



LEGAL UPDATE

AI TRAINING AND INDIAN COPYRIGHT LAW: NAVIGATING DATA SCRAPING AND INFRINGEMENT

INTRODUCTION

The constant development of generative artificial intelligence (AI), and the challenge of “web-scraped” data have become central topics in debates about IP. The data AI systems draw upon are vast, unorganised, and often composed of copyrighted articles, academic journals, creative blogs, published works, and high-quality photographs.¹ Strictly speaking, data scraping refers to the automated, unauthorised extraction of proprietary information from digital databases to feed into the machine learning model training process. The legal implications of such an extraction process carry significant threats of infringement of intellectual property rights, with questions regarding the legality of using such content for “non-consumptive” machine learning purposes remaining one of the most contentious and litigious legal areas.² What is at stake is the question of whether the use of these expressive works, whether for their expressive content or as statistical fuel for a machine, is an infringement of the author's monopoly on his/her work.³

THE INDIAN LEGAL FRAMEWORK AND WEB SCRAPING

The Copyright Act, 1957, was drafted with the idea of protecting the creative expression of literary, dramatic, musical and artistic works, but Indian jurisprudence is very unclear on the legality of high-volume data scraping. Section 14 of the Copyright Act gives the copyright owner the exclusive right to reproduce, store, and distribute their work to the public. The primary legal conflict is figuring out if the feeding of copyrighted works into a “latent” (mathematical) training algorithm of an AI model constitutes “reproduction” under current law.

Indian law does not say much about this particular technical application, but the law does provide that unauthorised copying as well as downstream republishing could lead to significant liability under Section 51, while Section 52 of the Act codifies the doctrine of fair dealing, which allows use of the copyrighted work for specific, enumerated purposes, including for research, private study, criticism, and news review.⁴ Large-scale commercial data scraping for the purpose of training profit-driven AI engines is no longer within the historical scope of these narrow, protective exceptions, and thus is increasingly likely to become the subject of legal challenges. As a result, the current notion of fair dealing for domestic AI developers is jeopardised.⁵





LEGAL UPDATE

THE MECHANICS OF COPYING: REPRODUCTION AND ADAPTATION

One of the key technical and legal concerns is the process of "copying" used in the AI training phase. As digital data are absorbed and analysed by modern machine learning systems, they make high-speed copies of the information. Lacking a specific statutory exception, these temporary copies may be considered "reproduction", which could mean that the process of "AI training" is actually a form of copyright infringement.

There was no specific definition of reproduction or copying in Indian copyright law to specifically address digital tokenisation, and the courts are now faced with the question of whether even storage of a work's expression, for machine-readable purposes, amounts to infringement. During the training of generative AI, the system breaks down data into tokens to learn the statistical patterns without really "reading" the work from a human perspective. But the judiciary must now determine whether to apply the broadest reading of reproduction rights on which the interests of both creators and users of digital tools could be based, or to limit the scope of the rights to traditional uses of expression.

Further, the Reproduction Right in Section 14 is based on the concept of "substantial similarity" that could replace the original in its main commercial market. After the precedent in *R.G. Anand vs. Deluxe Films* (1978), infringement is accepted only if there is an unmistakable similarity of the final output to the original work as a whole.⁶ Generative AI outputs will not come close to this unless they regurgitate the training data, which can be done intentionally through "prompt injections. The "ideas vs. expression" distinction will be the last major line of defence against over-protection, with AI models that learn styles and patterns of authorial expression, not necessarily the specific content, potentially falling under the fair-use exceptions.

INTERNATIONAL LAWSUITS AND LEGAL DISCUSSIONS

The legal landscape around AI training data is in flux worldwide, with several key court cases including The NY Times' blockbuster lawsuit against OpenAI shaping the current situation.⁷ This litigation focuses on the question of whether the use of newspaper articles in training generative models is infringing use or a "transformative technical purpose" protected by fair use. The debates that have been proposed in this section usually fall into two conflicting positions: defenders of the copyright claim will say that the AI is trying to make a substitute product for their copyrighted





LEGAL UPDATE

material, while advocates for AI will respond that the AI is simply trying to learn from a general pattern that does not violate copyright.⁸

LANDMARK LITIGATION IN INDIA: ANI VERSUS OPENAI

Asian News International (ANI) vs. OpenAI,⁹ in the Delhi High Court, is the first Indian case of an organisation taking on a businessman for employing training data used in the creation of AI algorithms. The plaintiff claims that OpenAI illegally scraped its articles of the news to use in the training of its large language models (LLMs), such as ChatGPT. The case was filed by the news agency ANI on 19 November 2024 in the Delhi High Court, and the judgment is presently reserved vide order dated 27 March 2026.

According to ANI, hacking their paywalls to access breaking news content and use it in Retrieval Augmented Generation (RAG) systems inflict irreparable damage.¹⁰ But, according to OpenAI, this is a self-contained process that can't harm the organic numbers of the news agency, and the company has not been liable in other key geographies for copyright infringements, and the procedure does not create a copy-and-paste reroll but rather combines answers from a wide range of publicly available sources of data.¹¹ Further, when the case reaches the Hon'ble Supreme Court of India, Sections 14 and 51 will provide much clarity on whether the automated compilation of proprietary news content for training AI constitutes a breach of the copyright owner's exclusive reproduction rights.

TRANSFORMATIVE USE

Indian law does not recognise the “transformative use” doctrine, which is contrary to the United States copyright laws. Delhi High Court, in *University of Cambridge vs. B.D. Bhandari* (2011), recognised another purpose for a work but at the same time emphasised that the reproduction right is limited to the original expressive purpose of the author of a work.

THE DOCTRINE OF EXTRACTIVE USE

The extractive use theory assumes that facts, information, or knowledge that is contained in a copyrighted work should not itself be protected by copyright. In the case of *Akuate Internet Services vs. Star India*¹², the court pointed out that to monopolise facts would hamper the constitutional pursuit of information dissemination. OpenAI has contended that its models learn facts, but not the expression. This, however, suggests that the underlying facts will have to be





LEGAL UPDATE

extracted from the entirety of the expressed work, raising other potential reproduction questions under the current law.¹³ OpenAI is big on the concept and expression dichotomy, a concept that they say they're only using unprotected facts to share the information democratically. This is grounded in the right to freedom of information and speech in the Constitution, Article 19(1)(a). In this argument, the debate about AI is framed as a new iteration of a “library” or a “search engine”, yet publishers have signalled that synthesising information is a new type of disruptive and fundamentally different activity from the one they have traditionally been selling.¹⁴ Further, existing technical solutions include “*Robots.txt*” protocols and opt-out mechanisms that developers/publishers use. These mechanisms can be used to indicate to users that they don't wish to be scraped, but their effectiveness is open to debate; they may be disregarded by illegal or violating users.¹⁵ Whether the bars can be legally enforced is another hot topic, with publishers saying that if they are not respected, it is interpreted as “wilful infringement”.¹⁶

JURISPRUDENTIAL PERSPECTIVES ON AI AUTHORSHIP AND THE RAGHAV STANDARD

Whether Artificial Intelligence can qualify as an “author” of a copyright work is a key dispute in IP law. The definition of an author is given in Section 2(d), which deems the author to be the person who created the work.¹⁷ In particular, Section 2(d)(vi) applies to computer-generated works, which refers to literary works, dramatic works, musical works, and artistic works where “author” means the person causing the work to be generated.¹⁸ Further, the incident is known as RAGHAV (Robust Artificially Intelligent Graphics and Art Visualizer). The artist, Ankit Sahni, an Indian, designed a work called ‘Suryast’ with RAGHAV AI and submitted it for registration to the Indian Copyright Office.

For the first time, they also double-listed the person and the AI tool as co-authors. The registration was accepted for a short time, but eventually the Copyright Office withdrew it, raising significant legal anxieties about the comprehension of a computer-generated work as the authorship of AI.¹⁹ Currently, Indian law does not have any specific provision granting an AI system authorship in its own right, and the judiciary has been reiterating that the presence of intellectual creativity and subjective intent is essential for authorship, and not something that an AI can have. Further, it is pertinent to note that the international position on AI as an author (or a co-author) is consistent; the majority are steadfastly against it.





LEGAL UPDATE

In *Thaler vs. Perlmutter*, it was similarly agreed by the courts that the Copyright Act assumes human authorship of copyrightable works, establishes a copyright term based on an author's life, and requires human intent in order to qualify a group of two people as joint authors.²⁰

Although mainly AI-created content is not subject to copyright, some jurisdictions are trending toward differentiating between works created solely by AI and works produced with the help of AI. If human input is substantial, either as detailed direction, specific elements or significant post-generation refinement, then the human can be noted as the only author of the final piece. In these cases, the AI becomes a powerful instrument akin to a camera or computer program, but it is no creative partner.

First, to be protected, a human has to show a degree of creative control and that the AI was a tool for human intellectual expression. Second, because AIs lack legal personality or legal intent, we cannot establish legal joint authorship.

WAY FORWARD

Given the scenario, a hybrid system can balance innovation with the protection of copyright owners. A promising concept would be to develop a '*One Nation, One License*' system through a Collective Management Organisation (CMO).²¹ With this model, the owners of AI models would dedicate a portion of their earnings and their usage towards a centralised pool, and we would see the rights holders' share going from the central collection according to any measurable metrics of scraping. This would move the burden from copyright owners to a "standardised" approach that would make compensation proportional to the usefulness of the data applied.²²

This would create an opportunity to check data usage on the fly through auditing and annual transparency reports. In addition, establishing a legally binding regime of 'Opt-Out Standard' (OS) would shift the industry from 'scrape by-default' to 'respect by-default'. The publishers could prevent unauthorised training easily by marking content with unambiguous tags, with the result of the violation being statutory damages. This could minimise the cost burden of individual litigation, be more cost-effective, and ensure the economic value of the AI ecosystem is distributed fairly to those who create the 'raw material' for technological development.





LEGAL UPDATE

Extractive use and fair dealing doctrines could afford some protection; however, their utility for big-picture AI training is undecided. Indian legal community will need to navigate with care to ensure that the principles of authorship and creativity are not trampled down to advance the speed of algorithms, and that the traditional role and importance of the creative class within the cultural and economic life of the Digital Age are not sacrificed.

* This article is written by Ms. Lamia Himayat Kidwai, Advocate, High Court of Delhi, under the supervision of Mr. Sami Ahmed, Partner, ARBIVIS PARTNERS LLP LAW OFFICES.

1. “Generative AI and LLM Learning Paths” (NVIDIA2025) <<https://www.nvidia.com/en-us/learn/learning-path/generative-ai-llm/>> accessed June 14, 2026.
2. Copyright, AI Training, and LLMs: Historical Perspective and Infringement Analysis” (Copyright Clearance Centre, January 8, 2025) <<https://www.copyright.com/blog/copyright-ai-training-and-llms-historical-perspective-and-infringement-analysis/>> accessed June 14, 2026.
3. Khurana & Khurana, “Copyright Protection in LLM AI Training Part 2” (Khurana & Khurana, January 27, 2025) <<https://www.khuranaandkhurana.com/2025/01/27/copyright-protection-in-llm-ai-training-part-2/>> accessed June 14, 2026.
4. Pallavi Rao, “‘Fair Use’ in the Age of AI” (India Corporate Law, April 29, 2025) <<https://corporate.cyrilamarchandblogs.com/2025/04/fair-use-in-the-age-of-ai/>> accessed June 14, 2026.
5. Shantanu Sood, “A Copyright Conundrum- AI and Fair Use - Quest IP” (Quest IP, August 21, 2024) <<https://questip.com/a-copyright-conundrum-ai-and-fair-use/#>> accessed June 14, 2026.
6. sauhritik8871, “RG Anand v. M/S Deluxe Films AIR 1978 SC 1613” (Jus Scriptum, March 30, 2024) <<https://www.juscriptumlaw.com/post/rg-anand-v-m-s-deluxe-films-air-1978-sc-1613-1>> accessed June 14, 2026.
7. Fabrice Guyot-Sionnest, “NYT v. OpenAI: The Times’s About-Face - Harvard Law Review” (Harvard Law Review January 13, 2026) <<https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/>> accessed June 14, 2026.
8. Stolen, “Columbia Undergraduate Law Review” (Columbia Undergraduate Law Review, June 9, 2025) <<https://www.culawreview.org/ddc-x-culr-1/nyt-v-openai-and-microsoft>> accessed June 14, 2026.
9. [CS(COMM) 1028/2024],
10. “ANI v. OpenAI: Where Will the Court Draw the Line? – DSK Legal: True Value, True Values” (Dsklegal.com2024) <<https://dsklegal.com/ani-v-openai-where-will-the-court-draw-the-line/>> accessed June 14, 2026.



LEGAL UPDATE

11. “Indian Journal of Integrated Research in Law PROTECTING CREATORS’ RIGHTS IN THE ERA OF AI: DEFENDING ANI IN THE CASE OF ANI VS. OPEN AI, INC” <<https://ijirl.com/wp-content/uploads/2025/07/PROTECTING-CREATORS-RIGHTS-IN-THE-ERA-OF-AI-DEFENDING-ANI-IN-THE-CASE-OF-ANI-VS-OPEN-AI-INC.pdf>>.
12. Harish Khan, “Case Study: Akuate Internet Services Pvt. Ltd. vs. Star India Pvt. Ltd.” (Legal Wires, August 22, 2024) <<https://legal-wires.com/case-study/case-study-akuat-internet-services-pvt-ltd-v-star-india-pvt-ltd/>> accessed June 14, 2026.
13. “AZB & Partners” (azb2025) <<https://www.azbpartners.com/bank/delhi-hc-navigates-uncharted-waters-in-landmark-ai-copyright-dispute/>> accessed June 14, 2026.
14. “AZB & Partners” (azb2025) <<https://www.azbpartners.com/bank/delhi-hc-navigates-uncharted-waters-in-landmark-ai-copyright-dispute/>> accessed June 14, 2026.
15. BTLJ, “Opt-Out Approaches to AI Training: A False Compromise - Berkeley Technology Law Journal” (Berkeley Technology Law Journal, April 18, 2025) <<https://btlj.org/2025/04/opt-out-approaches-to-ai-training/>> accessed June 14, 2026.
16. Altacit Global, “AI-Generated Works in India: Examining the ‘Raghav’ Signal in Copyright Law” (Medium, January 31, 2026) <<https://medium.com/@AltacitGlobal/ai-generated-works-in-india-examining-the-raghav-signal-in-copyright-law-b0b95325e783>> accessed June 14, 2026.
17. Shireen Yachu, “A Look at Generative AI in Terms of Intellectual Property Rights” (Vidhi Centre for Legal Policy, March 19, 2025) <<https://vidhilegalpolicy.in/blog/a-look-at-generative-ai-in-terms-of-intellectual-property-rights/>> accessed June 14, 2026.
18. “Recognition of AI Co-Author Raghav- an Anomaly?” (Centre for Study and Research in Intellectual Property Rights [CSRIPR], NUSRL, March 11, 2022) <<https://csriprnusrl.wordpress.com/2022/03/11/recognition-of-ai-co-author-raghav-an-anomaly/>> accessed June 14, 2026.
19. “United States Court of Appeals FOR THE DISTRICT OF COLUMBIA CIRCUIT” <<https://media.cadc.uscourts.gov/opinions/docs/2025/03/23-5233.pdf>>.
20. Sukanya Sarkar, “Exclusive: India Recognises AI as Co-Author of Copyrighted Artwork” (MIP, August 5, 2021) <<https://www.managingip.com/article/2a5bqo2drurt0bxl7ab24/exclusive-india-recognises-ai-as-co-author-of-copyrighted-artwork>> accessed June 14, 2026.
21. Ambika Aggarwal, “One Nation, Forced Licenses, Multiple Payments: (Un)Balancing AI Innovation & Copyright” (SpicyIP, January 4, 2026) <<https://spicyip.com/2026/01/one-nation-forced-licenses-multiple-payments-unbalancing-ai-innovation-copyright.html>> accessed June 20, 2026.
22. Editorial Staff, “DPIIT’s Proposed One Nation – One License – One Payment Policy: What Does It Mean for the Content Creators & Artists? - Trademark & Patent Attorneys for Registration to Litigation” (Trademark & Patent Attorneys for Registration to Litigation January 28, 2026) <<https://selvams.com/blog/dpiits-proposed-one-nation-one-license-one-payment-policy-what-does-it-mean-for-the-content-creators-artists/>> accessed June 20, 2026.