INDIAN JOURNAL OF LEGAL AFFAIRS AND RESEARCH

VOLUME 3 ISSUE 1

Peer-reviewed, open-access, refereed journal

IJLAR

+91 70421 48991
editor@ijlar.com
www.ijlar.com

DISCLAIMER

The views and opinions expressed in the articles published in the Indian Journal of Legal Affairs and Research are those of the respective authors and do not necessarily reflect the official policy or position of the IJLAR, its editorial board, or its affiliated institutions. The IJLAR assumes no responsibility for any errors or omissions in the content of the journal. The information provided in this journal is for general informational purposes only and should not be construed as legal advice. Readers are encouraged to seek professional legal counsel for specific legal issues. The IJLAR and its affiliates shall not be liable for any loss or damage arising from the use of the information contained in this journal.

Introduction

Welcome to the Indian Journal of Legal Affairs and Research (IJLAR), a distinguished platform dedicated to the dissemination of comprehensive legal scholarship and academic research. Our mission is to foster an environment where legal professionals, academics, and students can collaborate and contribute to the evolving discourse in the field of law. We strive to publish high-quality, peer-reviewed articles that provide insightful analysis, innovative perspectives, and practical solutions to contemporary legal challenges. The IJLAR is committed to advancing legal knowledge and practice by bridging the gap between theory and practice.

Preface

The Indian Journal of Legal Affairs and Research is a testament to our unwavering commitment to excellence in legal scholarship. This volume presents a curated selection of articles that reflect the diverse and dynamic nature of legal studies today. Our contributors, ranging from esteemed legal scholars to emerging academics, bring forward a rich tapestry of insights that address critical legal issues and offer novel contributions to the field. We are grateful to our editorial board, reviewers, and authors for their dedication and hard work, which have made this publication possible. It is our hope that this journal will serve as a valuable resource for researchers, practitioners, and policymakers, and will inspire further inquiry and debate within the legal community.

Description

The Indian Journal of Legal Affairs and Research is an academic journal that publishes peer-reviewed articles on a wide range of legal topics. Each issue is designed to provide a platform for legal scholars, practitioners, and students to share their research findings, theoretical explorations, and practical insights. Our journal covers various branches of law, including but not limited to constitutional law, international law, criminal law, commercial law, human rights, and environmental law. We are dedicated to ensuring that the articles published in our journal adhere to the highest standards of academic rigor and contribute meaningfully to the understanding and development of legal theories and practices.

WWW.ijlar.com
Volume 3 Issue 4 | April 2026 *ISSN: 3049-043X*

TRAINING AI ON COPYRIGHTED DATA: FAIR USE OR INFRINGEMENT

AUTHORED BY - A PRERNA MAHENDRA

Abstract:

The growing amount of generative AI technology has created a significant legal and ethical dilemma regarding the intersection of technological innovation. The central issue in this debate is whether training a large language model (LLM) and creating a diffusion algorithm from improperly using copyrighted works will be considered "fair use" or illegal copyright infringement if done to train the AI. This study examines the question of whether or not AI training is considered "fair use" under United States copyright laws by performing an analysis of the four (4) statutory fair use tests used to determine if a valid fair use exists. In order to do so, both the support of technology companies that advocate AI development by claiming that the models they are training are very transformative in that they are taking only non-copyrighted statistical patterns from the underlying works and using them to create new, statistical based output, and the opposition of copyright holders who argue that taking an entire copyrighted work from a copyright holder without permission to use the work undermines the creative human process, exploits the owner's proprietary IP, and creates a competing product, will be discussed through the application of the four (4) statutory fair use tests. Additionally, this paper will discuss existing judicial precedents, emerging licensing arrangements, and possible legislative solutions to the problem. This paper submits that although existing fair use doctrines are inadequate for addressing the systemic impact of AI training, a rigid copyright infringement framework will impede the further advancement of technology. To balance the need for continuing technological advancements with the need for fair compensation to copyright holders, this paper will present a hybrid regulatory proposal.

Keywords: Generative AI, Copyright Infringement, Fair Use, Large Language Models

Introduction:

With the emergence of generative artificial intelligence comes a major transformation of the global creative industry. The technological changes also created a major legal tension: large language models, generative art systems, etc., require enormous and a vast array of datasets to be effective but there's a large portion of these datasets that are comprised of copyrighted works used without permission, attribution or payment to copyright holders.

As a result, there has been an array of high-profile lawsuits involving tech companies, media companies, authors, visual artists, and software engineers. At the centre of these disputes lies the concept of "fair use"; a legal principle established to avoid rigid copyright laws from interfering with creativity and progress. After 2026, when the publicly available web data used to train many generative models will reach exhaustion, the legal system's need for high-quality, proprietary, human-generated data will be the main barometer for measuring the sustainability of AI systems.[1] The question itself is deceptively simple but also the subject of many court decisions: What are the issues of whether or not unauthorized scraping of copyrighted materials is "fair use" (leading to machine learning or use of that data for business) or if such scraping is an infringement of a copyright? To resolve the question, we will need to break down the methods used by companies engaged in using AI to develop and develop their products or services using the copyrights and property rights of others; (ii) assess the applicability of traditional or common law principles of copyright to new and futuristic digital/technological copyright, and (iii) consider the conflicting economic interests created by the technology industries versus those of creative industries and artists.[2]

Understanding How Ingestion Works: Machine Learning vs Copy

To assess the legality of AI training, it becomes crucial to clarify the line between the two types of actionable digital reproduction so they can then be evaluated based on mathematical theory. This will heavily depend on whether the deep learning process is modeled after cognitive human processes versus being an automatic form of reproduction.

[1] J.C. Charlesworth, *Generative AI's Illusory Case for Fair Use*, 27 Vand (2025).
[2] Pamela Samuelson, Unbundling Fair Uses, 77 Fordham L. Rev. (2009).

After AI can produce any sort of output, it still needs to have been provided data through an overwhelming number of "tears" for the AI to give back. AI have many “tears” on the net, but the most well-known are those from the Common Crawl. The import is to harvest data from the net into your own repository, so that you can uniquely process it.[3]

There are differing processes depending on the type of copyright (i.e., author’s intent, private rights, etc.) In short, registering an author’s intent to produce an output does leave some amount of author’s rights remaining, as the act of developing a copy of the copyrighted work from the harvested data will be taken as prima facie copyright infringement under Section 106 of the Copyright Act.[4]

Statistical Vectoring vs. Protectable Expression

AI developers reply to this argument by showing there is a difference between the act of training versus the end result of the trained model. The final, trained large language model (LLM) or diffusion model does not maintain a database of the scraped texts/images used to train it. Unlike a digital archive for long-term storage, trained models use sets of optimized data (weights/biases). When processing training data through a neural network structure, the model identifies multidimensional statistical relationships among tokens, and uses that information to identify the structural grammar, semantic syntax, and mathematical styles of human expression. Those styles don't fit within the meaning of standard copyright as defined by the idea-expression dichotomy (see § 102(b)). Because of this, AI companies argue that their algorithms do not copy a human's expression, but instead they analyze non-copyrightable structural vibrations to determine the most probable next logical token in any given sequence of tokens.[5]

The Statutory Tool: Fair Use Test 4 Factors

The Federal Copyright Act of 1976, Section 107 provides best defense against copyright infringement claims in U.S. Courts are required to evaluate fair use claims on an individual basis

[3] H. Didsbury & X.A. Zhu, *Transformative Training: An Analysis of AI Training Data and Fair Use in Authors Guild v. OpenAI Inc.*, 4 Publ'g Rsch (2025).
[4] Jane C. Ginsburg & Luke Ali Budiardjo, Authors and Machines, 34 Berkeley Tech. L.J. (2019).
[5] Mark A. Lemley & Bryan Casey, Fair Learning, 99 Tex. L. Rev. (2021).

utilizing a four-factor statutory standard of determination that looks at fair use in a balanced manner.[6]

First Factor: Purpose and Character of Use

The four-factor test considers whether the new use, will in fact, provide an alternative use that has a new purpose or character with some degree of alteration to the original work through an entirely new expression, meaning, or message. This factor is often referred to as "transformative use."[7]

The Technological Angle: In the context of copyright, technological teams have relied heavily on the Authors Guild v. Google, Inc. Decision in 2015, wherein it was ruled that the mass digitization of books create a searchable snippet database a transformative and hence, fair use. AI Developers argue that the use of language arts as a tool in building a multi-billion parameter engine that can summarize code and think provides an analytical capability that no longer exists through static print media.[8]

Conversely, creators of copyrightable materials will contend that generative AI fulfills the initial intent of using human expression, which is communication of ideas, via story-telling and visual creativity but simply shifts the manner in which their work is consumed as it is now done so via an automated algorithm. The fact that the majority of commercial companies that are expanding the functionality and usability of Foundation Models are heavily funded and pursuing a competitive advantage in market share diminishes the likelihood of fair use being established under Factor 2 due to the strong commercialization of data by those who created it. As a result of its systematic reproduction of prepaid materials, the Foundation Models have resulted in a significant increase in evidence supporting a higher level of creative expression associated with works created using Foundation Model technologies.[9]

[6] Authors Guild v. Google, Inc.
[7] Campbell v. Acuff-Rose Music, Inc. (1994).
[8] Y. Huang, *Limitations of Current Copyright Frameworks for Large Language Models Trained on Scientific Literature*, 9 Frontiers A.I. (2026).
[9] Harper & Row Publishers, Inc. v. Nation Enters. (1985).

Second Factor: Amount and substantiality of portion of work used

Generally, the greater the amount of copying from a work as whole, the stronger the disfavor of finding of "fair use" usage of that work as that substantially detracts from value and financially impacts the original creator.[10]

In ML systems at the time of consumption, almost everything is consumed 100%. Due to broken snippets, algorithms are unable to identify semantic contexts and will not be able to learn stylistic weights; thus they will need to consume the entire text or image. Artificial Intelligence companies will assert their reasoning for being able to utilize wholesale replication is that it's technologically necessary to support their intended transformative purpose in the same way that Google needed to copy all of the entirety of books so they can accurately index them for search purposes. The plaintiff believes that copying creators' entire life works, to train a commercial system created to replace the creator/founder, greatly exceeds historical fair use guidelines.[11]

Third Factor: What impact does use have on the potential market?

Historically, this was considered the most important factor in analyzing whether the authorization of unauthorized uses may diminish the value of an author or creator's work in the potential marketplace (current or future) or reduce the ability to license their work.

Market Substitution: This is the strongest argument from the perspective of creators. Whereas the Google Books platform contains only brief excerpts of material and then provides a link to buy the material, generative AI can create an entire piece of text, code, or illustrations that will compete with the creators of the material used to create the AI. For example, if a model can produce a custom work of fantasy fiction in the style of a currently working author in a few seconds, then it is a direct substitute in the marketplace for that author's work.[12]

[10] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith (2023).
[11] G.M. Lentner, *Generative AI and Copyright in the EU and the USA: A View from International Investment Law*, 21 J. Intell. Prop. L. & Prac. (2026).
[12] Pierre N. Leval, Toward a Fair Use Standard, 103 Harv. L. Rev. (1990).

Licensing Market Disruption: Rights holders believe that there is a large and mature market for AI data licensing, which is continuing to develop. By using the data above-mentioned, they ignore the legitimate and established commercial market for this type of data, depriving independent creators of income generated from royalties and creating harm to the greater creative economy.[13]

Intellectual Property Doctrine Evolution

To address the complexity of machine learning technologies, the judiciary is working to re-mould or reshape current intellectual property doctrines. A better understanding of how these doctrines are evolving is possible by analysing both more contemporary international legislative initiatives as well as examining examples of past digital cases.

Warhol Decision

One significant change in the law occurred with the Supreme Court's recent decision in Andy Warhol Foundation for the Visual Arts, Inc. v. Lynn Goldsmith, The Court held that whether a work produced after another can properly be considered transformative cannot simply be determined based solely on what new artistic meaning or message has been added to the original work, but must also consider the underlying commercial purpose of the second piece. Since the underlying commercial purpose of a generative AI tool closely mirrors/overlaps the underlying purposes of human creators within the commercial space (e.g. editorial writing, commercial illustrations, stock photography), this new precedent has created significant limitations to the technology sector's broader "transformative use" defence.

Developing Solutions: Closing the Gaps created by our System

While the Court System continues navigating through complex legal cases, the technology and entertainment industry are finding ways to create practical, systemic solutions that create resolution within this gap.

[13] A. Radeisen, *Open Foundation Models and TDM Exceptions to Copyright – Building Blocks for an AI Ecosystem*, 75 GRUR Int'l (2026).

Business-to-Business Licensing Frameworks

The growth of commercial licensing on a voluntary basis continues to be massive businesses today, with top A.I. Companies looking toward content marketplaces like news and social media to build large partnerships that secure the data required to succeed. While the establishment of these partnerships resolves issues for larger media companies, the independent freelance creator will continue to lack the leverage to create a private licensing agreement.[14]

Technical Protocols & Algorithmic Opt-Out

Web developers with the ability to have more control over their assets have created the use of new technological protocols. The utilization of traditional web standards such as GPTBot or Google-Extended within a website's robots.txt file has permitted these platforms to prevent web scrapers from utilizing their data to build future models. This solution is absolute going forward, but does not provide any recourse or erase the accumulated amount of data that is currently built into the models used today.

Collective Licensing for Statutory Compensation

For a sustainable, fair resolution to arise, a group of distinguished legal experts would be in favor of establishing an independent collective licensing agency similar to those already existing in the music industry, for example, ASCAP or BMI (Xiaoyun Xi, Xiaofu Liu). Within the established statutory scheme:

AI developers will pay a common licensing fee based on the amount of usage and the commercial aspect of their model into a central account. This central account will distribute the funds to creators of works used within AI's training dataset, for the purpose of paying royalties to those creators. Courts can determine the exact amount of damages for willfully/egregiously exploiting copyrighted works, whereas unintentional, minor infringements should be closely protected against extreme liability exposure. By developing this dynamic ecosystem, technology developers

[14] B. Zhang, *Unintentional Infringement by Generative AI: Protecting Niche Creators and Allocating Liability in China and the United States* (2026).

will have access to the breadth of training data required, while allowing for a continuing, predictable revenue stream for human creators.

Should courts rule that a commercial license agreement needs to be made before using any piece of copyrighted text or image; the open source software environment could end. There are only a few dominant tech companies that can afford to negotiate and secure licensing agreements with the major media companies around the globe.[15]

Start-ups, academic research groups and open source developers would be priced out of all training of foundational models. It is ironic that stringent copyright enforcement would merely reinforce the big tech monopolies that consumer advocates are speaking out against, and create a closed environment in which only the largest and wealthiest organizations will be able to develop artificial intelligence (AI).[16]

There are those who believe that artificial intelligence is capable of being trained in the same manner as a human author. The author has spent their entire life studying thousands of books and has learned all of the structure, pacing, character types and style of writing of each of those books. The author writes an original manuscript and does not owe any royalties to anyone whose book(s) they have read.[17]

Likewise, it is believed that to penalize software only because it is capable of utilizing computational power inaccurately assigns differing expectations and exemplifies a flaw that does not understand how an individual's intelligence is developed through observation and synthesis of experience.[18]

Conclusion: Toward an Equitable Synthesis

The question of whether or not to train AI systems on copyrighted material isn't merely a simple dichotomy of 'property rights at all costs' versus 'let's innovate with no limits.' Rather, it is an

[15] Neil Weinstock Netanel, Copyright and a Democratic Civil Society (1996).
[16] Sonia K. Katyal, Technoheritage, 105 Calif. L. Rev. (2017).
[17] X. Zhang, Digital Alchemy? Rethinking Copyright in the Age of AI-Generated Content: Lessons and Reflections from the AI Value Chain (2026).
[18] Ryan Abbott, The Reasonable Robot: Artificial Intelligence and the Law (Cambridge Univ. Press 2020).

essential dilemma that defines how we as a society determine, assess, and protect creative works in an automated age.[19]

When considering the four factors for fair use, the outcomes are directly opposed. Specifically, due to the analytical and non-recreative (structurally) nature of computational ingestion of data, fair use would apply in this case; however, the commercial application of these models to create output that competes with the creative (prior existing) output would constitute infringement because it threatens the continued economic viability of the creative class.[20]

To resolve this dilemma will require a much broader approach than simply relying on binary judicial decisions that serve as clouds for addressing the impact of significant structural industry changes. Rather, we need to develop new legislative frameworks, develop models for collective licensing, and implement sound technical standards for opt-outs. In doing so, we would develop a transparent ecosystem that recognizes the effort of human labor and provides appropriate credit and compensation to the creator of the work to ensure that the next generation of AI is developed with the creator rather than at the expense of the creator.

[19] Lawrence Lessig, Free Culture (Penguin Press 2004).
[20] W. Neil Terry, Copyright Challenges of Generative Artificial Intelligence (2024).