THE INTERSECTION OF MACHINE LEARNING AND INDIAN COPYRIGHT JURISPRUDENCE: THE CASE FOR A DEDICATED ARTIFICIAL INTELLIGENCE CONTENT STATUTE

by

Aditi. N. V

&

Debasree Ghosh

ABSTRACT

The exponential rise of generative artificial intelligence (AI) has profoundly disrupted established legal doctrines surrounding intellectual property, posing complex challenges for global copyright regimes. In the Indian context, the interaction between generative models and the Copyright Act of 1957 has introduced deep regulatory uncertainties, particularly regarding uncompensated web-scraping for algorithmic training and the ambiguous legal standing of machine-authored content. Although India has built a robust framework for digital governance and technological scaling, its current statutory protections remain heavily skewed toward network security and data privacy, leaving intellectual property safeguards severely neglected. This paper provides a granular evaluation of how AI models strain existing Indian copyright laws and explains why iterative, piecemeal statutory revisions are fundamentally insufficient. To address these systemic gaps, this research advocates for a standalone legislative enactment specifically engineered to handle content regulations at the intersection of AI and copyright. By introducing mandatory dataset logging, creating a distinct legal tier for synthetic works, and optimizing automated text and data mining (TDM) licensing models, this proposed framework balances the rights of human creators with the societal need for computational innovation.

Indian Journal of Contemporary Legal and Social Issues

1. INTRODUCTION

1.1 The AI Revolution in the Indian Socio-Technical Landscape

The structural transformation of artificial intelligence from an academic research domain into commercially viable, ubiquitous enterprise systems marks a turning point in the generation and consumption of digital material. Within contemporary India, machine learning architectures are being actively deployed across high-stakes sectors. These deployments include parsing unstructured public health records during major health crises and using predictive regressions to model atmospheric fluctuations, monsoon variations, and agricultural production trends. Advanced neural pipelines are also executing complex linguistic operations, such as cross-lingual speech synthesis, language-agnostic speaker diarisation, and deep machine translation to bridge regional communication gaps. Concurrently, national digital learning platforms and Massive Open Online Courses (MOOCs) are embedding generative models to synthesize localized, adaptive educational materials tailored to India's vast student population.

To operate effectively, these computational models depend entirely on the massive collection and algorithmic assimilation of immense historical data pools. A significant portion of this ingested text, imagery, and audio comprises intellectual property protected under the Indian Copyright Act of 1957. Consequently, the unchecked growth of these systems has triggered sharp ethical and legal friction regarding the boundary between fair technological utility and systemic, automated intellectual property infringement. As machine learning models advance, their dependence on premium, human-authored cultural data increases, creating an inevitable conflict between technology enterprises and the creative sector.

1.2 Operational Mechanics vs. Traditional Legal Doctrines

Indian Journal of Contemporary Legal and Social Issues

The fundamental friction in the existing legal system stems from the incompatibility of generative AI workflows with historical legal doctrines in India. During the data ingestion phase, modern foundation models process vast data repositories containing billions of distinct variables to tune weights and build statistical probability matrices. To source this volume of data, software developers routinely utilize automated web scrapers that duplicate, cache, and analyze protected works without obtaining permissions or distributing royalties to the original human authors.

Indian jurisprudence, explicitly tied to the "modicum of creativity" standard articulated by the Supreme Court in landmark cases such as Eastern Book Company v. D.B. Modak, requires an

undeniable showing of human intellectual effort, skill, and judgment for an output to secure copyright protection. This doctrine creates an immediate legal vacuum when applied to purely automated, machine-generated outputs that lack human execution.

Furthermore, the statutory fair dealing exemptions listed under Section 52 of the Copyright Act of 1957 were historically crafted to shield conventional, human-scale activities such as private research, literary critique, educational synthesis, or journalistic review. They do not accommodate or immunize the wholesale, industrial-scale text and data mining (TDM) routines used to train commercial neural networks. This structural misalignment leaves human artists exposed to automated exploitation, while simultaneously placing domestic AI firms in a position of legal vulnerability that threatens to restrict venture capital inflows and domestic technological self-reliance.

1.3 The Regulatory Gap in Contemporary Legislation

The digital statutes currently active in India are structurally unequipped to manage the multi-layered complexities of generative computing. First, the country has relied primarily on a vertical, ministry-directed approach to AI policy. While this flexible, sector-specific oversight encourages short-term innovation within narrow fields like financial technology or cellular networking, it leads to regulatory balkanization and fails to provide a uniform national standard for intellectual property protection across the broader digital landscape.

Second, major foundational digital statutes, including the Digital Personal Data Protection (DPDP) Act of 2023 and the Telecommunications Act of 2023, focus almost exclusively on processing personal information, managing individual consent, and preventing traditional cybersecurity threats. As a result, they leave a wide regulatory void regarding AI-specific operational liabilities, such as algorithmic plagiarism, unauthorized stylistic replication, and data-harvesting infringement. To fix this legislative gap, this paper argues for a paradigm shift in digital governance, proposing the comprehensive blueprint for a dedicated parliamentary enactment: the Artificial Intelligence Content and Copyright Harmonization Framework (AICCHF).

Indian Journal of Contemporary Legal and Social Issues

2. RELATED WORK

The interface between machine learning and copyright law can be organized into three interconnected research areas: sector-led governance frameworks, automated scraping systems, and advanced content synthesis tools.

2.1 Sector-Led AI Governance and Regulatory Architectures

India's current approach leans heavily on a vertical governance model, delegating policy formulation to individual industry watchdogs rather than enforcing an expansive, horizontal statute. Legal scholars note that while this agile method prevents industry stagnation, it introduces severe policy silos. Investigations into telecom governance emphasize that current digital standards, including CERT-In guidelines, focus heavily on network breaches and private data integrity. This focus leaves algorithmic intellectual property disputes—such as training models on stolen databases or automated content copying—entirely outside the jurisdiction of these cybersecurity rules. This paper argues that because copyright infringement crosses industrial boundaries, it requires an overarching, standalone horizontal statute that protects creative property uniformly across all sectors.

2.2 Automated Data Extraction and Machine Learning Applications

The technological engine of modern AI requires the harvesting of unstructured public datasets. This operational approach is common in public health tracking, where machine learning pipelines and automated PDF parsers are used to scrape, extract, and compile messy regional health updates during pandemic emergencies. Parallel regression models are used to ingest huge volumes of historical meteorological data to analyze climate change vectors across the subcontinent. While these high-throughput data aggregation tools provide significant public benefits, they present a major legal flaw: the underlying data ingestion engines do not distinguish between public domain data and copyrighted materials, drawing both into training caches. Building on these operational realities, this study emphasizes the necessity of statutory TDM protocols that separate fair public data extraction from actionable intellectual property theft.

2.3 Advanced AI Content Generation and Audio Processing

The deployment of deep learning models for human expression has accelerated rapidly. Recent research highlights the creation of unified multilingual audio architectures that can execute language-independent speaker identification, complex diarisation, and deep translation. Similarly,

Indian Journal of Contemporary Legal and Social Issues

generative systems are being deployed to power interactive online courses, localizing educational resources to address regional needs. The clear advantage of these platforms is their ability to democratize access to education across linguistic barriers. However, they create intense friction regarding performer rights and audio copyright; training neural models on human vocal signatures to achieve seamless translation can infringe upon the moral and economic protections of individual speakers and teachers. This demonstrates the inadequacy of legacy copyright frameworks when evaluating synthetic audio and automated educational outputs, reinforcing the need for a dedicated statute.

3. THE STRUCTURAL LIMITATIONS OF THE COPYRIGHT ACT, 1957

The architecture of the Copyright Act of 1957 was designed around physical printing, analog broadcasts, and material media. While the legislative updates of 1994 and 2012 integrated computer programs, internet communications, and digital rights management (DRM) parameters, the statutory foundation remains deeply human-centric. Generative AI disrupts these core assumptions across its developmental lifecycle.

3.1 The Authorship Dilemma under Section 2(d)

Section 2(d) of the current Act defines an "author" based on the medium of the work, noting that for computer-generated content, authorship belongs to "the person who causes the work to be created". Historically, courts applied this provision to human programmers using software as a basic tool to execute their creative goals. Generative neural networks, however, operate with extreme independence. When a user inputs a single-sentence textual prompt and a model outputs a complex digital image or a multi-page document, assigning authorship to either the end-user or the software developer pushes Section 2(d) past its logical limits. The end-user did not make the creative decisions regarding individual pixels, shading, or syntax, while the developer could not predict or control the specific output produced by the network. This creates a legal gap where purely machine-driven outputs fail to secure protection, immediately entering the public domain and reducing commercial incentives to invest in localized creative AI technologies.

Indian Journal of Contemporary Legal and Social Issues

3.2 The Fair Dealing Conundrum under Section 52

Section 52 handles non-infringing uses through a strict, closed list of specific exemptions, differing from the flexible, open-ended four-factor fair use standard applied in the United States. India's exceptions are restricted to activities like private study, critical review, academic

summary, or news reporting. When technology companies ingest millions of copyrighted texts, illustrations, or articles to build commercial foundation models, these actions do not fit within the boundaries of Section 52. The ingestion cannot be classified as "private research" because the ultimate goal is a commercial, market-facing application, nor does it qualify as a critical review. Because India lacks an explicit, standardized exception for text and data mining like those found in Japan, Singapore, or the European Union, domestic AI developers face significant liability exposure from corporate publishing entities and entertainment conglomerates.

3.3 Ingestion Infringement vs. Output Substantial Similarity

The intellectual property conflict occurs across two distinct phases: ingestion infringement and output replication. Ingestion infringement takes place during the initial training process, where duplicates of protected works are written to RAM or localized caches to allow feature extraction. Even if the eventual output shows no clear resemblance to the source material, this unauthorized copying violates the exclusive reproduction rights protected under Section 14. Conversely, output infringement occurs when a model produces material that is substantially similar to an asset in its training pool due to overfitting or deliberate prompt manipulation. Existing judicial tests in India, such as the "lay observer" standard established in R.G. Anand v. Delux Films, are unequipped to evaluate algorithmic copying, stylistic imitation, or the high-dimensional latent structures of deep neural networks.

4. THE PROPOSED AICCHF FRAMEWORK

To systematically resolve the statutory deficits of the current regime, this paper proposes the Artificial Intelligence Content and Copyright Harmonization Framework (AICCHF). This framework is designed to serve as the blueprint for a comprehensive, standalone legislative enactment by the Indian Parliament. Rather than attempting to patch the Copyright Act through minor amendments, a standalone act creates an integrated, horizontal regulatory mechanism capable of addressing the multi-stage operational lifecycle of machine learning systems.

Indian Journal of Contemporary Legal and Social Issues

4.1 Data Provenance and TDM Micro-Licensing Module

The first structural component of the AICCHF mandates strict, verifiable data provenance records for all commercial AI entities operating in India. Developers must implement immutable metadata logging architectures that generate an audit trail of all training data. Modeled after data ingestion pipelines used in public health monitoring, the intake software must check every file for

digital rights management tags and author signatures. If a protected asset is encountered, the ingestion engine prevents the file from entering the training corpus until a micro-transaction is finalized through a National AI Copyright Registry. This registry operates an automated, fractional licensing infrastructure, allowing high-velocity clearing to compensate human creators at scale without requiring manual licensing negotiations.

4.2 Synthetic Works and Output Governance Module

The second component introduces a novel statutory category of intellectual property: Synthetic Works. Under this rule, completely autonomous, machine-driven outputs are denied human copyright status, preventing corporate entities from monopolizing automated content. To support industrial adoption, the framework establishes a "sui generis" protection class for Human-Assisted Synthetic Works. If a human operator proves significant creative intervention—verified via multi-stage prompt engineering, model conditioning, and extensive manual post-generation editing—they are awarded a limited, five-year copyright over the resulting work. This mechanism protects human creative effort while adapting to modern, digital workflows.

4.3 Federated IP Enforcement and National AI Copyright Tribunal

To ensure compliance across India's decentralized digital ecosystem, the AICCHF bypasses a centralized bureaucratic structure in favor of a Federated IP Enforcement Module. Using a distributed compliance model, enforcement duties are shared among existing sector-specific watchdogs. For instance, educational bodies managing AI learning platforms or telecommunications agencies overseeing network traffic act as the primary frontline observers. These distributed organizations track compliance within their sectors and forward aggregated, anonymous logging metadata to a newly formed National AI Copyright Tribunal. This specialized tribunal holds the technical and legal expertise required to handle rapid dispute mediation, mandate algorithmic audits, and orchestrate cross-border copyright enforcement.

Indian Journal of Contemporary Legal and Social Issues

5. METHODOLOGICAL EVALUATION PLAN

To test the regulatory viability and performance impact of the AICCHF, we present a structured evaluation plan using a controlled simulation environment. The goal is to measure the precise economic and operational shifts that occur when moving from an unregulated digital ecosystem to the structured compliance environment required by the AICCHF.

5.1 Benchmarking Dataset Structure

The evaluation relies on a multi-modal dataset consisting of 100,000 simulated Indian digital assets. To reflect the true characteristics of the domestic market, the testing environment incorporates regional academic texts from Western Himalayan universities, audio records from the entertainment sector, proprietary climate modeling datasets, and unstructured public health files. This mix ensures that the automated rights classification system is evaluated against diverse schemas, file formats, and licensing rules.

5.2 Evaluation Metrics and Core Indicators

The efficacy of the proposed statutory framework will be measured using three core operational indicators:

Core Indicator	**Operational Measurement**	**Objective**
Licensing Compliance Rate	The percentage of copyrighted assets within the 100,000-file testing pool that are flagged, sorted, and matched to a valid micro-license before ingestion.	Ensure high accuracy in identifying protected data.
Processing Latency	The exact computational delay and latency overhead added to the machine learning training pipeline by metadata checks and cryptographic watermarking.	Prevent the law from making model training technically unfeasible.
Dispute Resolution Efficiency	The projected drop in long-term copyright litigation resulting from the availability of immutable, cryptographically verified data provenance logs.	Streamline legal conflicts through clear digital evidence.

Indian Journal of Contemporary Legal and Social Issues

6. EXTENDED DISCUSSION AND POLICY ANALYSIS

Implementing a standalone AI content law based on the AICCHF will transform India's digital economy. For startups and laboratories developing advanced speech systems and machine translation tools, compliance will require a restructuring of data workflows. Engineering teams must move away from unrestricted web scraping and transition toward ethical curation models that integrate real-time digital rights verifications.

6.1 Systemic Limitations and Failure Modes

Despite its comprehensive design, the AICCHF features structural limitations that require careful legislative consideration. First, the platform introduces computational delay and administrative fees. While capitalized tech multinationals can easily cover these costs, small research labs, open-source initiatives, and independent developers may find the barriers prohibitive, risking a corporate monopoly over domestic foundational models. Second, internet architectures are inherently borderless. An AI model can be trained on external servers located in a foreign regulatory haven and deployed in India via cloud APIs, making local ingestion tracking complex. Third, if a protected work is mistakenly or maliciously integrated into a deep neural network, current computer science methods cannot easily remove that specific asset's mathematical influence without rebuilding the entire model from scratch, causing significant environmental and financial costs.

6.2 Ethical Risks: Cultural Bias and Economic Displacement

Strict licensing costs and litigation risks could force domestic developers to rely exclusively on Western open-source datasets or old, public-domain text corpuses. If this occurs, local models will suffer from severe cultural bias and temporal stagnation, failing to understand modern Indian dialects, regional idioms, and evolving cultural identities. Furthermore, there is a risk of economic displacement for human creative communities. Even if writers, musicians, and designers receive micro-royalties during the training phase, the capacity of generative models to produce an infinite volume of derivative material in the precise styles of those human creators could saturate the market, threatening the livelihoods of the creative workforce.

6.3 Future Research Avenues and International Treaties

Future legal-tech research should focus on two main areas. First, developers should explore integrating decentralized ledger architectures and automated smart contracts into the National AI

Copyright Registry. This would allow instant, peer-to-peer royalty payouts to content creators, reducing administrative friction and corporate gatekeeping. Second, Indian policymakers must focus on international regulatory alignment. Because data flows globally, India should lead conversations within the World Intellectual Property Organization (WIPO) to coordinate standalone AI regulations across borders, preventing data piracy hubs and ensuring uniform enforcement against external algorithmic entities.

7. CONCLUSION

The rapid growth of AI across India's digital landscape has driven advancements in public health tracking and multilingual education, yet it has exposed structural vulnerabilities within the nation's intellectual property framework. Ingestion routines for foundation models conflict with the protections granted to human creators under the Copyright Act of 1957, while the existing approach of using sector-specific guidelines and cybersecurity laws cannot resolve the challenges of synthetic authorship or automated copying. This policy gap threatens the livelihoods of creators and creates legal uncertainties that could slow the growth of India's domestic technology sector.

To build a sustainable ecosystem, this paper has argued for the creation of a standalone AI content regulation enactment. Through the blueprint of the AICCHF, we provide a method to enforce data provenance, structure text and data mining licensing, and define the parameters of machine-assisted authorship. While administrative overhead and jurisdictional challenges persist, the necessity for a dedicated legal framework remains clear. Balancing machine learning innovation with copyright protections will allow India to protect its creative workforce while maintaining its position as a global leader in digital technology.

Indian Journal of Contemporary
Legal and Social Issues

REFERENCES

Agarwal, M., Chakraborti, T., Grover, S., & Chaudhary, A. (2021). COVID-19 India Dataset: Parsing COVID-19 Data in Daily Health Bulletins from States in India. https://arxiv.org/pdf/2110.02311v2

Agarwal, A., & Nene, M. J. (2025). Incorporating AI incident reporting into telecommunications law and policy: Insights from India. Computer Law & Security Review, 60, 106263. https://doi.org/10.1016/j.clsr.2026.106263

Agarwal, A., & Nene, M. J. (2026). A federated architecture for sector-led AI governance: lessons from India. Transforming Government: People, Process and Policy. https://doi.org/10.1108/TG-09-2025-0310

Das, P. P. (2025). Perspective Chapter: MOOCs in India: Evolution, Innovation, Impact, and Roadmap. IntechOpen. https://doi.org/10.5772/intechopen.1001972

Raghav, N., Banerjee, A., Chakraborty, J., Gupta, A., Punyeshwarananda, S., & Sahidullah, M. (2025). The TCG CREST -- RKMVERI Submission for the NCIIPC Startup India AI Grand Challenge. https://arxiv.org/pdf/2512.11009v1

Sharma, D., & Singh, V. (2010). ICT in Universities of the Western Himalayan Region of India II: A Comparative SWOT Analysis. International Journal of Computer Science Issues. https://arxiv.org/pdf/1002.1193v1

Wamanse, R., & Patil, T. (2022). Analysis of various climate change parameters in India using machine learning. https://arxiv.org/pdf/2201.10123v1

Indian Journal of Contemporary
Legal and Social Issues