom.

COMPARATIVE ANALYSIS: U.S. GENERATIVE AI COPYRIGHT DISCLOSURE ACT

To increase transparency regarding the use of copyrighted content for generative AI model training, Representative Adam Schiff presented the Generative AI Copyright Disclosure Act of 2024. The proposed Generative AI Copyright Disclosure Act of 2024 attempts to introduce new transparency requirements for AI developers. The primary goal of the bill is to ensure that copyright owners have visibility into whether their intellectual property is being used to train generative AI models. If enacted, the law would require companies to submit notices to the U.S. Copyright Office detailing the copyrighted works used in their AI training datasets.[13] After that, the Copyright Office would keep an open database of these notices so that authors could check to see if their work was listed. It is expected that increased openness will assist copyright holders in making better decisions on the licensing of their intellectual property and, when necessary, pursuing recompense. Strengths of the Act involve empowering creators by offering transparency into the utilisation of their work, which could lead to licensing arrangements and remuneration. It also encourages responsibility from AI developers. Weaknesses, however, include practical challenges to developers, particularly smaller organisations, in monitoring and reporting the vast datasets used in training AI. The provision of a "sufficiently detailed summary" is vague, and this could result in variable disclosures. Furthermore, considering its ongoing IT modernisation projects, the Copyright Office has a significant administrative burden managing the database.

In contrast, the U.S. strategy prioritises transparency without actually limiting the use of copyrighted material, unlike the European Union's AI Act, which mandates stricter compliance, including compliance with copyright legislation and possible licensing requirements. In the U.K., suggestions have gravitated towards permitting AI businesses to utilise copyrighted work without permission except where rights holders choose to opt out, a position that has come under massive objection from creators. France has experienced legal action against business entities such as Meta for inciting unauthorised utilisation of copyrighted work in AI training,

[13] Danner Kline, 'The Generative AI Copyright Disclosure Act of 2024: Balancing Innovation and IP Rights' (2024) 15 The National Law Review https://natlawreview.com/article/generative-ai-copyright-disclosure-act-2024-balancing-innovation-and-ip-rights accessed 8 July 2025

demonstrating a more enforcement-oriented stance. The U.S. Act is a compromise, seeking to find a balance between innovation and creators' rights by way of disclosure, whereas the other jurisdictions either have more stringent controls or confront a continuous debate on the desired level of control.

India has no special legal framework or statutory provision dealing with the employment of copyrighted works in training AIs. The Copyright Act, 1957, does not particularly envision the generative AI challenges, and questions like fair use, licensing, or permission from authors are still unclear in the context of machine learning. In contrast with the U.S. Act, which at least requires disclosure and allows creators to determine unauthorised use, India's legal vacuum provides no transparency or redress for rights owners. These disadvantages affect Indian creators, since they could unwittingly see their works used without acknowledgement or remuneration. Further, in the absence of disclosure, Indian regulators and courts struggle to determine whether AI systems violate protected content. Though the U.S. model is not ideal—focusing on transparency rather than restriction—it may provide a basis for India to develop its own disclosure-oriented or licensing-focused strategy, hopefully adapted to domestic content industries and digital rights issues. Essentially, though the U.S. has taken a proactive (albeit limited) regulatory step, India finds itself reactive and thus susceptible to uncontrolled AI research and possible copyright misuse.

ASSESSING INDIA’S READINESS FOR TRAINING DATA DISCLOSURE

With initiatives like Digital India, which seeks to develop digital infrastructure, encourage digital literacy, and deliver government services online, digitisation has taken front stage in India. The Digital India journey included e-governance projects like DigiLocker and DigiYatra. On August 12, 2023, the president signed the Digital Personal Data Protection (DPDP) Bill into law, continuing this journey. The Act will give individuals the ability to manage their data while enabling Indian companies to handle digital personal data responsibly. It is anticipated that it will significantly affect people, companies, and the economy as a whole. The Act offers Indian organisations the chance to improve customer and stakeholder trust while streamlining their data gathering and procedures.

The proposed Digital India Act is anticipated to repeal the IT Act in its entirety, along with other rules framed under it, such as the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 ("Intermediary Guidelines"), whereas the DPDP Act aims to omit or amend only specific sections of the IT Act. The Ministry of Electronics and Information Technology ("MeitY") may make changes to the Intermediary Guidelines about AI until the Digital India Act is implemented. Additionally, the government has been actively evaluating whether it would be wise to enact specific regulations on AI. A distinct chapter of the Digital India Act or particular provisions may be used to implement such a law. The government may establish guidelines for exchanging anonymised personal and non-personal data under this new law, including information obtained by intrusive devices, according to media sources.

Particularly in light of the Digital Personal Data Protection Act (DPDP Act), the Department for Promotion of Industry and Internal Trade (DPIIT) is essential to India's preparedness for training data disclosure. DPIIT has a direct influence on how companies handle and share training data because of its emphasis on ease of doing business, industrial growth, and fostering innovation within the startup ecosystem. The DPDP Act, which is enforced by the Ministry of Electronics and Information Technology (MeitY), affects the gathering and use of training data by requiring express consent for data collection and processing. Data privacy laws have an indirect impact on the innovation and growth that DPIIT's programs, including Startup India, seek to promote.

PRACTICAL CHALLENGES

Data Fragmentation and Accessibility: Different organisations have different schemas and metadata standards, and the available data is dispersed throughout them. For the large-scale development of AI systems, this could be a roadblock. Acknowledging this, the government offered certain programs aimed at facilitating data access and breaking down silos, but these have not been implemented well. The National Data Governance Framework Policy (NDGFP) seeks to standardise the administration of anonymised and non-personal data across government agencies to advance data-driven governance.

Privacy and Security Concerns: There are significant risks associated with the possible misuse of personal data and the absence of regulatory constraints around data anonymisation. Although the majority of businesses have privacy notifications, many are missing important components like breach notification procedures and informed consent. A website's privacy notice should include information on the personal data that will be processed, its intended use, the data subjects' rights, and how they can exercise those rights. It should also specify how the data principal can file a complaint with the board in the event of a disagreement.

Lack of Expertise and Awareness: The lack of qualified experts with both theoretical understanding and real-world experience in AI development and implementation is one of the main obstacles to the advancement of artificial intelligence (AI). There is a shortage of specialised education and training programs that can provide such expertise at scale in many areas, especially in developing nations. Additionally, companies and government organisations are not well-informed on the advantages, uses, and dangers of AI technologies. This ignorance frequently leads to a reluctance to embrace AI or to fully utilise its potential.

Resource Costs: A large financial commitment is required to use AI technologies, which many organisations, especially small and medium-sized businesses (SMEs), may find burdensome. In addition to purchasing hardware and software, the expenses also include data gathering and cleansing, system integration, employee training, continuing maintenance, and regulatory compliance. Additionally, huge datasets and high-performance computer infrastructure—both of which are costly and resource-intensive are frequently needed to construct trustworthy AI models. Many potential adopters are left behind in the absence of grants, funding, or affordable AI-as-a-service models, which exacerbates the digital gap between big tech companies and smaller organisations.

Need for Collaborative Approach: The development of AI is a multidisciplinary topic that benefits greatly from cooperation between many stakeholders, including government organisations, business executives, academic institutions, startups, and civil society. But without unified approaches and common frameworks, AI projects often become dispersed, resulting in redundant work, inefficient use of resources, and uneven standards. A cooperative strategy guarantees that ethical considerations are incorporated from the outset, public policy fosters innovation, and research is in line with industry demands. Countries and organisations may advance AI while tackling issues like bias, transparency, and fair access to technology by establishing collaborations and open innovation platforms.

RECOMMENDATIONS FOR LEGAL REFORM IN INDIA

Amend the Copyright Act: The Copyright Act will need to be updated to incorporate clauses that specifically include AI-generated works and the training data that goes into creating these systems. In the first place, the Act should have precise legal definitions for concepts like "generative AI," "machine learning models," and "training data disclosure." These definitions will give a legal basis to gauge compliance and responsibility. In addition, a new subsection should be added mandating that AI developers and platforms provide detailed reports of their training data to a specified regulatory agency. This would involve revealing whether they used copyrighted material, the type of data, and the proportion of its usage. This addition would enhance responsibility and guarantee that rights holders are not exploited unknowingly.

Establish a Training Data Registry: There must be a centralised Training Data Registry, either in the Copyright Office or in the Ministry of Electronics and Information Technology (MeitY). It would be a repository where AI developers would have to register and make public datasets used for training high-risk AI models, particularly large language models (LLMs), generative visual models, and music generation tools. The emphasis has to be on models with broad social reach, high commercial value, or high capacity for content generation. The registry would enhance transparency, aid in tracking possible infringement, and benefit regulators in policy implementation and research management.

Penalties: Compliance with AI copyright law should be ensured through a system of tiered penalties. The severity and commercial extent of the infringement would determine how degree of penalties. For example, increased fines would be levied on firms that purposefully train models on copyrighted works without authorisation for profit-making purposes. Along with financial penalties, injunctions could be granted to stop the deployment of infringing AI models, and takedown mechanisms would enable the speedy removal of outputs or datasets proven to infringe copyright. These enforcement mechanisms would act as a deterrent and safeguard the interests of creators.

Exemptions: To find a balance between innovation and protection of rights, the law must acknowledge exemptions for specific uses. These might consist of de minimis use—where copyrighted material used is inconsequential or incidental—and use of material already public domain. Non-commercial and academic AI research must also be covered with safeguards, particularly projects engaged in ethical, safety, or public-interest objectives. Such exemptions would safeguard innovation while not losing sight of copyright limits.

Transparency-by-Design: A future-proof legal infrastructure needs to infuse transparency-by-design principles into AI regulation. There should be an encouragement, or even mandatory publication, of model cards that provide information on the model's architecture, training data, known biases and limitations, and intended applications. Provenance-tracing tools (i.e., tracking the origin and processing of data throughout the model life cycle) should be encouraged to provide traceability and accountability. Policymakers must also provide incentives, like regulatory fast-tracks or funding preference, to open-source AI projects that are transparent and follow copyright standards. This would promote a more ethical and open development culture for AI.

CONCLUSION

India has come to a pivotal point in its pursuit of global leadership in the field of artificial intelligence. Although the technical momentum is powerful, the legislation and regulation under which AI operates, especially in terms of copyright and transparency of data, are still lagging behind and in need of update. The absence of explicit provisions in the Copyright Act of 1957 concerning AI-generated works and the use of copyrighted content to train AI models exposes creators and hinders the development of a fair and responsible AI ecosystem. The obscurity of training data not only erodes authors' and artists' rights but also generates legal insecurity for AI developers and users.

By borrowing from global models like the U.S. Generative AI Copyright Disclosure Act and harmonising with international best practices, India can produce a futuristic legal framework. It should entail explicit statutory definitions, mandatory disclosure requirements of data, establishing a centralised registry of training data, graduated penalties for abuse, exemptions that protect innovation, and institutionalisation of transparency-by-design. In addition, India's preparedness—spearheaded by digital agenda items like the DPDP Act, the upcoming Digital India Act, and the functioning of MeitY and DPIIT—indicates a supportive policy climate. To make this work, however, the government will need to overcome real barriers like data fragmentation, dearth of talent, cost of resources, and the requirement for inter-institutional coordination. A revised copyright and AI regulation regime will not only safeguard creative rights but also enable responsible innovation, making India a model for the world in balanced, rights-centric AI development.