INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 5 AND ISSUE 8 OF 2025

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

DATA WITHOUT CONSENT: THE COPYRIGHT DILEMMA IN AI DEVELOPMENT

AUTHOR - MEGHNA NAIR, STUDENT OF LL.M - IP, AMITY UNIVERSITY, NOIDA

BEST CITATION - MEGHNA NAIR, DATA WITHOUT CONSENT: THE COPYRIGHT DILEMMA IN AI DEVELOPMENT, *INDIAN JOURNAL OF LEGAL REVIEW (IJLR),* 5 (8) OF 2025, PG. 856-863, APIS – 3920 – 0001 & ISSN - 2583-2344

I. ABSTRACT

This paper critically examines the role of data mining in the development of artificial intelligence (AI), especially in the context of copyright law. As AI systems increasingly rely on large-scale datasets, many comprising copyrighted works for training, the practice of text and data mining (TDM) has become a double-edged sword. On the one hand, it serves as a cornerstone of innovation, enabling machines to simulate human-like reasoning and generate sophisticated outputs. On the other, it raises serious legal and ethical concerns regarding the unauthorized use of protected intellectual property. The legal vacuum that exists in jurisdictions like India, and the ramifications for authors' economic and moral rights are explored along with the evolution and mechanics of data mining in AI development. It delves into critical jurisprudential debates, discussing real-world legal disputes such as the ANI v. OpenAI case to illustrate the urgent need for regulatory clarity. By analysing both the supportive and critical perspectives on data mining in AI, the necessity of a balanced framework, one that fosters innovation without undermining the foundational principles of copyright and authorship is pressed upon.

II. INTRODUCTION

Have you ever noticed that after searching for a product just once on Google, whether it's headphones, a coffee maker, or running shoes, your online ads immediately start reflecting that search? Suddenly, every website you visit seems to show suggestions for that exact item or something closely related. This isn’t a coincidence. Advertising algorithms are designed to track your search and browsing behaviour, and they adjust quickly, tailoring the content you see to your most recent interests and online activity. What’s even more striking is that these ads are constantly evolving, updating in real time to reflect your changing needs and preferences.

In today’s digital landscape, data has emerged as the most valuable currency in the world, arguably even surpassing traditional commodities like oil in influence and economic power. Initially, the rise of the internet and smart technologies promised better communication, enhanced accessibility, and more personalized experiences. However, what began as simple data collection, intended for improving services and analytics, rapidly expanded into a vast and largely hidden economy centred around the continuous harvesting and monetization of personal information. As digital platforms became integral to everyday life, facilitating everything from shopping and social networking to learning and working. Corporations, especially large tech conglomerates, began to recognize the immense value in tracking user behaviours, preferences, and habits. This data wasn't only used to personalize ads; it also became the backbone of predictive algorithms, artificial intelligence training, and strategic business decisions.

What makes this system particularly concerning is the often opaque and covert manner in which data collection takes place. While companies may claim they are personalizing services for user benefit, the reality is that vast quantities of sensitive data are being gathered, analysed, and frequently sold, often without the user’s informed consent. Lengthy, convoluted terms and conditions serve more to shield corporate interests than to genuinely inform users, and most people agree to them simply to access basic services. This isn't limited to private corporations anymore, governments, third-party advertisers, and data brokers are also active participants in this data-driven ecosystem. The result is a complex and largely invisible global network in which personal data is continuously exchanged, often without oversight. In many ways, this unchecked flow of information represents a serious threat to individual privacy in the modern age.

III. DATA MINING

In recent years, data mining has evolved into one of the most powerful tools used by technology companies to train artificial intelligence systems. While on the surface, this practice is often framed as a catalyst for technological advancement, much like how automobiles replaced horses in the name of progress, it also conceals a far more concerning reality.

Data mining broadly refers to the practice of analysing large datasets to uncover meaningful patterns, trends, or correlations that can be applied to inform future decisions.[950] This process has gained substantial traction in recent years due to technological advancements in data collection and storage, the proliferation of machine learning techniques, and the steady decline in computational costs.[951] At its core, machine learning relies heavily on data; it is the mechanism through which systems "learn" from past events to make informed predictions or generate new content.[952] Through statistical analysis, algorithms uncover insights embedded in vast quantities of data insights that are often inaccessible to traditional analytical methods.[953]

Machine learning has become the backbone of most contemporary AI systems, enabling machines to simulate human-like reasoning, learning, and decision-making.[954] A critical component of this process is the use of large-scale datasets, particularly text-based datasets in the domain of Natural Language Processing, to train AI models to understand and produce language-based outputs.[955] These applications range from automated translation and information extraction to chatbots and intelligent search systems.[956] Text and data mining (TDM) has emerged as a cornerstone of this process. TDM refers to the automated analysis of large corpora of text or other datasets to extract information and build intelligent systems. While the practice is crucial for AI advancement, it raises pressing legal questions, particularly surrounding the use of copyrighted material.

In essence, a typical AI development pipeline begins with identifying a problem; what the model is intended to predict or perform. Developers then curate and prepare the necessary data, often requiring extensive labelling or annotation.[957] This data is used to train a model using a selected algorithm, which outputs a machine learning model capable of interpreting or predicting based on new, unseen data.[958]

AI developers frequently utilize pre-trained models to reduce the need for massive training datasets, sometimes narrowing data requirements from millions of entries to just thousands.[959] Nevertheless, fine-tuning these

[950] **Foster Provost & Tom Fawcett**, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (O'Reilly Media 2013).

[951] ibid

[952] **Pedro Domingos**, *The Master Algorithm* (Basic Books 2015)

[953] ibid

[954] **Stuart Russell & Peter Norvig**, *Artificial Intelligence: A Modern Approach* (4th ed. Pearson 2020).

[955] **Jacob Devlin et al.**, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805 (2018), https://arxiv.org/abs/1810.04805.

[956] ibid

[957] **Aurélien Géron**, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2d ed. O'Reilly 2019).

[958] **European Commission**, *Guidelines on Text and Data Mining* (2020)

[959] **OpenAI**, *GPT Fine-Tuning Guide*, OpenAI Documentation (2023), https://platform.openai.com/docs/guides/fine-tuning.

models still necessitates downloading or sourcing new data. Since this step is typically automated, especially in fields like computer vision or NLP developers often scrape data from online platforms such as Google Images, YouTube, Wikipedia, or even Gmail.[960] This scraping, while efficient, presents significant copyright concerns when it involves the unauthorized use of protected content.

IV. DATA MINING VIS-À-VIS COPYRIGHT

The collection and utilization of vast amounts of data are not only central to AI development, but they also come at the cost of significant legal and ethical compromises. One of the most critical of these is the rampant infringement of copyright through indiscriminate text and data mining practices.

For AI to function at the level it does today, i.e. recognizing complex patterns, responding to nuanced prompts, and generating human-like content, it must be trained on enormous volumes of data. These datasets are compiled from a multitude of sources, including books, articles, images, videos, and even creative literary works, many of which are protected under copyright law. Unfortunately, this data is often scraped and extracted without the knowledge or consent of the original creators. Major AI companies deploy automated tools that harvest this information at scale, frequently ignoring the origin, ownership, or protected status of the content. In doing so, they blur the line between publicly available information and proprietary intellectual property.

This indiscriminate data mining is especially concerning in jurisdictions like India, where there currently exists no specific legal framework that addresses the use of copyrighted material in AI training. As a result, authors and creators find themselves at a severe disadvantage, with no practical means to seek redress or enforce their rights. The absence of a clear regulatory structure allows AI developers to exploit literary and artistic works without securing licenses or compensating rights holders. This undermines the exclusive rights granted to authors under copyright law particularly the right to reproduce, distribute, and communicate their work to the public. As AI systems continue to improve, they are increasingly capable of generating content that closely mirrors or replicates original works, effectively releasing these derivative creations into the public domain. This severely impacts the author's economic rights, as their ability to control and commercially benefit from their work is eroded.

Beyond financial concerns, such practices also strike at the heart of an author's moral rights. AI-generated outputs can sometimes misattribute works or associate false information with a particular creator. This misattribution not only distorts the intent and originality of the creator's work but also affects their reputation, a key concern protected by moral rights such as the right to attribution (paternity) and the right to preserve the integrity of one's work. For instance, if an AI model is trained on a copyrighted novel and later produces derivative content without proper attribution, or worse, attributes it incorrectly, the original author's personal and professional standing may suffer. These breaches go beyond technical violations; they represent a deeper moral and ethical failure to respect the individual creator's voice, identity, and legacy.

In essence, while data mining may be justified as a necessity for technological innovation, its unchecked and opaque implementation has led to a landscape where creators' rights are routinely trampled. This is especially problematic in developing nations where legal systems have yet to evolve to keep pace with the rapid expansion of AI technologies. Without stringent safeguards, creators remain vulnerable, their work quietly absorbed into training datasets and re-emerging in ways they cannot foresee, control, or benefit from. If left unregulated, this unchecked exploitation of creative labour will not only weaken the intellectual property framework but also

[960] **Abhishek Gupta**, *The Ethical Machine* (Routledge 2022).

threaten the very foundation of authorship, originality, and ethical innovation.

A. Flip side of the debate

One perspective in the ongoing debate around copyright and artificial intelligence asserts that data mining, particularly when used to train machine learning models, should not be viewed as a form of copyright infringement. Proponents of this view argue that the nature of data mining is fundamentally non-expressive and functional, thereby placing it outside the scope of copyright protection. They contend that a legal "safe harbour" for such activity is warranted, not only to enable the continued growth of artificial intelligence but also because, under current copyright jurisprudence, the act of mining data does not necessarily equate to infringement.

This reasoning finds roots in established copyright doctrine, notably in the U.S. Supreme Court decision in *Baker v. Selden*. In this seminal case, the Court clarified the distinction between a work's expressive content and its utilitarian aspects. It held that the mere use of a copyrighted work's material form for a non-expressive purpose, such as learning or extracting functional information does not, in itself, amount to copyright infringement.[961] This distinction is crucial in the context of AI training where large datasets comprising of text, images, or other expressive works are analysed not for their aesthetic or expressive value, but for their embedded patterns and statistical structures. When developers download copyrighted material to train AI models, they do so to enable machines to learn linguistic structures, visual classifications, and behavioural patterns, not to replicate or publicly disseminate the original works.[962] The use is technical, non-communicative, and wholly separate from any expressive or creative reuse. As such, these copies are not distributed, displayed, or performed publicly, which are key criteria for infringement under copyright law.[963]

Additionally, this line of argument relies on the idea-expression dichotomy central to copyright doctrine: copyright protects the original expression of ideas, not the ideas themselves. AI model training is understood to extract unprotectable ideas, patterns, or data structures from input material, rather than replicating expressive content. Thus, it is posited that even a fair use analysis might not be necessary for such acts, as they may fall outside the protective boundaries of copyright altogether.[964]

Under this reasoning, the need for a legislative or judicially recognized safe harbour becomes evident. As artificial intelligence continues to evolve, denying developers the ability to mine data, even from copyrighted sources would significantly hinder innovation and slow technological progress.[965] Supporters of this viewpoint therefore advocate for clearer legal provisions that distinguish between functional and expressive uses of copyrighted content in the realm of AI, ensuring that legitimate, transformative, and non-expressive data mining is not unjustly penalized.

B. Fair Use or Fair Dealing: a possible defence?

Thanks to the consumption of websites like YouTube, almost everyone is aware of the Section 52 of the Copyright Act, 1957, which provides specific exceptions to copyright infringement, allowing limited use of copyrighted material without the owner's authorization. The author wishes to explore into the idea whether this provision potentially defends the use of copyrighted content by AI tools, such as Large Language Models. Though the law does not explicitly define "fair dealing," it permits the use of literary, dramatic, musical, or artistic works for purposes like research, private

[961] *Baker v. Selden, 101 U.S. 99 (1879); see also* ***James Grimmelmann****, There's No Such Thing as a Computer-Authored Work — And It's a Good Thing, Too, 39 Colum. J.L. & Arts 403 (2016).*

[962] **M. Feldman**, *The Uncharted Territory of AI and Copyright*, Harv. J.L. & Tech. Dig. (2023).

[963] ibid

[964] **J. Gervais**, *AI and Copyright: Fair Use is Not the Answer*, 24 Vand. J. Ent. & Tech. L. 609 (2022).

[965] **European Parliament**, *Artificial Intelligence Act: Text and Data Mining Exceptions*, Legislative Briefing (2022).

study, criticism, or review, which may not constitute infringement.

In India, the test for copyright infringement focuses on two key questions: whether the content is "substantially similar" to the original copyrighted work and whether it falls under the exceptions listed in Section 52. According to this framework, literary works generated by AI systems, such as LLMs, may not be considered infringement if used for private consumption, research, review, or educational purposes, as these uses typically fall within the exceptions outlined in the law. However, the situation changes when AI-generated content is used for commercial purposes or in ways that do not meet the exceptions. For example, if a person uses an LLM to generate a translation of a copyrighted book or to create an alternate ending to a novel and then publishes it under their own name, this would infringe the original author's copyright, as it goes beyond the scope of fair dealing.

Determining what qualifies as fair dealing ultimately rests with the courts, which evaluate each case based on the specific facts, circumstances, and established guidelines. Courts assess whether the use of the copyrighted material is justified within the context of the exceptions and whether the original work has been transformed in such a way that it constitutes an infringement. This legal standard is particularly relevant in cases involving AI, as the lines between creative input and replication of copyrighted material can sometimes blur, leading to potential infringement claims.

V. CRITICISM

The legal dispute between Asian News International (ANI) and OpenAI, the creators of ChatGPT, marks a significant turning point in India's evolving approach to intellectual property law, particularly in the context of artificial intelligence. While jurisdictions such as the United States and the United Kingdom have already seen litigation concerning the intersection of generative AI and copyright, this case represents the first of its kind before the Indian judiciary, bringing to the forefront critical questions about data usage, copyright infringement, and technological advancement.

At the heart of the case is ANI's allegation that OpenAI has unlawfully utilized its vast archive of news reports, interviews, and exclusive statements, built over five decades for training its large language models and generating outputs in response to user queries. Under Section 14 of the Indian Copyright Act, 1957, ANI claims exclusive rights to reproduce, store, and distribute its original works. ANI asserts that OpenAI's actions using these materials without licensing or permission amount to copyright infringement at two distinct stages: (i) during the training phase of the AI model using copyrighted material, and (ii) through the generation and reproduction of those works as outputs during user interaction[966].

The Delhi High Court has issued notice to OpenAI and its local affiliate, Microsoft India, and has posed preliminary questions to be resolved, such as whether training a model on copyrighted data constitutes infringement, and whether the reproduction of similar content via AI-generated responses also violates copyright law[967].

One of the pivotal legal doctrines under scrutiny is the principle of "fair use" or "fair dealing," enshrined under Section 52 of the Copyright Act. This provision allows limited use of copyrighted materials for purposes such as private use, criticism, review, or the reporting of current affairs. However, applying this exemption to AI training is complex. ANI has contested the applicability of this doctrine to OpenAI, especially since OpenAI offers a commercial version of its tool - ChatGPT Plus, thus complicating claims that the use is purely for research or educational purposes[968]. This case also brings into focus the concept of Text and

[966] *ANI Media Pvt. Ltd. v. OpenAI OpCo LLC*, CS(COMM) 1028/2024 (Del. H.C.).
[967] *Delhi High Court Issues Notice to ANI and OpenAI in Copyright Infringement Matter*, Bar & Bench (Apr. 2024), https://www.barandbench.com.
[968] Ibid

Data Mining, a process that involves the automated extraction and analysis of data from vast datasets. While the UK allows TDM for non-commercial research under its copyright framework, India currently lacks any specific legal exception or regulation dedicated to TDM, making this case all the more consequential. The lack of clarity leaves content creators vulnerable, especially when their work is used without consent in high-stakes commercial contexts like AI model training[969].

OpenAI, in its defence, has argued that it has blocklisted ANI's domain (www.aninews.in) from future training datasets and that it provides an "opt-out" mechanism for content creators who do not wish for their work to be used. While this mirrors evolving practices in the European Union where rightsholders can prevent TDM through machine-readable signals. It raises further questions in India: should the burden to prevent infringement lie with the copyright holder or the entity performing data mining? Moreover, ANI has highlighted that even after opting out, data already stored on servers remains and cannot be deleted. This raises further legal and ethical questions about the indefinite storage of copyrighted material and whether such long-term retention constitutes a continuing infringement.

A. Opt-out is not the way out

The opt-out mechanism primarily, it shifts the burden of copyright protection onto the author, contradicting the principle that copyright is an automatic right and should not require affirmative action to be enforced under Section 14 of the Indian Copyright Act, 1957. Additionally, once an AI model is trained on copyrighted data, the knowledge cannot be "unlearned" even if a creator opts out later, the model retains the patterns and insights extracted during training[970]. Thus, the infringement has already occurred and cannot be reversed through opt-out procedures.

A more rights-preserving alternative would be an *opt-in model*, where content is included in training datasets only with explicit authorization. This approach aligns better with the principles of informed consent and safeguards the exclusive rights granted to authors[971]. Further complicating the matter, tools used for opting out, such as *robots.txt* or domain blocklisting, are technically voluntary, not legally binding, and rely on the good faith of AI developers[972]. Moreover, many smaller creators may not possess the technical know-how or awareness to implement these safeguards effectively. From a legal standpoint, post-infringement opt-outs offer no retroactive remedy. Courts will still consider whether content was accessed and utilized without prior consent, potentially causing irreparable harm, especially when it involves misattribution or economic displacement. The opt-out model is insufficient and inadequate. A robust framework for consent-based data usage, where rights are proactively respected, is necessary to balance technological advancement with authorial protection.

B. Unjust enrichment and market harm

In addition to copyright violations, ANI has also made claims under tort law, citing unjust enrichment and unfair competition. ANI contends that OpenAI's use of its content allows it to replicate ANI's reporting potentially even faster than ANI can publish its own updates thus siphoning away its readership and harming its core business model. This argument is particularly potent given the time, effort, and resources ANI invests in content creation.

Another critical dimension raised by ANI pertains to moral rights, especially the right to attribution and protection from false attribution. ANI notes instances where ChatGPT falsely attributed news to ANI that it never published. Such misattributions not only harm ANI's reputation but also raise concerns about the

[969] *Directive (EU) 2019/790 of the European Parliament and of the Council on Copyright and Related Rights in the Digital Single Market*, 2019 O.J. (L 130) 92.
[970] **Casey Newton**, *AI Can't Unlearn What It Knows*, Platformer (2023), https://www.platformer.news.
[971] **Daniel Gervais**, *Data and the Future of Copyright*, 45 Colum. J.L. & Arts 1 (2022)
[972] *Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market*, 2019 O.J. (L 130) 92.

reliability of generative AI tools. These issues highlight the broader risks of misinformation and reputational damage when AI models are trained on unvetted or misattributed data.

This case represents more than a private dispute between two entities; it is poised to define how Indian copyright law will interact with rapidly advancing AI technologies. Courts will need to strike a balance between protecting creators' rights and enabling innovation through AI. This also raises broader policy questions: Should India adopt specific exceptions for TDM? Should there be stronger consent frameworks or opt-in mechanisms? And most importantly, how can the law ensure fairness, accountability, and transparency in the AI ecosystem? Whatever the outcome, the ANI vs. OpenAI case is expected to shape India's stance on copyright protection in the digital age, setting the stage for how AI and authorship will co-exist in the years to come.

VI. CONCLUSION

Data mining lies at the heart of modern AI development, yet its legal and ethical implications remain highly contentious, particularly in the realm of copyright. As evidenced in the ANI v. OpenAI litigation, the unregulated use of copyrighted content for AI training exposes a systemic gap in India's intellectual property framework. This vacuum not only leaves creators unprotected but also encourages practices that undermine both their economic rights and moral agency. The current reliance on opt-out mechanisms and voluntary safeguards is woefully inadequate, shifting the burden of enforcement onto authors instead of holding developers accountable. At the same time, legitimate concerns from the AI development community regarding the need for non-expressive and transformative data use cannot be ignored. As such, a nuanced legal approach is imperative. One that distinguishes between expressive replication and functional learning, and that embeds informed consent as a prerequisite for data use. Whether through legislative amendment or judicial interpretation, India must move toward a rights-preserving, innovation-friendly copyright model. If not addressed promptly, the unchecked exploitation of creative labour through data mining could erode not just individual authorship, but the very integrity of copyright law in the digital age.

VII. REFERENCES

Books

- Abhishek Gupta, *The Ethical Machine* (Routledge 2022).
- Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2d ed. O'Reilly 2019).
- F. Provost & T. Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (O'Reilly Media 2013).
- Pedro Domingos, *The Master Algorithm* (Basic Books 2015).
- Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach* (4th ed. Pearson 2020).

Journal Articles & Academic Papers

- Daniel Gervais, *Data and the Future of Copyright*, 45 Colum. J.L. & Arts 1 (2022).
- James Grimmelmann, *There's No Such Thing as a Computer-Authored Work — And It's a Good Thing, Too*, 39 Colum. J.L. & Arts 403 (2016).
- J. Gervais, *AI and Copyright: Fair Use is Not the Answer*, 24 Vand. J. Ent. & Tech. L. 609 (2022).
- M. Feldman, *The Uncharted Territory of AI and Copyright*, Harv. J.L. & Tech. Dig. (2023).
- Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805 (2018).

Cases

- *ANI Media Pvt. Ltd. v. OpenAI OpCo LLC*, CS(COMM) 1028/2024 (Del. H.C.).

- *Baker v. Selden*, 101 U.S. 99 (1879).

Legislation & Government Documents

- European Commission, *Guidelines on Text and Data Mining* (2020).
- European Parliament, *Artificial Intelligence Act: Text and Data Mining Exceptions*, Legislative Briefing (2022).
- *Directive (EU) 2019/790 of the European Parliament and of the Council on Copyright and Related Rights in the Digital Single Market*, 2019 O.J. (L 130) 92.

Online & Other Sources

- Casey Newton, *AI Can't Unlearn What It Knows*, Platformer (2023), https://www.platformer.news.
- Delhi High Court Issues Notice to ANI and OpenAI in Copyright Infringement Matter, Bar & Bench (Apr. 2024), https://www.barandbench.com.
- OpenAI, *GPT Fine-Tuning Guide*, OpenAI Documentation (2023), https://platform.openai.com/docs/guides/fine-tuning.