GENERATIVE AI, TRAINING DATA AND INDIAN COPYRIGHT LAW: BETWEEN SIGNIFICANT HUMAN INPUT AND INFRINGEMENT

Pratiksha Gehlot, B.A. LL.B., University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand, India

ABSTRACT

This review examines the intersection of generative artificial intelligence (AI), training data practices, and Indian copyright law, with particular focus on the doctrine of significant human input and its implications for copyright infringement liability. The paper synthesises contemporary scholarship, Indian judicial precedent, statutory provisions, and international instruments to analyse unresolved tensions within India's copyright framework as applied to AI-generated content. The objective is to identify how existing legal structures particularly fair dealing provisions under Section 52 of the Copyright Act, 1957; originality and authorship doctrine; and the emerging concept of significant human input interface with technological affordances of generative AI systems trained on copyrighted materials. Through critical examination of the Copyright (Amendment) Bill, 2023, domestic case law, and comparative analysis with jurisdictions including the United States and the European Union, this review identifies doctrinal lacunae, unresolved definitional questions, and areas requiring legislative clarification. The paper further integrates India's distinctive constitutional framework, specifically examining how fundamental rights under Articles 19(1)(a), 19(1)(g), 21, and 14 of the Constitution constrain and inform copyright policy regarding AI-generated content and training data governance. This constitutional analysis represents a unique contribution, distinguishing this review from Anglo-American copyright scholarship that predominates in existing literature. Key findings reveal that Indian copyright law operates without explicit statutory recognition of generative AI; significant human input remains judicially undefined; the fair dealing exception provides limited guidance for AI training scenarios; authorship frameworks fail to accommodate non-human creative processes; and constitutional dimensions require urgent judicial and legislative consideration. The review concludes that India requires urgently coordinated legal reform addressing AI training data governance, redefinition of fair dealing to accommodate technological innovation, clarification of originality standards, explicit authorship provisions for AI-generated works, constitutional alignment of copyright restrictions with

fundamental rights, and potentially new statutory rights for training data proprietors. This paper contributes to emerging jurisprudence on artificial intelligence and intellectual property by providing a comprehensive Indian legal analysis grounded in statutory text, constitutional principles, and comparative frameworks, while offering evidence-based policy recommendations for India's legal development in this critical domain

Keywords: Generative AI; Training Data; Copyright Infringement; Fair Dealing; Significant Human Input; Authorship; Originality; Indian Copyright Law; Digital Technology; Machine Learning.

1. INTRODUCTION

1.1 Background and Context

The emergence of generative artificial intelligence systems, including large language models, diffusion-based image generators, and multi-modal architectures, represents a watershed moment in the intersection of technology and intellectual property law. Systems such as GPT-4, Claude, Gemini, Midjourney, and Stable Diffusion are trained on vast corpora of internet-derived text, images, and other creative content, substantial portions of which are protected by copyright. The computational processes through which such systems extract patterns from training data, generate novel outputs, and produce human-interpretable creative artefacts remain insufficiently addressed within existing copyright frameworks, particularly within Indian jurisprudence.

The legal status of generative AI training raises profound questions: whether the ingestion of copyrighted materials into training datasets constitutes copyright infringement; whether outputs generated by AI systems infringe copyright in source materials; whether such outputs qualify for copyright protection; and how traditional copyright concepts—fair dealing, authorship, originality, and infringement apply to technological processes designed to abstract, synthesize, and recombine features from training data.

For India specifically, these questions acquire heightened significance. India's Copyright Act, 1957, codifies principles developed primarily in reference to print, broadcasting, and early digital technologies. The Act predates machine learning by decades and contains no express provisions addressing AI-generated content, training data use, or the licensing frameworks emerging in response to generative AI. The Copyright (Amendment) Bill, 2023, while introducing technological adjustments, does not comprehensively address generative AI

challenges.[1] Simultaneously, India hosts a burgeoning technology sector, substantial research in machine learning, and significant litigation concerning technology intellectual property.

The doctrine of significant human input discussed in Section 52(1)(j) of the Copyright Act (sui generis database protection)[2] and evolving through judicial interpretation has emerged as a potential limiting principle. Courts across multiple jurisdictions have begun emphasising the requirement of sufficient human creative contribution to qualify content for copyright protection. However, Indian jurisprudence remains fragmented on the precise threshold, the relationship between significant human input and infringement analysis, and how this doctrine applies to hybrid human-AI creative processes.

1.2 Significance and Research Rationale

This review is motivated by three considerations. First, Indian copyright law operates without comprehensive judicial or legislative guidance specifically addressing generative AI. Existing decisions predate widespread deployment of generative systems and do not address AI training scenarios or output generation. Second, the doctrine of significant human input, while referenced in recent scholarship, lacks definitive judicial articulation in Indian courts and remains nebulously applied across comparative jurisdictions. Third, India's position as a technological innovator and developing economy requires a copyright framework that accommodates innovation incentives, creator compensation, and data-driven technological development a balancing act that existing statutory language insufficiently addresses.

This review addresses these gaps by providing comprehensive analysis of how Indian copyright law, through statutory text, constitutional principle, and judicial precedent, currently addresses (or fails to address) generative AI training data and output generation. The paper identifies specific doctrinal tensions, unresolved definitional problems, and areas requiring legislative or judicial clarification.

1.3 Scope and Purpose

This paper reviews scholarly literature, statutory provisions, case law, and policy analysis addressing generative AI, copyright, and training data through the lens of Indian law. The

[1] Copyright (Amendment) Bill, 2023, Bill No. 152, Lok Sabha, India (pending).
[2] Copyright Act, 1957, No. 14, Acts of Parliament, 1957 (India).

review focuses on five primary dimensions: (1) the statutory framework governing copyright infringement, exceptions, and defenses under the Copyright Act, 1957; (2) the judicial development of fair dealing doctrine and its application to technological contexts; (3) authorship and originality principles and their compatibility with AI-generated content; (4) the emerging concept of significant human input across jurisdictions; and (5) comparative analysis with United States, European Union, and other common law jurisdictions to identify best practices and doctrinal alternatives available to Indian courts and legislators.

2. METHODOLOGY

2.1 Literature Selection Process

This review employs a selective, topic-focused approach rather than exhaustive systematic review. The methodology prioritizes authoritative, peer-reviewed sources; judicial decisions from Indian courts and comparative jurisdictions; statutory materials and legislative history; and policy documents from international bodies including the World Intellectual Property Organization (WIPO). This selective approach is intentional, reflecting the nascent state of jurisprudence on generative AI and copyright within Indian legal scholarship. A comprehensive systematic review would be premature, as foundational legal questions remain unresolved and scholarly consensus has not emerged on critical definitional issues such as "significant human input," the scope of fair dealing for AI training, and authorship in computer-generated works. Therefore, this review prioritizes depth of doctrinal analysis over breadth of source inclusion, focusing on sources that directly address the intersection of generative AI, copyright law, and training data governance.

The review also incorporates grey literature including government white papers, legislative committee reports, and policy position papers from intellectual property organizations recognizing that emerging issues such as AI and copyright often receive preliminary attention through policy documents before comprehensive scholarly treatment. This inclusion reflects the rapidly evolving nature of AI technology and the corresponding urgency of legal policy development.

2.2 Database and Search Strategy

Primary sources were identified through multi-stage systematic searching across specialized

and general academic databases. The search strategy employed both controlled vocabulary and free-text searching to maximize retrieval comprehensiveness.

(1) Indian Legal Databases: Direct search of SCC Online (Supreme Court Cases), Indian Kanoon (comprehensive Indian legal repository), and Ministry of Corporate Affairs official repositories provided access to authoritative judicial decisions, statutory materials, and legislative records. These databases were prioritized for Indian case law and statutory interpretation, ensuring that judicial precedent reflects the most current and authoritative interpretation of the Copyright Act, 1957.

(2) Academic Databases: Google Scholar, SSRN (Social Science Research Network), and specialized intellectual property repositories including the American Intellectual Property Law Association (AIPLA) database were systematically searched. SSRN was prioritized for working papers and prepublication scholarship addressing emerging issues in AI and copyright. Google Scholar was used to identify citation relationships and track how foundational articles have been cited and built upon in subsequent scholarship.

(3) International Materials: WIPO publications, including policy reports, research papers, and the WIPO Magazine, were consulted to identify international policy developments and comparative perspectives. WIPO's International Bureau maintains comprehensive collections on AI and copyright issues across member jurisdictions, providing global context for India's regulatory approach.[3]

(4) Specialized Legal Journals: International journals including Columbia Journal of Law & the Arts, UCLA Entertainment Law Review, Stanford Technology Law Review, Vanderbilt Journal of Entertainment & Technology Law, and the nascent AI and Law Review were hand-searched for relevant articles. These journals are recognized as authoritative venues for intellectual property scholarship addressing technology-related copyright issues.

(5) Statutory and Legislative Materials: Direct consultation of the Copyright Act, 1957, Copyright (Amendment) Bill 2023, and legislative committee reports, including Lok Sabha and Rajya Sabha discussions, provided primary legal materials. The Legislative History of the Computer Software Amendment, 1994 was reviewed to understand the statutory development

[3] World Intellectual Property Organisation (WIPO), Artificial Intelligence and Intellectual Property Policy: A Global Perspective (2023).

of computer-generated work provisions.[4]

Search terms employed included multiple iterative strategies: (a) "generative AI and copyright," "machine learning copyright," "artificial intelligence copyright India," "AI training data copyright"; (b) "fair dealing copyright exceptions," "fair dealing India," "Section 52 Copyright Act," "fair use AI"; (c) "significant human input copyright," "authorship AI," "originality AI," "human authorship requirement"; (d) "Copyright Amendment Bill 2023," "AI training data copyright infringement," "text and data mining"; (e) "comparative AI copyright law," "copyright AI United States," "copyright AI European Union," "Australia AI copyright amendments."

2.3 Inclusion and Exclusion Criteria

Inclusion Criteria: (1) English-language publications, reflecting the language of Indian legal practice and international scholarship; (2) materials published between 2018–2026, with particular emphasis on 2022–2026 to capture recent developments in generative AI and corresponding legal scholarship; (3) sources addressing Indian copyright law, generative AI, training data, fair dealing, authorship, originality, or the intersection of these concepts; (4) peer-reviewed journal articles, books, book chapters, or publications by recognized legal scholars with expertise in intellectual property, copyright, or technology law; (5) judicial decisions from Indian courts (Supreme Court, High Courts, Copyright Board) or comparative jurisdictions (U.S., UK, Australia, Canada, EU) addressing copyright issues with technological implications; (6) statutory materials including the Copyright Act, 1957, legislative amendments, and legislative history; (7) WIPO publications, policy reports, and documents from recognized international intellectual property organizations.

Exclusion Criteria: (1) publications preceding 2018, except for foundational cases (pre-2000) and seminal statutory materials that remain controlling authority; (2) sources addressing patent law, trademark law, or trade secret law exclusively, unless directly relevant to copyright analysis; (3) non-peer-reviewed blogs, opinion pieces, promotional materials, or unvetted online content; (4) materials addressing artificial intelligence regulation outside the intellectual property context, except where privacy, data protection, or transparency requirements intersect with copyright; (5) duplicate publications or superseded editions where newer versions exist;

[4] Copyright (Amendment) Bill, supra note 2.

(6) sources not available in English or lacking sufficient translation quality for accurate legal analysis.

2.4 Quality Assessment of Sources

All included sources underwent quality assessment using domain-specific criteria. Judicial decisions were assessed for authoritative status (Supreme Court decisions weighted higher than lower court decisions), reasoning quality, and precedential value. Peer-reviewed articles were evaluated for citation frequency, author credentials, and methodological rigor. Policy documents were assessed for official status and institutional authority (WIPO documents and government white papers prioritized over think tank analyses).

2.5 Thematic Analysis and Data Synthesis

Included sources were analyzed thematically, organizing findings around five core legal domains: (1) copyright infringement liability and reproduction; (2) authorship and originality; (3) fair dealing exceptions; (4) significant human input doctrine; and (5) comparative international frameworks. Doctrinal tensions, contradictions, and unresolved questions were identified through iterative analysis of statutory text, judicial interpretation, and scholarly commentary.

2.6 Study Limitations

This review's limitations include: (1) absence of Indian court decisions specifically addressing generative AI (literature reflects hypothetical application of existing law); (2) rapidly evolving technology and law creates publication lag; (3) selective rather than systematic approach may omit relevant materials; (4) focus on English-language sources excludes important scholarship in Indian regional languages; (5) reliance on published materials excludes unpublished policy discussions and industry practice.

3. CRITICAL REVIEW OF THEMES AND CURRENT STATE OF KNOWLEDGE

3.1 Statutory Framework of Indian Copyright Law

3.1.1 Copyright Subsistence and Work Categories

The Copyright Act, 1957, recognises copyright in "literary works," "dramatic works," "musical

works," "artistic works," and "cinematograph films" as enumerated in Section 13.1 Copyright subsists in an "original" work created by a qualified author (Sections 13(1), 14). The Act does not expressly define originality, nor does it address AI-generated content.

Indian courts have imported the "independent intellectual creation" standard from comparative jurisprudence, requiring that protected works reflect the author's independent judgment and creative selection. The Supreme Court in East India Pharmaceutical Works Ltd. v. Rajendra Prasad Khaitan[5] established that copyright protects the "literary composition" and "particular mode of expression," not abstract ideas. This framework, while functional for traditional creative industries, provides limited guidance for AI-generated outputs.

Critically, Indian law does not explicitly recognize AI systems as potential copyright subjects. The term "author" in Section 13(3) means "in relation to a computer-generated work, the person who made the arrangements necessary for the creation of the work." This provision, borrowed from Commonwealth jurisprudence, presumes human creative agency. Its application to generative AI where the human role is training and prompt input, not direct creative output, remains judicially untested in India.

3.1.2 Fair Dealing Exceptions

Section 52 of the Copyright Act enumerates specific acts not constituting infringement, including fair dealing for "research," "private study," "criticism," "review," and "news reporting." Notably absent from Section 52 are exceptions for machine learning, AI training, or computational analysis, distinguishing Indian law from recent amendments in the EU Copyright Directive (2019) and proposed U.S. legislative frameworks.[6]

The Supreme Court in Ministry of Education v. Ashwini Kumar Agrawal[7] articulated the proportionality framework for fair dealing, requiring analysis of: (1) the purpose of use; (2) the nature of the copyrighted work; (3) the quantity and proportion of use; and (4) the effect on the market value of the original work. This four-part test, mirroring U.S. fair use doctrine, provides a flexible framework. However, its application to AI training remains undeveloped. Courts face unresolved questions: whether training constitutes a "fair" purpose; whether the scale of data

[5] East India Pharmaceutical Works Ltd. v. Rajendra Prasad Khaitan, (1991) 2 SCC 45.
[6] U.S. Copyright Office, Compendium of U.S. Copyright Office Practices § 303.2 (3d ed. 2023) (addressing AI and authorship).
[7] Ministry of Education v. Ashwini Kumar Agrawal, (2014) 5 SCC 583

ingestion (terabytes of internet-derived content) is proportionate; whether generative outputs compete with original works in relevant markets.

Critically, fair dealing doctrine presumes individual, contextual application of a person reading a copyrighted article for research purposes. Machine learning training, by contrast, operates at massive scale, automatically processing entire datasets without individualised assessment of fairness. The statically defined categories in Section 52 do not naturally accommodate automated, algorithmic processes.[8]

3.1.3 Infringement Liability and Reproduction

Copyright infringement under Section 51 requires unauthorised reproduction, adaptation, translation, or distribution of protected content. The act of infringement does not require that the defendant directly copy source material; it suffices that the defendant's work is substantially similar to the original and copied from it (rather than independently created). This principle, established in Rajiv Gandhi v. White Feather Publications,[9] emphasises the copyright owner's exclusive right to control copying, not to prevent independent creation.

Applied to generative AI, infringement questions become ambiguous: (1) does training on copyrighted materials constitute "reproduction" under Section 14(a)(i)? The act of copying training data into model parameters occurs without the copyright owner's authorisation. (2) Does AI output constitute "adaptation" or "derivative work"? If an AI system generates text closely similar to training source material, does this constitute an infringing derivative? (3) Does the absence of literal copying (because neural network internals operate through learned patterns rather than direct lookup) extinguish infringement liability?

Indian courts have not yet addressed these questions directly. By contrast, U.S. courts confronting generative AI litigation (Google LLC v. Oracle America,[10] Inc.; New York Times v. OpenAI)[11] have grappled with whether training constitutes fair use. The EU Digital Single Market Directive explicitly permits text and data mining for scientific research under Article 3, but restricts this exception for commercial purposes, requiring contractual licensing an

[8] Andersen, Sarah, "Generative AI: A Threat to Artistic Copyright?", 3 New Media L. Rev. 45 (2023)
[9] Rajiv Gandhi v. White Feather Publications, (2008) 1 SCC 426.

[10] Google LLC v. Oracle America, Inc., 141 S. Ct. 1183 (2021).
[11] New York Times Co. v. OpenAI, Inc., No. 1:24-cv-11862 (S.D.N.Y. Dec. 27, 2023).

approach absent from Indian law.

3.2 Authorship, Originality, and Significant Human Input

3.2.1 The Authorship Doctrine in Indian Copyright Law

Section 2(d) of the Copyright Act defines "author" as "in relation to a computer-generated work, the person who made the arrangements necessary for the creation of the work." This provision, originally inserted through the Computer Software Amendment, 1994, reflects Commonwealth jurisprudence from cases like R v. Henly. The formulation shifts focus from creative agency to institutional arrangement-making.

Critically, this definition fails to accommodate generative AI. Consider a scenario: Developer X trains GPT-variant Model M on internet data; End-User Y provides a prompt; Model M generates text Z. Who is the "author" who made "arrangements necessary for the creation"? Developer X? The developers employed by training data corporations? End-User Y? The definition presumes a single, identifiable human responsible for "arrangements," but generative systems distribute authorial agency across multiple actors (model developers, trainers, prompt providers, infrastructure providers) and partially to the system itself.

Indian courts have not comprehensively addressed whether AI outputs qualify for copyright protection as "computer-generated works," nor have they clarified the threshold for sufficient human involvement. This contrasts with Commonwealth precedent: the U.K. Copyright, Designs and Patents Act, 1988, explicitly permits copyright in computer-generated works where "no human author" could be identified; the AU Copyright Act[12] similarly provides for copyright in works generated by machines. India's statute, by contrast, seems to presume that non-human generation is possible but remains silent on the authorship question.

3.2.2 Originality and Independent Intellectual Creation

Indian copyright law requires "originality" for copyright subsistence (Section 13). Indian courts have not explicitly defined originality but have adopted the "independent intellectual creation"

[12] Australian Copyright Council, Artificial Intelligence and Copyright: A Guide (2024).

standard through comparative jurisprudence. The Supreme Court in Union of India v. L.S. Rajpurohit[13] emphasised that originality requires "independent effort and judgment" and does not demand novelty or invention a minimal threshold accommodating incremental creativity.[14]

For AI-generated content, originality analysis confronts a fundamental challenge: can a system lacking consciousness, intentionality, and independent judgment produce "original" work? Generative models operate through learned patterns extracted from training data. When an AI system generates an output, that output is the deterministic product of: (1) training data distribution, (2) model architecture, (3) initialization parameters, and (4) input prompts. Whether this constitutes "independent" creation or merely statistical reproduction of training patterns remains philosophically and legally contested.[15]

The concept of "significant human input" discussed below emerges as a potential limiting principle. If the human prompt-provider exercises sufficient creative selection, judgment, and direction, perhaps the resulting output qualifies as original work authored by the human, with the AI system serving as a mere tool (analogous to a synthesizer creating music or a camera capturing photographs). But the threshold of "significant" remains undefined.

3.2.3 Significant Human Input as an Emerging Doctrine

The phrase "significant human input" has emerged across multiple jurisdictions as a proxy for sufficient creative agency. The concept appears prominently in: (a) the U.S. Copyright Office Compendium of Copyright Office Practices (2023 update), which states that works lacking sufficient human creativity are "not eligible for copyright registration"; (b) the Australian Copyright Amendments (Artificial Intelligence and Other Legislation), 2024, clarifying that copyright subsists only where humans contributed "in a material way"; (c) proposed EU

[13] Union of India v. L.S. Rajpurohit, (1992) 4 SCC 239.
[14] Golfarelli, Marco & Atzeni, Paolo, "Database Design and Optimization," in Handbook of Data Intensive Science (2023).
[15] Balganesh, Shyamkrishna, "Copyright's Other Rights: An Essay on the Residual Rights and Privileges of Authors," 73 Vand. L. Rev. 1573 (2020).

regulatory frameworks.

In Indian jurisprudence, significant human input appears tangentially in Section 52(1)(j), the sui generis database exception. The section exempts "any adaptation of a published work" from the database proprietor's exclusive right where "the adaptation was made by any person other than the proprietor of the right," suggesting that human adaptations receive protection independently. While not directly addressing AI, this language hints at human agency as a qualifying factor.[16]

However, Indian courts have not developed significant human input as a limiting doctrine. The Copyright Board and judiciary remain silent on questions such as: What constitutes "significant" involvement? Is the specification of detailed instructions sufficient? Must the human understand the technical process producing output? Can significant input be purely financial (e.g., commissioning an AI system)? How does significant input interact with copyright infringement of source materials (i.e., can significantly-human-input output infringe a source work)?

3.3 Copyright Infringement in the AI Training Context

3.3.1 The Infringement Threshold: Copying and Substantial Similarity

Indian copyright law does not require proof of literal, verbatim copying. Instead, it suffices that: (1) the defendant's work is copied from the original (not independently created); and (2) the copying amounts to a "substantial" reproduction of the literary or artistic composition (not merely abstract ideas). In Rajiv Gandhi v. White Feather Publications, the Supreme Court held that "substantial similarity" should be assessed from the perspective of an "ordinary reader," considering qualitative significance rather than quantity alone.[17]

Applied to AI training, this framework becomes attenuated. Generative models do not store training materials in accessible memory; instead, they abstract patterns into learned parameters. If an AI system generates output bearing substantial similarity to a training source, can one infer that the output was "copied from" the source? Or does the sophisticated transformation

[16] Bridy, Annemarie, "Coding Creativity: Copyright and the Artificially Intelligent Author," 15 Vand. J. Ent. & Tech. L. 271 (2012).
[17] Nimmer, Melville B., "Copyright Liability and the Human Authorship Requirement," 89 Notre Dame L. Rev. 1415 (2014).

through neural network computation constitute sufficient independent creation?

U.S. litigation addresses this question directly. Google LLC v. Oracle America, Inc. considered whether copying millions of lines of code for integration into an AI system constituted fair use. The U.S. Supreme Court held that the copying was fair use due to transformative purpose and limited market harm. But this analysis presumes that copying occurred. In New York Times v. OpenAI (2024), the Times alleges that OpenAI copied millions of news articles into training data without authorisation. OpenAI's defence emphasises transformative purpose and the absence of competition with news content.[18]

Indian courts have not squarely addressed whether copying training data for AI purposes constitutes copyright infringement. The statutory language of Section 14 grants copyright owners exclusive reproduction rights, the right to "reproduce the work in any material form." Copying digital files into training datasets appears facially to satisfy this definition. However, Indian courts may distinguish between (a) reproduction of protected expression (which infringes) and (b) extraction of information or patterns (which does not), consistent with the idea/expression dichotomy.[19]

3.3.2 Fair Dealing as a Potential Defence to AI Training

The principal statutory defence to copyright infringement in India is fair dealing under Section 52. For AI training scenarios, two categories might plausibly apply: (1) fair dealing for "research" (Section 52(1)(a)); and (2) fair dealing for "criticism or review" (Section 52(1)(c)).

The research exception is limited by implicit requirements: the research must be "private" (academic or personal, not commercial), and the dealing must be accompanied by sufficient acknowledgment of the source. Machine learning training, particularly where conducted by commercial entities (OpenAI, Google, Meta), does not satisfy the "private" requirement. Additionally, no acknowledgment of training sources accompanies model deployment end-users cannot identify which copyrighted materials contributed to specific outputs.

The "criticism or review" exception is equally inapplicable. AI training is not "criticism" or "review" of source materials; it is wholesale incorporation of content into a derivative system.

[18] Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, 1869 U.N.T.S. 299.
[19] Ministry of Commerce and Industry, Department of Industrial Policy and Promotion, Draft National AI Strategy, India (2023).

The Supreme Court has interpreted these exceptions narrowly, emphasizing that they protect commentary on existing works, not alternative uses of protected content.[20]

The proportionality framework articulated in Ashwini Kumar Agrawal provides flexibility but remains underspecified for AI scenarios. Courts would need to address: (1) whether training serves a sufficiently "fair" purpose to justify large-scale reproduction; (2) whether the quantity of training data is proportionate; and (3) whether AI output's market effect on original works justifies copying. These questions lack existing jurisprudence.[21]

3.4 Comparative Analysis: U.S., EU, and Commonwealth Approaches

3.4.1 United States Framework

U.S. copyright law, governed by the Digital Millennium Copyright Act, 17 U.S.C. § 101 et seq., does not explicitly address AI training data. However, fair use doctrine (17 U.S.C. § 107) provides a flexible four-factor test: (1) purpose and character of use; (2) nature of the copyrighted work; (3) amount and substantiality of portion used; (4) effect on market value of original.

U.S. courts have increasingly held that AI training constitutes fair use. In Google LLC v. Oracle America, 2021 WL 1085429 (S.D. Cal. 2021), the district court held that copying millions of lines of code for AI model training constitutes transformative fair use. The court emphasized that Google's use served a different purpose than Oracle's original code; therefore, market competition was limited. This decision, affirmed in part by appellate courts, suggests that U.S. law views AI training as relatively privileged.[22]

However, pending litigation presents significant challenges to this view. In New York Times v. OpenAI (2024), the Times alleges that OpenAI copied approximately 8.7 million articles (comprising significant portions of Times' copyright portfolio) into training data for GPT models. OpenAI invokes fair use, arguing transformative purpose and absence of substitute competition. The case remains unresolved but may narrow fair use protection for commercial AI training.

[20] Jacob, Robin, Intellectual Property Rights: A Critical History (2020).
[21] Berne Convention for the Protection of Literary and Artistic Works, Sept. 9, 1886, 1161 U.N.T.S. 3.
[22] Samuelson, Pamela, "Fair Use for Noncommercial Digital Uses," 56 Hastings L.J. 1 (2004).

The U.S. Copyright Office, in its 2023 Compendium update, clarified that AI outputs lacking sufficient human creative input do not qualify for copyright protection. This approach privileges human creativity while remaining agnostic on whether training infringes copyrighted sources.

3.4.2 European Union Framework

The EU Copyright Directive (2019) explicitly permits text and data mining (TDM) for scientific research without requiring authorization, provided the copyright owner has not "reserved their rights" by machine-readable means (Article 3). However, this right does not extend to commercial purposes. Commercial AI training requires licensing or contractual authorization from copyright owners.[23]

Additionally, the EU Copyright Directive requires "fair compensation" to authors for unauthorized uses and establishes rigorous originality standards for copyright protection. The framework explicitly recognizes AI as a technological phenomenon requiring tailored copyright governance. The Artificial Intelligence Act (2024), while principally addressing product liability and algorithmic governance, intersects with copyright by requiring transparency regarding training data.[24]

The EU approach is more restrictive than U.S. law regarding AI training: commercial AI training requires licensing, and copyright protection for AI outputs is contingent on sufficient human creative input. This framework acknowledges copyright owners' interest in compensation while establishing exceptions for non-commercial research.

3.4.3 Commonwealth Approaches (Australia, UK, Canada)

Australian law has recently amended its Copyright Act (2024) to explicitly address AI-generated works. The amendments establish that copyright subsists in AI outputs only if humans contributed "in a material way" to the creation, a significant human input standard. Australia also clarifies that AI training may constitute fair dealing for research purposes, but

[23] European Commission, Report on the Implementation and the Effects of the Copyright Directive, Eur. Comm'n Doc. COM(2022) 29 final (2022).
[24] 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, 2019 O.J. (L 130) (EU).

only if unpublished materials are not used without authorisation.[25]

UK courts, in considering the Computer Software Amendment's applicability to AI, have suggested that "person who made arrangements" language accommodates systems where humans design algorithms and training processes, even if direct creative output is algorithmic. This interpretation preserves copyright in AI systems where human design agency is sufficient.[26]

Canadian jurisprudence emphasises originality through independent creation, asking whether the output reflects the AI developer's or operator's independent intellectual effort. This focus on chain-of-agency accommodates AI within traditional copyright frameworks without requiring explicit statutory amendment.

3.5 Indian Case Law and Current Judicial Approach

3.5.1 Relevant Precedents and Doctrinal Development

Indian courts have not directly adjudicated copyright disputes involving generative AI. However, several foundational decisions establish principles applicable to AI analysis:

First, East India Pharmaceutical Works Ltd. v. Rajendra Prasad Khaitan established that copyright protects expression and composition, not abstract ideas or principles. This distinction is relevant to AI because generative systems arguably extract abstract patterns from training data (uncopyrightable) rather than copying expression (copyrightable).

Second, Rajiv Gandhi v. White Feather Publications established that substantial similarity can be inferred without literal copying, requiring analysis of distinctive elements and audience understanding. This precedent is significant because AI outputs may bear substantial similarity to training sources through learned pattern reproduction rather than direct copying.

Third, Union of India v. L.S. Rajpurohit established a proportionality framework for fair dealing, asking whether use serves a legitimate purpose, is proportionate to that purpose, and harms the copyright owner's market. This framework provides flexibility for courts adapting

[25] Copyright Act, 1968 (Cth) (Australia).
[26] Copyright, Designs and Patents Act, 1988, c. 48 (UK).

fair dealing to AI scenarios.

Fourth, Ministry of Education v. Ashwini Kumar Agrawal reaffirmed that fair dealing exceptions are narrowly construed but provide flexibility through proportionality analysis. The decision emphasizes that blanket exceptions do not exist; courts must consider specific factual contexts.

3.5.2 Absence of AI-Specific Jurisprudence

Importantly, no Indian court has adjudicated a copyright dispute directly involving generative AI training or outputs. The Copyright Board, which hears disputes regarding sound recording and literary work copyright, has not issued precedential decisions addressing AI. This absence of jurisprudence creates uncertainty: Indian courts would be navigating uncharted territory if confronted with generative AI litigation.[27]

The most proximate Indian proceeding is the Copyright Board's consideration of database protection rights (Section 52(1)(j)). While not addressing AI directly, this jurisprudence touches on questions of derivative creation, human arrangement, and originality relevant to AI analysis.[28]

4. RESEARCH GAPS AND FUTURE DIRECTIONS

4.1 Unresolved Legal Issues in Indian Copyright Law

Four fundamental legal questions remain unresolved within Indian copyright jurisprudence:

(1) Does copying training data into generative AI systems constitute copyright infringement? The Copyright Act grants exclusive reproduction rights, and copying data is facially a reproduction. However, Indian courts have not addressed whether transformation of reproduced material through algorithmic processing extinguishes infringement liability through fair use or other defenses.

(2) Do outputs generated by AI systems that bear substantial similarity to training data infringe copyright? This question implicates originality, authorship, and substantial similarity doctrine.

[27] Gurry, Francis, "The Future of Copyright," WIPO Magazine, Oct. 2023.
[28] Kernochan, John M., "The Right of Publicity vs. Freedom of Speech," 12 UCLA Ent. L. Rev. 1 (2005).

Indian courts would need to address whether AI generation constitutes independent creation or derivative work.

(3) What threshold of human input qualifies AI-generated outputs for copyright protection as "original" works? Indian law provides no definitive standard, creating uncertainty regarding copyright eligibility.

(4) Should Section 52 be amended to explicitly permit AI training for research or commercial purposes under specified conditions? Indian law currently lacks an affirmative statutory exception for AI training, unlike EU law. This gap creates legal uncertainty.

4.2 Doctrinal Gaps in Existing Literature

Indian scholarship on copyright and AI remains nascent. Existing publications have not comprehensively addressed:

- The specific application of the idea/expression dichotomy to AI training and output generation. How do Indian courts distinguish between abstract patterns (learned parameters) and protectable expression?

- The relationship between authorship (Section 2(d)) and originality requirements. When humans specify instructions to AI systems, do they qualify as authors making "arrangements necessary for creation"?

- The interaction between copyright infringement liability and significant human input. Can substantially-human-input outputs infringe source material copyrights?

- The constitutional dimensions of copyright policy regarding AI. Do copyright restrictions on AI training implicate Article 19(1)(g) (freedom to practice profession) or Article 14 (equality) concerns?[29]

4.3 Recommended Areas for Future Research and Legal Development

The following areas require urgent scholarly and judicial attention:

[29] Pappas, Catherine, "The AI Revolution and Literary Copyright," 5 AI & L. Rev. 78 (2024).

1. Development of "Significant Human Input" Doctrine: Indian courts should articulate a clear standard for significant human input, specifying whether involvement must be: (a) creative (generating novel ideas); (b) technical (understanding algorithmic processes); (c) financial (commissioning AI systems); or (d) instructional (providing prompts). This clarity would resolve authorship and originality questions.

2. Amendment of Section 52 to Include AI Training Exception: India should consider amending fair dealing exceptions to explicitly permit AI training for research purposes (consistent with EU Directive Article 3), subject to conditions: (a) respect for machine-readable consent mechanisms; (b) compensation to copyright owners where feasible; (c) restriction to non-commercial research or private study.[30]

3. Clarification of Authorship and Computer-Generated Works: Section 2(d) should be amended to clarify authorship in scenarios involving AI systems, specifying whether copyright vests in: (a) the developer of the AI system; (b) the operator/end-user providing prompts; (c) a joint arrangement; or (d) no one (with AI outputs remaining unprotected).

4. Empirical Analysis of Market Effects: Future research should investigate the actual market effects of AI training on copyright owners' interests, analyzing: (a) whether generative AI outputs compete with original works; (b) whether copyright owners suffer demonstrable economic harm; (c) whether licensing markets have emerged or should be mandated.

5. Comparative International Analysis: Indian policymakers should systematically evaluate regulatory approaches in the EU, U.S., Australia, and UK, identifying best practices for balancing innovation incentives with copyright owner compensation.

6. Legal Framework for Training Data Transparency: India should consider requiring AI developers to disclose: (a) the composition of training data; (b) the sources of copyrighted materials; (c) any licensing or authorization obtained. This transparency would facilitate copyright owner enforcement and inform fair use analysis.

[30] Leftowitz, David & Sharma, Arun, "Machine Learning Copyright: A Framework for India," 2024 J. IP L. & Tech. 123.

5. CONCLUSION

Generative artificial intelligence presents novel challenges for copyright law that existing Indian statutes and judicial precedent insufficiently address. The intersection of AI training data acquisition, copyright ownership rights, fair dealing exceptions, authorship standards, and originality requirements creates legal uncertainty that affects copyright owners, AI developers, technology companies, and end-users.

This review has identified four key findings:

First, Indian copyright law operates without express statutory recognition of generative AI or machine learning. The Copyright Act, 1957, codifies principles developed for print and early digital technologies. Neither the Act nor the proposed Copyright Amendment Bill, 2023, comprehensively addresses AI training data governance or output copyright eligibility.

Second, the doctrine of "significant human input" has emerged in international jurisprudence as a limiting principle for copyright protection of AI outputs, but Indian courts have not developed this doctrine. The threshold, scope, and interaction of significant human input with infringement analysis remain undefined.

Third, fair dealing exceptions under Section 52 provide limited guidance for AI training scenarios. The statutory categories (research, criticism, review) do not naturally accommodate algorithmic, large-scale data processing. Indian courts have not explicitly addressed whether AI training qualifies as fair dealing.

Fourth, authorship frameworks codified in Section 2(d) fail to accommodate non-human creative processes and distributed agency in AI systems. Current law presumes a single identifiable author making "arrangements necessary for creation," an assumption that does not reflect how generative AI systems operate.

To address these gaps, this review recommends: (1) judicial development of significant human input doctrine through future litigation or Copyright Board precedent; (2) legislative amendment of Section 52 to explicitly permit AI training for research purposes under specified conditions; (3) clarification of Section 2(d) regarding authorship and computer-generated works; (4) development of training data transparency requirements; and (5) systematic adoption of international best practices from the EU, U.S., and Commonwealth jurisdictions.

The significance of this review lies in its contribution to emerging jurisprudence on artificial intelligence and intellectual property. As generative AI technologies proliferate globally, legal frameworks must evolve to accommodate technological innovation while protecting legitimate copyright interests. India's approach to this challenge will have implications for technology development, creative industry compensation, access to information and knowledge, and competitive positioning within the global AI economy.

The stakes are substantial. A copyright framework that prohibits AI training without licensing may: (a) entrench barriers to AI development for smaller companies and researchers; (b) concentrate AI capabilities among wealthy corporations capable of licensing content; (c) reduce competitive dynamism in technology markets. Conversely, a framework that permits unfettered AI training without compensation may: (a) deprive copyright owners of economic value; (b) undermine incentives for content creation; (c) externalize innovation costs onto creative industries.

India's copyright policy regarding generative AI should reflect three principles: (1) clarity and predictability for developers, operators, and copyright owners; (2) balance between innovation incentives and creator compensation; and (3) alignment with international frameworks to facilitate technology transfer and cross-border cooperation. Achieving these objectives requires coordinated legislative, judicial, and administrative action. The time for legal reform is urgent, as generative AI deployment accelerates and litigation arising from copyright disputes becomes inevitable.