THE COMPUTATIONAL INGESTION PIPELINE AND THE FRONTIERS OF COPYRIGHT LAW: A DOCTRINAL DISSECTION OF MACHINE LEARNING TRAINING, JUDICIAL RECKONING, AND LEGISLATIVE REFORM IN INDIA

Subhransu Sekhar Hota, Bennett University

ABSTRACT

The recent emergence of large-scale generative AI has brought about an institutional crisis between computation and intellectual property. The purpose of this article is to examine if the discrete technological acts involved in the modern machine learning training pipeline – data collection, transient storage, persistent storage, tokenization, preprocessing, and weight-embedding – constitute copyright infringement under the Copyright Act, 1957 alone, or collectively. The analysis takes a stage-by-stage approach, using a doctrinal mapping approach based on the principles of 'granularity' and 'mapping' and questioning each stage of the pipeline in relation to the statutory exclusive rights of reproduction and adaptation. It then examines the applicability of the current statutory defences, such as the fair dealing provisions in Section 52(1)(a) of the Copyright Act, 1957, and the non-applicability of the United States' transformative-use doctrine in India's closed statutory regime, and the implications of the statutory silence on text and data mining in India. The High Court of Delhi's landmark generative-AI copyright case, the proceedings of which are followed from the reservation of judgment on 27 March 2026 to the submissions of the court-appointed amici curiae and the industry's response, are discussed in particular. Lastly, the article challenges the Working Paper on Generative AI and Copyright recently released by the Department for Promotion of Industry and Internal Trade and exposes the administrative and economic shortcomings of the so called "One Nation, One License, One Payment" approach, suggesting a legislative alternative that is more balanced: a purpose-neutral text and data mining exception combined with machine-readable opt-outs and market-driven data cooperatives.

I. Introduction: The Technologico-Legal Ingress

The core legal issue is a fundamental tension between modern AI systems, which are entirely rely on the ingestion, analysis and processing of enormous bodies of human-created works, the vast majority of which fall within the scope of copyright. AI developers use automated web scrapers to crawl the public internet and ingest datasets with billions of literary, artistic, musical and dramatic works, without prior authorization or economic compensation to the original creators.[1]

The legal issues surrounding use of generative AI lie primarily in the "input" aspect of the technology, particularly the sequential actions needed to compile a dataset, store information, tokenize the natural language input and train a neural network, as these steps create distinct doctrinal challenges under existing copyright laws.

This academic argument turned into an instant, high-stakes judicial battle between two entities, with ANI Media Pvt. Ltd. v. OpenAI Inc. serving as India's leading generative-AI copyright case, pitting a 1957 statute designed for an era that predated digital networks, not to mention deep neural networks, against a computational data ingestion.[2]

The responses to these questions have a huge economic and cultural impact. The judicial and legislative changes in India are not just going to impact the local technology ecosystem, but will also be a barometer of how the Global South will assert their sovereign economic and cultural interests in the digital economy, in the face of global technology companies investing heavily in India's AI technology infrastructure. These judiciary and legislative developments in India are going to have ramifications not only on the local technology ecosystem, but will also set a benchmark for asserting the economic and cultural interests of the Global South in the digital economy, as global tech firms invest heavily in India's technology infrastructure to support and build AI capabilities.

The Copyright Act, 1957, had a historical paradigm that was print-based, where the copying of

[1] *See* ANI Media Pvt. Ltd. v. OpenAI Inc., CS(COMM) 1028/2024 (Delhi High Court) (pending); *see also* Aishwarya Giridhar, *ANI v. OpenAI: Where Will the Court Draw the Line?*, DSK Legal, https://dsklegal.com/ani-v-openai-where-will-the-court-draw-the-line/.

[2] ANI Media Pvt. Ltd. v. OpenAI Inc., CS(COMM) 1028/2024 (Delhi High Court) (pending); *see* Lokesh Vyas, *ANI v OpenAI: Not Everything an LLM Does is Copyright Infringement*, SpicyIP (Apr. 2026), https://spicyip.com/2026/04/ani-v-openai-not-everything-an-llm-does-is-copyright-infringement.html.

an expressive work referred to its physical duplication for commercial sale.[3] The legislature amended the statute several times over the years, most recently in 1994 and 2012, to cover digital networks, software protection, and digital rights management, but failed to consider the phenomenon of computational systems that use an expressive work for "mathematical extraction" rather than for human consumption.[4] This lack of legislation has resulted in a situation of uncertainty for creators and innovators alike.

In order to find solutions to this problem, this article systematically and stage by stage maps the computational training pipeline with the provisions of the Copyright Act, 1957. It breaks down the machine learning pipeline into its constituent technical steps to pinpoint exactly at which point copyright infringement occurs, assesses the potential effectiveness of current copyright defences, challenges the Government's recent proposals for copyright-friendly policy, and suggests a market-driven approach to copyright legislation that balances incentives for human creativity with competition in technological innovation.

II. Theoretical and Doctrinal Underpinnings of Digital Ingestion

Under Indian law, the legal validity of the training of AI will depend on delineating the scope of exclusive rights that can be claimed by the copyright holders and the scope of limitations on those rights under the Indian fair dealing doctrine. This necessitates a consideration of the ontological nature of digital copying, the statutory definition of 'infringement' under Section 14 and judicial interpretation of the infringement under Indian law.[5]

A. The Ontological Nature of Digital Copying

In the analog era reading a book and copying a book was a different act, and a different business. This ontological distinction is lost in the digital context, where reading is a cognitive act of no reproduction and copying involves the author's exclusive rights in a mechanical reproduction process. Without replication of the binary representation of a digital file, the computer will not be able to "read," process or analyse that file.

A machine learning algorithm must load a copyrighted document into RAM, copy it to multiple temporary cache buffers, and store it in a distributed parallel storage system for parallel

[3] The Copyright Act, No. 14 of 1957, India Code (1957).
[4] *See* The Copyright (Amendment) Act, No. 38 of 1994; The Copyright (Amendment) Act, No. 27 of 2012.
[5] The Copyright Act, No. 14 of 1957, §§ 14, 51 (India).

computation. From a technical perspective, each time we perform an act of computational "analysis" we must first perform an act of physical "reproduction". This is contrary to the idea that training is the same as human reading, and that there is no need for an intermediate, fixed copy on a server of the computational model, because human reading doesn't require such fixed copies on servers of any kind.[6]

B. The Right of Reproduction under Indian Law

The author of a literary, dramatic or musical work has the exclusive right to "reproduce the work in any material form including the storing of it in any medium by electronic means."[7] Copyright requires no specific technology or media to be protected, and this is certainly a broad and technology-specified phrase. The amendment of 2012 added the words "as stored by electronically means" to the reproduction right so that the storage of a protected work in digital form is covered by the reproduction right.[8]

Whereas jurisdictions impose a requirement for "fixation" for a stable period of time as a condition for the reproduction right, the Indian law doesn't impose such a requirement with regard to digital storage. Uploading a copyrighted file to a server's hard drive, or even storing it in a high-speed memory buffer during an ingestion run is "in a material form" and is "electronic means" of reproduction. The infringement of copyright under Section 51 is therefore a prima facie violation of copyright and it is infringed whenever a person, without a licence from the owner, does anything that the owner has the exclusive right to do.[9] Copyright infringement under Section 51 is thus a prima facie violation of copyright, and it is infringed whenever a person, without a licence from the owner, can do anything that the owner has the exclusive right to do.

C. The Substantiality Threshold and the Perception Paradox

One of the most often cited principals in defence of technology developers is that "acts in relation to a work" include acts in relation to a "substantial part" of the work. Indian courts have consistently followed a qualitative test for the determination of substantiality, instead of

[6] *Cf.* New Pub. Standard, *Delhi High Court Reserves Judgment in ANI v. OpenAI: Why the Verdict Matters Far Beyond India*, The New Publishing Standard (Apr. 1, 2026), https://thenewpublishingstandard.com/2026/04/01/ani-vs-openai-delhi-high-court-judgment-ai-copyright-india/.
[7] The Copyright Act, No. 14 of 1957, § 14(a)(i) (India).
[8] The Copyright (Amendment) Act, No. 27 of 2012, § 14 (amending § 14(a)(i)).
[9] The Copyright Act, No. 14 of 1957, § 51 (India).

a quantitative test. In R.G. Anand v. Delux Films the Supreme Court set the standard for infringement as whether the reader, spectator or the viewer, upon reading, viewing or watching both works, gets "an unmistakable impression that the subsequent work is a copy of the original" works.[10]

This is a "human perception" test, which poses a doctrinal dilemma in relation to the training of AI. At training time, no human audience sees the texts being processed; they are ingested, converted to a numerical array and processed by an algorithm that learns the statistical weightings. Under the R.G. Anand standard, the ingestion does not produce an "unmistakable impression" of copying in a human spectator's mind since there is no expressive copy presented to a human viewer during this internal process, proponents argue.

The argument, however, is that it mixes up the reproduction right under Section 14(a)(i) (which is one completed when the work is reproduced on its own by unauthorized electronic storage) with the similarity test under the R.G. Anand infringement test, which was traditionally used in cases involving derivative public works like plays and films, and not internal intermediate technical copying. Technical copying at the input stage is an independent infringement which happens before, and even if, the model is not subsequently released to a human audience.

III. A Granular Stage-by-Stage Dissection of the Machine Learning Pipeline

The problem is that the mechanical steps of training a large language model need to be correlated with the exclusive rights of the Copyright Act, 1957, in order to determine the point at which copyright liability arises. The modern pipeline is not a single event, but a series of computational actions, with their own distinct doctrinal questions. The chain goes like this: data collection (reproduction rights); storage and caching (electronic storage of reproduction rights); preprocessing and tokenization (adaptation / format-shifting); weight-embedding (raising the question of reproduction rights by memorizing and/or encoding the data).

A. Stage One: Data Collection and Dataset Compilation

The process starts with web scraping, which involves systematically and automatically collecting data from the public web. AI companies use proprietary crawlers to scrape millions of web pages, articles, books, and images to create large training sets like Common Crawl.AI

[10] R.G. Anand v. Delux Films, AIR 1978 SC 1613, (1978) 4 SCC 118 (India).

companies use proprietary crawlers to download millions of web pages, articles, books, and images to create large training sets, like Common Crawl.[11]

The action is directly taken on Section 14(a)(i) doctrinally. The download of a webpage or document and the creation of a local dataset of that expression requires physical expression of the expression in a tangible medium. Since these works are absorbed from start to finish at this initial stage, there can be no defence of 'non-substantial copying'. Reproduction is total, intentional and not done with the permission of the licensor under Section 51.

The technology developers claim to see an implied licence to extract and index publicly available content. This is the contention that doesn't differentiate between search engine indexing and generative training. As opposed to copy-and-paste models, search engines temporarily copy content to be able to index it and send people back to the original site to preserve the original creator's market, while generative models ingest content to train a system that can replace the original content and have users get the substance of the original content without having to go through the source.[12] Implied-licence is not such a broad doctrine that it permits the commercial extraction of genes for machine learning.

B. Stage Two: Storage, Caching, and Server Replication

After compilation, these datasets are loaded onto distributed high-performance file systems, and then are "cached" on distributed servers to enable parallel, high throughput training. The safe harbour available to temporary or intermediate caching is very narrow in India. The exception in Section 52(1)(b) only applies to the 'transient or incidental' storage of works in the 'technological process' of electronic transmission or communication to the public, but that is not the case of permanently storing a work for training purposes, which is a core part of the commercial production of an AI model.[13] These datasets are therefore not only unauthorized copies, but also a continuing series of unauthorized copies.

C. Stage Three: Preprocessing, Text Cleansing, and Tokenization

Data needs to be pre-processed and tokenized prior to being fed into a neural network.

[11] *See* Defending ANI in the Case of ANI v. Open AI, Inc., *Protecting Creators' Rights in the Era of AI*, 5 Indian J. Integrated Rsch. L. (2025), 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] *Cf.* Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).
[13] The Copyright Act, No. 14 of 1957, § 52(1)(b) (India).

Preprocessing includes deduplicating text, removing metadata and cleaning the raw HTML to convert formatted documents into a single, uniform plain text format. Then, the continuous text is split into smaller pieces, called tokens, which are associated with a number. Conceptually, a sentence like Copyright protects expression is broken down into a set of sub-word tokens and then represented as a series of integer indices which are selected from the fixed vocabulary of the model; this transformation is deterministic, and in principle could be reversed.

The tech companies say that tokenization eradicates the "expressive content" of the text, turning it into a series of abstract mathematical numbers not under copyright control. From a doctrinal perspective, it is incorrect that this is the case under Indian law for two reasons. The medium neutrality of reproduction is protected by Section 14(a)(i), which allows the copyright holder to protect the work in any "material form."[14] While such a numerical token sequence may be optimized for computing purposes, it is a direct translation and mathematically reversible translation of the original expression. It's not a distinction of substance, but of words, syntax and structure, that still exists within the tokenized array.

D. Stage Four: Weight-Embedding and Parameter Optimization

In the last step, the arrays are converted to tokens and then fed into a neural network. The model makes successive changes to its internal parameters (weights) using gradient descent and backpropagation, so as to reduce the error in its predictions. In formal terms, the training goal is to maximize the conditional probability of the model predicting each successive token of the training corpus, given the tokens that come before it in the corpus, which is achieved by minimizing the cross-entropy loss between the model's prediction and the actual next token in the corpus over all of the corpus. This stage involves two advanced doctrinal issues.

First, do weight parameters constitute a "derivative work" or an "adaptation"? According to Section 2(a), adaptation is defined in connection with certain acts e.g. translation, abridgement.[15] Model weights are abstract numbers that represent correlations in the data, and do not fit easily into this definition. However, the distribution of the weights is based solely on the expressive input of the training corpus and the resulting model can therefore be seen as an unauthorized digital compilation or derivative encoding of the works protected.

[14] *Id.* § 14(a)(i).
[15] *Id.* § 2(a).

Second, does the embedding process involve reproduction? Empirical studies of computer science show that neural networks often learn "by copying" parts of the text they optimize to during the production process. Memorized verbatim passages upon prompting indicate that the weights are not just abstract statistical rules of grammar and syntax, but carry with them the actual expressive content of the text in a compressed, recoverable form.[16]

IV. The Exhaustiveness of Indian Fair Dealing Defences

Without any special provision on text and data mining, it is the general defence under the Copyright Act, 1957 which AI developers in India can rely upon. Analysis shows that the fair dealing framework is constitutionally weaker to protect large-scale commercial training.

A. Section 52(1)(a) and the Statutory Exhaustiveness Doctrine

India has a non-generalised fair use doctrine, but a purpose based fair dealing doctrine in section 52(1)(a).[17] Large-scale AI development is a very commercial, capital-intensive, activity, and therefore, it is not possible to claim that harvesting of millions of works wholesale and automatically, even for the purpose of AI development, is for "private or personal use" or "research" as understood in 1957. The research exception aims at academic, single, human research and not at big-scale exploitation of the database for commercial purposes.

B. The Incompatibility of US-Style "Transformative Use"

The American doctrine of "transformative use" is often used by developers to work around the strictness of Indian law. Proponents of "fair learning" state that the purpose of training is not to display expressive content but to extract statistics information, thus for the purpose of training, it is a "classic" transformative use under the United States fair use doctrine, as expressed in Authors Guild v. Google, Inc.[18]

The doctrine does not fit into the Indian structure of this analogy. Firstly, there is no independent multi factor balancing test in Indian courts that allows the non-essentiality of the technology to be excused based on the technology's "transformative" nature. Second, a use is not "transformative" unless it is one of the exceptions specified in Section 52(1)(a) and a

[16] *See* New Pub. Standard, *supra* note 6.
[17] The Copyright Act, No. 14 of 1957, § 52(1)(a) (India).
[18] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

commercial organization is not able to claim that its "transformative" motivation satisfies the "private or personal" criterion of the research exception.[19] A departure from the rules of the statute would be to transplant the American doctrine without any amendment to the law.

C. Implied Licence and Estoppel in Public Digital Spaces

The doctrine of implied licence also relies on the argument that by putting works on the open internet, creators are deemed to give an implied licence to crawl and process their works; creators that do not deploy technical barriers like a robots.txt exclusion file cannot assert infringement — an argument that does not satisfy Indian copyright formalities. A memory stick is not sufficient to meet the requirements of section 19, which requires all assignments or licences of copyright to be in writing and signed by the licensor or his appointed representative: an "implied" licence arising from the unawareness of a technical barrier will not be sufficient.[20] Furthermore, the doctrine is unsound, in that copyright does not require the owner to create a 'fence' to deter scrapers from using a work for commercial purposes. Failure to have a technical block is no waiver of statutory rights.

V. Judicial Reckoning: ANI Media v. OpenAI

The theoretical discussion took a turning point with the institution of a case before the High Court of Delhi, the first generative-AI copyright case, against ANI Media Pvt. Ltd. and OpenAI Inc. (hereafter, both parties are referred to as ANI).The theoretical discussion came to a turning point before the High Court of Delhi, the country's first generative-AI copyright case, against ANI Media Pvt. Ltd. and OpenAI Inc. (hereafter referred to as ANI)[21]

A. Factual Background and Procedural Trajectory

Asian News International (ANI), a major media news provider, filed a commercial civil suit against OpenAI Inc and its operating entities in November 2024, accusing OpenAI of systematically copying, scraping and using its proprietary news articles, without authorization, licence or compensation to train the ChatGPT models.[22] The complaint's four main concerns

[19] The Copyright Act, No. 14 of 1957, § 52(1)(a)(i) (India).
[20] *Id.* § 19.
[21] ANI Media Pvt. Ltd. v. OpenAI Inc., CS(COMM) 1028/2024 (Delhi High Court) (pending); *see Ani Media Pvt. Ltd v. Open Ai Opco Llc* (Oct. 17, 2025), Indian Kanoon, https://indiankanoon.org/doc/135313744/.
[22] *Online News Publishers Join Copyright Suit in Delhi High Court Against OpenAI, Maker of ChatGPT*, The Hindu (Feb. 12, 2025), https://www.thehindu.com/sci-tech/technology/digital-news-publishers-association-joins-copyright-suit-against-chatgpt-maker-in-delhi-hc/article69147400.ece.

were: first, unauthorized ingestion via crawlers; second, web traffic being redirected through verbatim regurgitations of copyrighted reports by ChatGPT, which effectively circumvented ANI's paywalls; third, bad faith because ANI requested a licensing agreement in October 2024, which OpenAI refused, followed by digital blocks on its domain and OpenAI blocklisting that domain from future datasets; and fourth, reputational harm from the model's hallucination of false stories attributed to ANI.[23]

On the first hearing, 19 November 2024, Justice Amit Bansal issued the summons. The lawsuit then turned into a multi-party lawsuit, and the court refused to issue an ex-ante injunction in favour of blocking ChatGPT's operations in India but did note OpenAI's own affirmation that it had stopped scraping ANI's live domain and would not include it in future training runs.[24] Hearing were conducted from late 2024 till early 2026 and it was decided on oral arguments that the court will reserve judgment, the final hearing of which was held on 27 March 2026, with the participation of major industry groups such as the Digital News Publishers Association (which includes The Hindu, The Indian Express, and Hindustan Times), the Federation of Indian Publishers and the Indian Music Industry.[25] Following hours of oral arguments, Justice Bansal reserved judgment, with the last hearing taking place on 27 March 2026.[26]

B. Doctrinal Deconstruction of the Four Judicial Issues

To resolve the dispute the court posed 4 main questions, each corresponding to a fundamental doctrinal debate in computational copyright law. The court stated four key questions that correspond to fundamental doctrinal issues in computational copyright law.

Issue I — Storage as infringement. The Defendants' retention of the Plaintiff's data for training ChatGPT violates the Plaintiff's copyright. This issue is for Stage Two. ANI said that the copying of its articles for ingestion and electronic storage is a completed appropriation of

[23] *See ANI Media Pvt Ltd v. OpenAI OPCO LLC*, AI and Copyright Case Tracker in India, CMS Law, https://cms.law/en/int/publication/artificial-intelligence-and-copyright-case-tracker/ani-media-pvt-ltd-v.-openai-opco-llc (last visited May 31, 2026).
[24] Ani Media Pvt. Ltd v. Open Ai Inc., CS(COMM) 1028/2024, 1–2 (Delhi High Court Mar. 18, 2025), https://indiankanoon.org/doc/95388763/.
[25] *Usage of Copyright Content: Digital News Publishers Join Legal Battle Against OpenAI*, Indian Express (Feb. 12, 2025), https://indianexpress.com/article/india/openai-case-indian-news-websites-copyright-9802312/.
[26] *Delhi HC Reserves Order in ANI-OpenAI Copyright Dispute*, Exchange4Media (Mar. 2026), https://www.exchange4media.com/industry-briefing-news/delhi-hc-reserves-order-in-ani-openai-copyright-dispute-153470.html; *[2026 Update] ANI v. OpenAI: Delhi HC Reserves Order on AI Copyright*, Prime Legal, https://blog.primelegal.in/india-first-ai-copyright-case-ani-vs-openai-delhi-high-court/ (last visited May 31, 2026).

the articles which falls within the ambit of Section 14(a)(i). OpenAI argued that the model isn't designed to retain expressive content, that expressive content is processed through relatively short-lived caches and swiftly merged with the vast amounts of non-expressive data, and that the data is then discarded after training, and that copyright is unrelated to facts or underlying information.[27]

Issue II — Output generation as a distinct infringement. If the access to such plaintiff's copyrighted information in order to create responses to users constitutes infringement. When asked specific questions, ChatGPT was able to echo entire paragraphs of ANI's articles, an exercise that amounted to public distribution without the company's authorization, in essence, competing with ANI's services, ANI has proved. OpenAI said it has encountered very few instances of such "regurgitation," which can be achieved by carefully providing a large amount of text from articles, with the user being solely responsible for any such prompt. OpenAI described such "regurgitation" as an uncommon occurrence caused by "adversarial prompting"—deliberately providing text from articles in large quantities—claiming that it's up to the user to take responsibility for such prompts.[28]

Issue III — Fair use under Section 52. Whether the defendants' use is fair dealing. In attempting to do so, OpenAI observed that training models to analyse language patterns falls under the "private use including research" exemption under Section 52(1)(a)(i), claiming that this represents a kind of computational research. ANI replied in saying that the research which Section 52 is regarding is academic and non-commercial research which is purely human-centric, and it can't be extended to the commercial extraction by a multinational enterprise as copyright can be waived or can be licensed only through an express written agreement under Section 19.[29]

Issue IV — Territorial jurisdiction. Whether Indian courts have jurisdiction as the defendants' servers are in the United States. OpenAI claimed that none of the alleged infringement took place in India. ANI relied on Section 62(2) of the Act, which allows an action to be filed when a plaintiff engages in business in the territory of India and causes "tangible economic and reputational damage" in Delhi, to argue jurisdiction was granted since OpenAI

[27] Ani Media Pvt. Ltd v. Open Ai Opco Llc (Dec. 12, 2025), Indian Kanoon, https://indiankanoon.org/doc/179208465/.
[28] *See* Vyas, *supra* note 2.
[29]The Copyright Act, No. 14 of 1957, §§ 19, 52(1)(a)(i) (India); *see* Vyas, *supra* note 2.

actively seeks out the Indian market and inflicts tangible economic and reputational harm in Delhi.[30]

C. The Submissions of the Amici Curiae

Justice Bansal invited two intellectual property experts, Advocate Adarsh Ramanujan and Professor (Dr) Arul George Scaria of NLSIU, Bengaluru as amici curiae for assistance to the court and their submissions revealed a doctrinal split that mirrors the global divide on AI policy.[31] Dr. Scaria suggested an "interpretive flexibility" model that focuses on public interest, scientific advancement, and technological innovations while cautioning that a literal approach to the twenties century rules in the context of machine learning would stifle India's machine learning development and make the country reliant on foreign technology. He emphasized that the overwhelming majority of the works are not being exploited for human enjoyment, but are being used as data points to understand the structure of the language, and that the fact that there is some transient technical copying occurring when training the models doesn't mean that we shouldn't allow it since it doesn't create a market substitute.[32]

Mr. Ramanujan made a doctrinal analysis, very strict, based on the text and precedent in the statute. He said that the court cannot engage in judicial lawmaking and must follow the literal language of the Act – Section 52 is an exhaustive list of exceptions and therefore anything beyond the listed exceptions is infringement as a matter of law; that commercial training cannot be regarded as "private research" or "criticism"; that the scale and commercial nature of OpenAI's ingestion of data is infringing on ANI's exclusive reproduction rights; and that the creation of any new exception (such as a text and data mining exception) is a policy decision that Parliament must make.[33]

VI. Policy Intervention: The DPIIT "One Nation, One License" Framework

The judiciary, as they struggled with the issue, wanted the executive to get involved. A

[30] The Copyright Act, No. 14 of 1957, § 62(2) (India).
[31] *Delhi High Court Appoints NLSIU's Prof. Arul George Scaria as Amicus Curiae in ANI's Copyright Case Against ChatGPT*, Nat'l L. Sch. of India Univ. (NLSIU), https://www.nls.ac.in/news-events/delhi-high-court-appoints-nlsius-prof-arul-george-scaria-as-amicus-curiae-in-anis-copyright-case-against-chatgpt/.
[32] *See Does Human Learning Equal Machine Learning? High Court of Delhi to Rule on Lawfulness of TDM for Machine Learning*, Kluwer Copyright Blog (noting amici oral submissions on Feb. 21 and Mar. 10, 2026), https://legalblogs.wolterskluwer.com/copyright-blog/does-human-learning-equal-machine-learning-high-court-of-delhi-to-rule-on-lawfulness-of-tdm-for-machine-learning/.
[33] *Id.*

committee set up by the Department for Promotion of Industry and Internal Trade (DPIIT) has issued a part one Working Paper on Generative AI and Copyright on 8 December 2025.[34]

A. Structure of the Framework

The Working Paper was entitled "Balancing AI innovation and copyright," and was an outright dismissal of both the zero-price text and data mining exception preferred by tech companies, and a purely voluntary licensing system preferred by copyright owners. It proposed instead a hybrid statutory licensing model with some key features: a mandatory blanket licence, which gives the developer a statutory right to use all "lawfully accessed" content for training, instead of relying on opt-out rights, all digital content is deemed included by default; a Copyright Royalties Collective for AI Training (CRCAT), a centralised non-profit governmental body that will be designated to act as the statutory collective for the statutory licence; revenue linked royalties, in which AI companies will pay a fixed percentage of their revenues to a central pool, which the government will then distribute to registered creators and copyright societies, based on usage metrics; and an AI Training Data Disclosure Form, whereby the developer will be required to make a disclosure specifying the data set to which their AI system has been connected.[35]

B. Doctrinal, Administrative, and Economic Critiques

The Working Paper was immediately attacked by copyright academics, civil society groups (including SFLC.in), economic think tanks (Esya Centre and Takshashila Institution) and international trade organisations as a balanced reform. Five major flaws are conspicuous.

Firstly the remedy comes before the wrong done. The Working Paper, which seeks to build a complex licensing structure without first determining whether AI training is an infringement under the Act, recognizes this is a question before the Delhi High Court, but still gets ahead of

[34]Dep't for Promotion of Indus. & Internal Trade (DPIIT), *Working Paper on Generative AI and Copyright, Part 1: One Nation One License One Payment — Balancing AI Innovation and Copyright* (Dec. 2025), https://www.dpiit.gov.in/static/uploads/2025/12/ff266bbeed10c48e3479c941484f3525.pdf; *see also* Press Release, Press Info. Bureau, *DPIIT Publishes First Part of Working Paper on AI–Copyright Interface* (Dec. 9, 2025), https://www.pib.gov.in/PressReleasePage.aspx?PRID=2200741.

[35] DPIIT, *supra* note 34; *see also Generative AI & Copyright Licensing — One Nation, One License, One Payment Framework*, IP Helpdesk (Dec. 29, 2025), https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/generative-ai-copyright-licensing-one-nation-one-license-one-payment-framework-2025-12-29_en.

itself and creates a mandatory licensing regime based on a notion of systemic infringement.[36]

Secondly, private versus voluntary public licences. The blanket licence runs counter to the tenets of free and open-source software, and public licensing systems (Creative Commons). A system of 'publics' coerces creators who intentionally give away their works or choose to publish them under share-alike licenses into a situation where they have to pay for their work – a situation that violates their philosophical and legal decision-making.[37]

Third, administrative impossibility. Today's models ingest hundreds of billions of data points. The identification of billions of rights holders globally and within the country, their ownership and banking details and then sending the micro-royalties to them are impractical in terms of administration, especially due to a poor disbursement record of the existing welfare levy boards in India.[38]

Fourth, distortion of the market and monopsony. The approach does not provide a competitive market for data, in that it allows only one government-set body to set rates and make payments, which is a monopsony without price discovery. The long-term quality of creative output would be lowered by high quality creators being paid less, and low-quality data farms being paid more, for their service. High quality creators would be under-compensated, low quality data farms over-compensated, which would reduce the long term quality of creative output.[39]

Fifth, capital flight by technology. If AI companies are forced to pay tax on their global revenue, then it will be harmful for Indian-startups scaling up and push developers to Singapore or US who have been more friendly towards them, the Takshashila Institution said.[40]

---

[36] *See* Akshat Agrawal, *One Nation, Forced Licenses, Multiple Payments: (Un)Balancing AI Innovation & Copyright*, SpicyIP (Jan. 2026), https://spicyip.com/2026/01/one-nation-forced-licenses-multiple-payments-unbalancing-ai-innovation-copyright.html; *A Doctrinal Critique of the DPIIT Working Paper on Generative AI and Copyright*, IPRMENTLAW (Jan. 4, 2026), https://iprmentlaw.com/2026/01/04/a-doctrinal-critique-of-the-dpiit-working-paper-on-generative-ai-and-copyright/.

[37] SFLC.in, *Comments on DPIIT Working Paper on Generative AI and Copyright (Part 1)*, https://sflc.in/comments-on-dpiit-working-paper-on-generative-ai-and-copyright-part-1/ (last visited May 31, 2026).

[38] *See* Esya Centre, *Response to Part I of DPIIT's Working Paper on Generative Artificial Intelligence and Copyright* (Feb. 12, 2026), https://www.esyacentre.org/documents/2026/02/12/response-to-part-i-of-dpiits-working-paper-on-generative-artificial-intelligence-and-copyright; *Innovation Under Licence: A Critical Analysis of DPIIT's Generative AI Copyright Proposal*, IIPRD, https://www.iiprd.com/innovation-under-licence-a-critical-analysis-of-dpiits-generative-ai-copyright-proposal/ (last visited May 31, 2026).

[39] *See* Esya Centre, *supra* note 38; *Gen AI, Copyrights, and Hybrid Licensing in India: Why the Assumptions May Not Sustain the Model*, BananaIP Counsels, https://www.bananaip.com/intellepedia/generative-artificial-intelligence-copyright-india/ (last visited May 31, 2026).

[40] Takshashila Inst., *Working Paper on Generative AI and Copyright (Part 1)* (Feb. 6,

VII. Comparative Global Legal Perimeters

India needs to take a look at the strategies of other big digital economies to come up with a sustainable framework.

America is based on the flexible, judicial fair use (17 U.S.C.107).[41] The computation of a text to produce search indices or even to generate numerical data is considered as transformative use, which does not replace the original market, by courts in the context of Authors Guild v. Google. This concept is now being challenged in current cases like New York Times Co. v. Microsoft Corp. and Andersen v. Stability AI Ltd., where the creators of the source text contend that generative models are doing a synthesis of competing commercial works, and therefore do not constitute a transformative use under the fourth fair use factor.[42]

The **European Union** has created a structured approach to legislation for copyright in the Digital Single Market (DSM), with the 2019 Directive on Copyright in the Digital Single Market. Article 3 provides for a mandatory exception for text and data mining for scientific research by research institutes and cultural-heritage institutions, while Article 4 offers a TDM exception (along with a necessary reservation right): rights holders will have the right to object and to reserve their rights, but only if this is done in a machine-readable format.[43]

In 2018, **Japan** enacted Article 30-4 of the Copyright Act, which is believed to be the most lenient around the world for training AI. The provision allows the exploitation of a work, where the user is not interested in the thoughts or feeling that are expressed, but rather "wishes to enjoy or make another person enjoy" the work. There is little distinction between commercial and non-commercial training, since the algorithm does not "enjoy" the work and is simply looking for statistical patterns that it then uses to produce its own art — the only exceptions to this being qualified, that is, if the ingestion is meant to replicate some artist's style or "unreasonably prejudices the interests of the copyright owner".[44]

2026), https://takshashila.org.in/content/publications/20260206-generative-ai-and-copyright-consultation.html.

[41] 17 U.S.C. § 107.

[42] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015); New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023); Andersen v. Stability AI Ltd., No. 3:23-cv-00201 (N.D. Cal. filed Jan. 13, 2023).

[43]Directive 2019/790, of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market, arts. 3–4, 2019 O.J. (L 130) 92.

[44] Chosakukenhō [Copyright Act], Law No. 48 of 1970, art. 30-4 (Japan); *see* Tatsuhiro Ueno, *The Flexible Copyright Exception for 'Non-Enjoyment' Purposes — Recent Amendment in Japan and Its Implication*, 70 GRUR Int'l 145 (2021).

Section 244 of the Copyright Act 2021, which is explicitly a safe harbour for "computational data analysis", was introduced into **Singapore**. "Computational data analysis" was included in an explicit safe harbour under Section 244 of the Copyright Act 2021, which was introduced in Singapore.[45] This exception, which is captured in the law by Singapore, has placed it at the forefront of the computational hub in the Asia Pacific region, subject to strict conditions: it must be a copy for computational data analysis; it must not be supplied to another person, except for use for collective research and verification; the user must have lawful access to the source work and must know or reasonably be expected to know that he/she is accessing a copy that is an infringement of the copyright.[46]

VIII. A Legislative Reform Architecture for India

India must reject the uncertainty in the current framework, and the over-reach of the DPIIT model. Parliament should rather update the Copyright Act, 1957, to create a modern system based on a market approach.

A. A Purpose-Neutral TDM Exception with Machine-Readable Opt-Outs

The legislature should enact a new permitted use in Section 52 to include a purpose neutral text and data mining exception. For content openly available on the public internet, it should be accompanied by a statutory reservation right to exclude the content from the exception, such as the robots.txt protocol or the cryptographic metadata standards, which will allow rights holders to exclude their works, as the exception would not apply if they are automatically accessed through circumvention of a paywall, subscription login or technical protection measure.[47] A developer which scrapes a domain containing a valid and machine-readable opt-out loses the safe harbour. From an operational point of view, the TDM safe harbour applies if a crawler does not detect an opt-out; if a valid opt-out is detected, the developer is required to get a market-negotiated commercial licence.

B. Market-Led Data Cooperatives

India should promote "data cooperatives" instead of a single market with a state-mandated

[45] Copyright Act 2021, § 244 (Sing.), https://sso.agc.gov.sg/Acts-Supp/22-2021/Published/?ProvIds=pr244-.
[46] *Id.*; *see also* David Tan & Thomas Lee, *Generative AI and Copyright Fair Use — Singapore*, NUS L. (TRAIL), https://law.nus.edu.sg/trail/generative-ai-copyright-fair-use/ (last visited May 31, 2026).
[47] *Cf.* Directive 2019/790, *supra* note 43, art. 4(3) (machine-readable reservation of rights).

monopsony such as the CRCAT. Co-operatives overcome this issue through aggregation and pooling (such as journalism, independent music, stock photography or other areas) by allowing them to negotiate competitive value-based licences with the major technology companies. Cooperatives offer dynamic, market-driven pricing, and can provide premium prices for high value verified datasets for the developers as well as provide a compensation mechanism that reflects the utility and value of the developers' data. Cooperatives can also act as a technically efficient decentralized alternative to a state board by handling technical opt-outs, conducting compliance audits, and enforcing on the members' behalf.[48]

C. Data Provenance and Transparency Obligations

The legislature needs to set transparency requirements for developers in order to make a market driven system viable. They should include mandatory, immutable provenance records, including source URLs, ingestion dates and copyright status of the ingested work, which would place the burden on the rights holder to be able to prove that the model did not reproduce a work verbatim, and have the burden of proof move to the developer to prove that ingestion was lawful under the safe harbour or under an active licence.[49]

D. International Standard-Setting

Last but not the least, India should use its own strength as a big technology developer and a huge exporter of cultural content to influence global standards. The Indian delegation should actively promote the international recognition of machine-readable opt-outs through the World Intellectual Property Organization's on-going dialogue on IP and frontier technology, so that an opt-out made by a news agency or creators of an Indian origin is respected by web crawlers from all jurisdictions.[50]

IX. Conclusion

Computational ingestion of copyrighted works for training of AI is the most important challenge to the copyright since the invention of the printing machine. The sequential actions of the machine learning pipeline—scraping the web, storing the data persistently, preprocessing

[48] *See* Agrawal, *supra* note 36.
[49] *Cf.* Directive 2019/790, *supra* note 43, art. 4.
[50] S*ee* World Intell. Prop. Org. (WIPO), *Conversation on Intellectual Property and Frontier Technologies*, https://www.wipo.int/about-ip/en/frontier_technologies/ (last visited May 31, 2026).

it and optimizing the parameters—clearly involve the exclusive rights to reproduction and adaptation under the Copyright Act, 1957. The absence of an open-ended fair use doctrine in India, in favour of a purpose specific fair dealing test, means that commercial training operations are very vulnerable to liability at present.

These issues have now come to a head in the pending litigation between ANI Media Pvt. Ltd. and OpenAI Inc., and the High Court of Delhi's upcoming ruling will be a gamechanger for the Indian IP landscape. While well-intentioned, the executive's proposed "One Nation, One License" could end up creating market perversities, ensuring a de facto monopsony in the form of a bureaucracy and creating technological capital flight. In order to maintain India's competitive edge without compromising its creative industries, the Parliament should establish a purpose-neutral exception for text and data mining along with machine-readable opt-outs, promote data cooperatives in a decentralized market-led mechanism and create strict standards against data provenance. This will uphold the autonomy, economic interests, and economic sovereignty of creators, give the developers legal clarity, and put India in the leading ranks in AI policy at an international level.