Christ University Law Journal
2025, Vol. 14, No. 2, *23-47*
ISSN 2278-4322/https://doi.org/10.12728/culj.27.2

Global Perspectives on Fair Use and AI Training Data

Rishabh Tomar*

Abstract

The increasing adoption of artificial intelligence (AI) has recently posed important questions regarding the propriety of copyrighted data in training datasets. Under fair use and similar exceptions and limitations as legal doctrines, it is important to evaluate the permissions for such usage. Despite this, its meaning and implementation differ greatly from one jurisdiction to another, raising legal ambiguity when developers engage in AI businesses across geographical borders. The article addresses a pressing issue. It offers a comparative examination of fair use and analogous doctrines in the United States, the European Union, and emerging markets, set against a rapidly evolving legal landscape. Recent landmark litigation in the United States involving AI developers, alongside the adoption of the EU AI Act, has brought renewed urgency and clarity to debates on the lawful use of copyrighted material in the context of AI development. This study intends to show how these frameworks can be used to explain questions concerning copyright, transformative use, and public benefit. Finally, this study examines regulatory trends and concerns about AI from a legal and policy perspective, assesses the alignment and misalignment of regulatory frameworks, and explores the challenges and possible solutions.

Keywords: AI Ethics, Copyright, Digital Single Market, Text and Data Mining, Licensing

* UILS, Chandigarh University, Mohali, Punjab, India; tomar.rishabh1996@gmail.com

1. Introduction

The fast pace at which artificial intelligence (AI) has developed raises an urgent legal issue - under what circumstances should the application of copyrighted material in AI training data sets be considered as lawful act of fair use? Although the legal underpinnings of fair use in such jurisdictions as the United States are highly established, their use with AI training, which involves consuming and processing massive amounts of copyrighted text, images, and code, is legally immature and controversial.[1]

The doctrine of transformative use lies at the centre of this modern controversy. The main problem with transformative use as the fundamental principle of the fair use analysis in the U.S. is whether the new work is performed in a way that a new purpose, different character, expression, meaning or message has altered the original work. According to the argument based on a permissive view of AI training, the transformation of copyrighted content into numerical representations (embeddings) with the aim of identifying statistical patterns and creating an original and non-infringing work is transformative in nature.[2] This postulates that the aim of utilizing the data, i.e., to train a model to execute a new type of functionality is fundamentally opposed to the original expressive intent of the work.[3]

On the other hand, certain scholars are arguing that the non-transformative unauthorized copying during the training phase per se is a commercial exploitation that poses the threat of devaluing their creative markets.[4] Such tension is enhanced by major jurisdictional mismatch. The European Union, specifically, does not have a general doctrine of fair use but has certain narrow exceptions of Text and Data Mining (TDM) under the Digital Single Market (DSM) Directive, which does not depend on the transformative use analysis, and is subject to opt-outs by rightsholders .[5]

The paper will offer a comparative analysis of the applications of the notion of transformative use and their functional counterparts, to AI training datasets in major jurisdictions. Going beyond the general theoretical discussion of fair use, this paper addresses the issue of the changing and disputed understanding of the concept of transformative use. It will also

[1] NICOLAS SUZOR, THE CONTESTED GOVERNANCE OF AI AND THE RULE OF LAW, IN THE OXFORD HANDBOOK OF AI GOVERNANCE (Oxford Univ. Press 2023).

[2] Mark A. Lemley & Bryan Casey, Fair Learning, 99 TEX. L. REV. 743 (2021).

[3] Pamela Samuelson, Generative AI Meets Copyright, 381 SCIENCE 158 (2023).

[4] Daniel J. Gervais, The Machine as Author, 18 OHIO ST. TECH. L.J. 1 (2022).

[5] Eleonora Rosati, The Exception for Text and Data Mining (TDM) in the Proposed Digital Single Market Directive: Technicality or Policy Change?, 52 IIC INT'L REV. INTELL. PROP. & COMPETITION L. 150 (2021).

will examine the existing compatibility and non-compatibility in copyright systems of specific jurisdictions, the implications of this in developing AI on a cross-border scale, and the possibilities that exist to maintain a more compatible and balanced approach.

Some of the key questions surrounding fair use and similar exceptions in AI training are characterizing the use of copyrighted material in AI datasets in terms of legal factors, including transformation, purpose, and public interest. Furthermore, how do jurisdictions diverge in implementing these doctrines? Legal provisions differ greatly, with the U.S. employing a wide-open four-factor test focusing on transformative use, while the E.U. or one of its member states depends on very limited Data Mining exceptions and limitations under the Digital Single Market (DSM) Directive. As the existing guidelines are at a nascent stage in many emerging economies, we get to see a clear disparity in what is being implemented across borders and what this means for harmonization. This paper uses doctrinal method to compare the fair use doctrine and the analogous copyright exceptions in the application of AI training data within the jurisdictions of the United States (U.S.), the European Union (EU), and the emerging markets of India and South Africa.

2. Literature Review

2.1. The Foundation of Fair Use and Transformative Use

The fair use doctrine in U.S. copyright law, according to 17 U.S.C. § 107, allows the limited use of copyrighted material without the owner's consent for the purpose of criticism, comment, news reporting, teaching, scholarship, or research. The doctrine employs a flexible four-factor test, with the "purpose and character of the use" being the most important factor. Within this factor, the notion of "transformative use" has emerged as the crucial point of judicial inquiry. This concept was originally laid down in the Supreme Court case of *Campbell v. Acuff-Rose Music, Inc.* (1994),[6] which ruled that a use is transformative only when it "adds something new, with a further purpose or different character, thereby turning the original into a new one with new expression, meaning, or message." This rule was more emphatically confirmed and broadened in *Google LLC v. Oracle America, Inc.* (2021),[7] where the Court ruled that the copying of software code for the transformational purpose of making a new platform was fair use. In the context of AI and machine learning, some academic circles argue that the ingestion of copyrighted data for training purposes can be perceived as being quite transformative because it does not take over the original work but rather examines it to recognize statistical patterns for a completely different

[6] 510 U.S. 569 (1994).

[7] 593 U.S. 1 (2021)

functional aim.[8] The basic premise of transformative use as characterized by the courts is crucial for the assessment of the fair use exception in the case of AI training datasets.[9] Recent analyses emphasize the fact that the in-depth studying of generative AI models based on large data sources is mostly "quintessentially transformative," because it identifies patterns for new outputs instead of just reproduction of the old expression.[10] Nevertheless, transformative use is a relative concept, and in some instances, commerciality or market harm could be a factor against fair use.[11] This changing viewpoint points out the flexibility of fair use in the face of technological changes, while at the same time, protecting the creator's rights.

2.2. The European Model: Exceptions over Fair Use

Unlike the U.S. model, the European model is concise in that it has a list of exceptions and limitations. The idea of implementation of Text and Data Mining (TDM) exceptions in Articles 3 and 4 of the Digital Single Market (DSM) Directive is widely criticized in the literature. Although considered a step, such scholars as Rosati (2021)[12] and Geiger and others (2023)[13] note that they have certain limitations, such as the opt-out right of Article 4 that introduces legal uncertainty in cross-border AI projects and their implementation fragmentation between member states.[14] It is cited as a major structural impediment to innovation in AI, as the lack of an overarching and loosely construed fair use doctrine in the EU is driving developers to more complicated and expensive licensing frameworks. developers to more complicated and expensive licensing frameworks.[15]

2.3. Emerging Economies and Global Harmonization Standpoints

The emerging economy discourse focuses on the special issues and opportunities these jurisdictions encounter. Researchers observe that some

[8] Matthew Sag, Copyright Safety for Generative AI, 61 HOUS. L. REV. 295 (2023).

[9] Mark A. Lemley & Bryan Casey, supra note 2, at 2.

[10] Bartz v. Anthropic PBC, No. 3:25-cv-02710, 2025 WL 2874752 (N.D. Cal. Mar. 12, 2025).

[11] U.S. COPYRIGHT OFF., COPYRIGHT AND ARTIFICIAL INTELLIGENCE: PART III – GENERATIVE AI TRAINING (2025).

[12] Eleonora Rosati, Copyright and Artificial Intelligence in the EU: Rethinking Text and Data Mining Exceptions, 43 EUR. INTELL. PROP. REV. 219 (2021).

[13] Christophe Geiger et al., Copyright and AI: Challenges and Opportunities, 31 INT'L J.L. & INFO. TECH. 45 (2023).

[14] João Quintais et al., The DSM Directive in National Courts: Toward Copyright Harmonization?, 17 J. INTELL. PROP. L. & PRAC. 12 (2022).

[15] MAURIZIO BORGHI & STAVROULA KARAPAPA, COPYRIGHT AND MASS DIGITIZATION: A CROSS-JURISDICTIONAL PERSPECTIVE (Oxford Univ. Press 2020).

of the most populous nations such as India and South Africa are walking a fine line between following global standards of copyright and the promotion of homegrown AI development.[16] Nair (2023) research on India examines the possibilities of the Indian judiciary to develop flexible interpretations of the concept of fair dealing[17] and Nkosi (2022) looks into the South African Copyright Amendment Bill, which suggests a U.S.-style fair use clause.[18] The international aspect of the issue has triggered the increasing literature on harmonization. Researchers are also considering the possibility of multilateral treaties on AI-specific copyright exceptions with some analogies to the success of such treaties as the Marrakesh Treaty.[19]

2.4. Ethical Dimensions and Alternative Models

In addition to a very strict legal analysis, the literature touches more and more on ethical concerns. Such authors as Binns (2022) suggest transparency and accountability in the creation of datasets and associate ethical AI practices with the necessity to comply with copyright.[20] Moreover, there is a controversial issue concerning the role of licensing as an alternative or a supplement to fair use as a market-based mechanism. Feng and others (2023), consider the new AI-specific licensing models,[21] but the author Samuelson (2023) warns about the possible stagnation of innovation and the entrenchment of dominance of large tech companies on the basis of licensing.[22]

3. Fair Use and AI Training Data: Legal Foundations

3.1. Definition and purpose of fair use

The idea of fair use in many countries involves fair use and related exceptions and limitations of copyright materials without having to obtain permission from the writer. It is to protect the holder's interests while simultaneously promoting the interests of the general public in gaining access to information

[16] Vandana Singh, Harmonizing Copyright Frameworks in Emerging Economies, 10 GLOB. INTELL. PROP. REV. 250 (2023).

[17] Priya Nair, Fair Use in the Context of AI: Legal Challenges in India, 28 J. INTELL. PROP. RTS. 45 (2023).

[18] Thandeka Nkosi, The Copyright Amendment Bill: A Step Towards Fair Use in South Africa, 139 S. AFR. L.J. 295 (2022).

[19] Andres Guadamuz, Copyright and Artificial Intelligence: Navigating Transformative Use in AI Training, 45 EUR. INTELL. PROP. REV. 15 (2023).

[20] Reuben Binns, Transparency and Accountability in Artificial Intelligence: Ethical Challenges and Legal Frameworks, 2 AI & ETHICS 123 (2022).

[21] Xiao Feng et al., Licensing Models for AI Training: Opportunities and Challenges, 5 J. AI POL'Y & REG. 89 (2023).

[22] Pamela Samuelson, Generative AI Meets Copyright, 381 SCIENCE 158 (2023).

and furthering learning and development.[23] The fair use doctrine began in the United States and developed in its jurisprudence before becoming part of the Copyright Act of 1976.

The four-factor test stated in Section 107 of the Copyright Act 1976 is the basis for evaluating fair use. It requires evaluating:

1. The nature and extent of the use and the purpose of the use, whether for commercial or non-profit educational purposes. Although there are restrictions, expressions that add new uses or create new meanings are preferred.

2. Purpose of the copyrighted work, where non-fiction works are more likely to be protected under fair use than artistic works.

3. Proportion of the portion taken, firstly, to highlight the amount of material used and secondly, the profound importance of the portion used.

4. Market impact determines whether the use actually displaces the market for the copyrighted work or the potential market for the work.[24]

3.2. The Blurred Lines of Transformative Use: A Comparative Case Law Analysis

The doctrine of transformative use has emerged as the focal point over which the amount of copyrighted material used in AI training may qualify as a fair use or not. The primary issue in the fair use analysis of the U.S. is often transformative use wherein one using a copyrighted work makes a different use or with a new expression, meaning, or message. Nevertheless, the distinction between transformative fair use and infringement appropriation is becoming more and more blurry with recent landmark decisions in the field of AI and non-AI.

3.2.1. Pre-AI Precedents: Establishing the Transformative Use Spectrum

Before the AI age, courts were establishing the foundation of what it is that transformative use is. In *Authors Guild v. Google, Inc.* (2015), the Second Circuit found that Google digitizing millions of books to form a searchable database was a highly transformative fair use.[25] The court believed that the end result of the copies, to facilitate text mining and search which did not seem to disclose any expressive content to a human reader, was completely

[23] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).

[24] Copyright Act of 1976, 17 U.S.C. § 107 (2023).

[25] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

unrelated to the original aesthetic and expressive intent of the books. The case is a great testament to the argument that a non-expressive use of a copyrighted work such as a data analysis is transformative.

On the other hand, the recent ruling of the Supreme Court in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith* (2023) limited the area of transformative use in a commercial situation.[26] The Court believed that Warhol silkscreen portrait of Prince, which was created on the basis of Lynn Goldsmith photograph, was not distinctive enough to qualify as fair use as both works had essentially the same commercial role the purpose of licensing magazine illustrations. This decision highlights that a new aesthetic or meaning might not be sufficient where the secondary use competes with the original either in the same market or a similar market.

3.2.2. AI-Specific Litigation: Direct Application to Training Data

The ethics of these precedents are being directly challenged in AI litigation. In *The New York Times Company v. Microsoft Corp. and OpenAI Inc.* (2023), the plaintiffs assert that the application of their copyrighted articles to be trained on large language models (LLMs) such as ChatGPT is not transformative but a mass reproduction that produces a competing product.[27] They believe that in case an AI model is capable of producing output that recreates or summarizes the style of a Times article, it replaces the original directly and injures its market.[28] This case will be the ultimate outcome of the argument of commercial interest against AI training.

Conversely, defendants of such cases, relying on Google Books, claim that such training process is a non-expressive, intermediate use. They argue that it is of no significance to ingest in order to learn the statistical patterns of the language, rather than to republish or communicate the safeguarded expression of the works.[29] The resulting AI model, they suggest, is a novel good, which serves other purposes (e.g., code generation, conversational assistance), and thus is transformative.

3.2.3. Public Interest vs. Commercial Interest: The Core Tension

This is due to the fact that the legal ambiguity lies within the basic tension between the public and commercial interests, which is at the core of copyright law.

[26] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023).

[27] N.Y. Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023).

[28] Benjamin L. Sobel, The Fight Over Generative AI and Copyright Has Begun, 67 COMMC'NS ACM 22 (2024).

[29] Pamela Samuelson, *Supra* Note 19, at 6.

1. **The Public Interest Case on AI Training:** Advocates of the wide fair use exception on AI training focus on the vast social good of AI innovation. They claim that the use of stringent copyright imposing restrictions on the training data would be the death of advancements in such important areas as medical research, climate studies and educational resources, since these areas are not copyrightable due to the fact that they are patents.[30] In this regard, the act of using copyrighted data to train is a contemporary scholarship and research just like a scholar going through thousands of books to come up with a new theory. It is the transformative nature in the making of a new and powerful knowledge generation tool rather than reproducing the training texts themselves.
2. **The Commercial Interest Argument of Holders of Copyright:** According to the argument of the rights holders, the commercial implications of AI development and its magnitude cannot be overlooked. They speculate that AI corporations are making commercial ventures worth billions of dollars on the intellectual property of their clients, without their consent or payment.[31] As soon as AI model can create content that will be competitive to even the works that it was trained on, the fourth fair use factor, which is market harm, will take center stage. The issue is that the commercial success of the creators of AI is being cross-subsidized by the unpaid use of creative works that undermine the value of the market of these works and discourages the creation of new ones.

This tension is less pronounced under the EU's text and data mining exceptions, where the boundaries are clearly defined in law. Articles 3 and 4 draw a deliberate distinction between uses undertaken for non-commercial research, which reflect broader public interest objectives, and a wider exception that accommodates rightsholders' interests. By contrast, the United States relies on a case-by-case fair use analysis. While this approach is more flexible, it also creates significant uncertainty, as courts are now required to weigh the broad societal value of AI technologies against the legitimate commercial interests of individual creators and firms. The outcome of ongoing litigation, including The New York Times v. OpenAI,[32] is therefore likely to be decisive in shaping clearer, albeit still contested, boundaries for the AI industry.

[30] James Grimmelmann, Copyright for Literate Robots, 108 IOWA L. REV. 1681 (2023).

[31] NICOLAS P. SUZOR, THE LAW OF AI FOR GOOD, IN THE OXFORD HANDBOOK OF ETHICS OF AI 617 (Markus D. Dubber et al. eds., Oxford Univ. Press 2023).

[32] No. 1:23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023)

3.3. Current Regulatory Challenges in Global AI Data Protection and Copyright Law

Copyright and data protection laws both shape how datasets for artificial intelligence are created and used. Copyright law governs the use of creative works such as text, images, and music within training data, while data protection law regulates the collection and handling of personal information, much of which is embedded in modern datasets. In the context of AI training, this overlap creates a dual compliance challenge, requiring developers to navigate both regimes at the same time.

In the present digital environment, national copyright laws generally require prior authorization before copyright-protected works can be included in AI training datasets, unless such use is allowed under recognized exceptions. In the United States, this takes the form of the fair use doctrine, while in the European Union it is addressed through text and data mining exceptions. Recent judicial developments, particularly the Andy Warhol Foundation v. Goldsmith[33] precedent in the United States, have narrowed the scope of transformative use, placing AI developers who rely on unlicensed copyrighted material in an increasingly uncertain legal position. The GDPR places substantial constraints on the processing of personal data within the EU, creating a complex compliance landscape for the development of AI systems that rely on such data for training.[34]

This means that balance is the key when it comes to practising it. To address these challenges, initiatives for harmonised AI-specific exception proposals are due in an attempt to foster innovation while considering the legal framework. Academics also stress the need to make dataset creation open and accessible, for which licensing systems and stringent processes of anonymisation should be established to solve copyright and privacy issues at once.[35]

4. Comparative Analysis of Fair Use Doctrines

4.1. Fair use in the U.S. Copyright Act

In the United States, there is a favoured approach known as the fair use doctrine laid down in Section 107 of the US Copyright Act. The doctrine assesses four factors.

[33] 598 U.S. 508 (2023)

[34] Pamela Samuelson, Copyright's Fair Use Doctrine in the AI Age, 37 BERKELEY TECH. L.J. 987, 987-1012 (2022).

[35] Christophe Geiger, Giancarlo Frosio & Oleksandr Bulayenko, Copyright and AI: Challenges and Opportunities, 32 Int'l J.L. & Tech. 45, 45-67 (2023).

1. The nature and extent of use, especially that of transformational nature.
2. The field of endeavour of the copyrighted work.
3. Whether it amounted to a noticeable or substantial portion of the whole.
4. Therefore, the impact of use on the market value of the original work must be considered.[36]

There are key juridical treatments that define fair use. In *Authors Guild v. Google Inc.*,[37] the federal appeals court supported Google's utilization of copyrighted books in its Google Books undertaking by arguing that the book digitization project is transformative. The court stated that the discussed service of Google provided public benefit without replacing the original works.[38] Similarly, in *Google LLC v. Oracle America Inc.* (2021)[39] the Supreme Court supported Google by noting that its utilization of Oracle's Java API for Android was transformative, thus amounted to fair use and encouraged innovation.[40] All these cases epitomize how the US focuses on transformative use and public interest in fair use assessments.

4.2. Applicability to AI training datasets

AI is clearly in the crosshairs of the transformative use doctrine, as the training datasets are nearly always derived from copyrighted works to create new attributes. While deciding on the issue, the nature of use has been further examined on the basis that it produces something different from the purpose under consideration. For example, AI models, such as those used for writing or creating images, might claim fair use if AI changes the content and if this change purpose is beneficial for society, such as in education or research, or for persons with disabilities.[41]

Simultaneously, public benefit corresponds to the objectives of the AI industry, further development of AI, and its increased use in areas such as diagnostics, medicine, and environmental control. However, general knowledge about, for instance, datasets that have to be used for machine learning is missing, which leads to legal risks.[42]

[36] Copyright Act, 17 U.S.C. § 107 (2023).

[37] 804 F.3d 202 (2d Cir. 2015),

[38] Matthew Rimmer, Google Books and the Future of Fair Use, 45 J. INTELL. PROP. 256, 256-270 (2020).

[39] Google LLC v. Oracle America Inc., 593 U.S. 1 (2021)

[40] Pamela Samuelson, Google v. Oracle: What It Means for Fair Use and Software Development, 24 STAN. TECH. L. REV. 87, 87-104 (2021).

[41] P. Bernt Hugenholtz, Fair Use and Artificial Intelligence: A Comparative Perspective, 34 J. COPYRIGHT L. & PRAC. 123, 123-140 (2022).

[42] David McGowan, AI and Fair Use: Navigating the Legal Landscape for Training Datasets, 36 HARV. J.L. & TECH. 45, 45-78 (2023).

The current legal landscape is marked by uncertainty and a surge in litigation. This uncertainty has been compounded by the U.S. Copyright Office's decision to initiate a policy study on artificial intelligence, inviting public comments on how copyright law should apply to the training and output of AI models. Together, ongoing lawsuits and regulatory scrutiny expose developers to significant legal risk, effectively requiring them to make significant judgments about how courts will ultimately rule on whether AI training qualifies as transformative use.

The doctrine of transformative use lies at the centre of current debates on AI, as most training datasets rely on copyrighted material to produce new outputs. Courts assessing this issue focus closely on the purpose and character of the use, particularly whether the material is employed for a function different from its original intent. For instance, AI systems that generate text or images may seek to rely on fair use where they meaningfully alter the source material and where such use serves broader social interests, such as education or research. There is however a litigation being put to test of this legal theory now. In *Anthropic, PBC v., OpenAI, Inc.* (2024), a collective of authors and publishers accused the defendants of using their AI models to provide service on their copyrighted materials without their consent, which amounted to colossal copyright violations.[43] The result of this case will become a precedent that may help to establish whether the ingestion and use of the text under copyright to train LLMs can be defined as a non-infringing fair use of the text or it should be licensed (Samuelson, 2023).[44] Simultaneously, the social good is linked to the goals of the AI sector; the further evolution of AI and its greater application in such spheres as diagnostics, medicine, and environmental control. Nonetheless, there is no overall information regarding, say, data sets that must be employed to train machines, which makes them legally dangerous (McGowan, 2023).[45]

4.3. European Union: A Restrictive Framework

4.3.1. Absence of a fair use doctrine in EU copyright law

The European model can be described as being devoid of fair use and instead favouring selected exceptions, which are narrow-minded and focused, and has been applied to its landmark AI regulation. The 2024 EU AI Act introduces mandatory transparency obligations for providers of general-purpose AI models, requiring them to publish a sufficiently detailed

[43] Anthropic, PBC v. OpenAI, Inc., No. 3:24-cv-04555 (N.D. Cal. filed July 5, 2024).

[44] Pamela Samuelson, Copyright Law and AI: Reconciling Licensing with Innovation, 75 STAN. L. REV. 987 (2023).

[45] David McGowan, AI and Fair Use: Navigating the Legal Landscape for Training Datasets, 36 HARV. J.L. & TECH. 45 (2023).

summary of the data used for training. While these provisions do not amend copyright law directly, they create a clear link between AI regulation and copyright compliance. By compelling developers to disclose their data sources, the Act indirectly pressures them to demonstrate that their training practices comply with the text and data mining exceptions under the DSM Directive.

TDM Exceptions under the DSM Directiv*e*

While Article 3 expressly allows TDM for otherwise lawful research purposes by non-commercial parties for scientific research purposes, Article 4 of the DSM Directive also allows wider discretion for all parties where certain conditions are satisfied regarding the manner of access. However, these exceptions are spread across member states, and therefore, their management is inconsistent.[46]

Limitations of the TDM Framework for AI Training

TDM exceptions are a positive trend, but we cannot consider them sufficient for training AI on copyrighted datasets. The opt-out provision provided under Article 4 raises a much greater concern for AI developers since there is a constraint implementation that restricts the rights holder from allowing access to data. In addition, the 'lawful access' angle of the directive does not consider the practical challenges of achieving this on a large scale, which is indispensable for training massive LLMs.[47] The practical issues of the lawful access requirement and the opt-out provision are important. As an illustration, the opt-out mechanism, as applied by certain rightsholders using metadata tags, has been discussed by leading European research institutions and AI start-ups as potentially creating an unadvisable motive of isolating large chunks of the common internet, making it impossible to train large-scale models on European data. This poses a contrast between the desire of the EU to lead the world, in terms of AI and the confining aspect of the copyright system.

4.3.2. Challenges for AI developers in the EU

The restrictive nature of EU copyright law presents significant challenges for AI developers.

[46] MAURIZIO BORGHI & STAVROULA KARAPAPA, COPYRIGHT AND MASS DIGITIZATION: A CROSS-JURISDICTIONAL PERSPECTIVE (Oxford Univ. Press 2020).

[47] Eleonora Rosati, Copyright and Artificial Intelligence in the EU: Rethinking Text and Data Mining Exceptions, 43 EUR. INTELL. PROP. REV. 219, 219-225 (2021).

1. **Fragmentation:** The lack of uniformity of TDM exceptions and limitations across EU member countries makes this issue challenging for multinational AI projects.[48]
2. **Licensing Burdens:** In the absence of fair use doctrine, developers rely on expensive and time-consuming licensing, which stifles innovation, particularly in emerging start-ups and SMEs.[49]
3. **Legal Uncertainty:** The opt-out and non-uniformity issue and all the inconsistencies themselves discourage investment in artificial intelligence research and development within the EU.

Finally, even though the DSM Directive pays lip service to the necessity of TDM exceptions and limitations, the flaws inherent in the concept and the complete lack of fair use doctrine led to the EU's problems stated in its inability to remain competitive in the field of AI. These are the areas that future reforms need to consider to support innovation as well as copyright law.

4.4. Emerging Economies: A Developing Narrative

The case of AI and copyright in emerging economies such as India and South Africa is inherently connected to their socio-economic agenda and the path of digital development. In contrast to the U.S. and the E.U. which are established technological powerhouses, these countries are both trying to build a strong domestic AI innovation ecosystem to drive economic growth, and at the same time, demand equitable access to international knowledge and technology.[50] This is their economic stance which is a key determinant in their attitude towards copyright exceptions. However, compared to the developed economies, which discuss the specifics of transformative use, in emerging markets, it is usually more important to focus on the policies to enable access to information, substitute technological dependence, and foster local innovation, especially towards small- and medium-sized enterprises (SMEs) and uses of public interest.[51] The lack of obvious, AI-related exemptions, therefore, poses a substantial obstacle, rather than an ambiguity of the law, that might impede their participation in the AI world race.

[48] João Quintais, Giancarlo Frosio & Stef van Gompel, The DSM Directive in National Courts: Toward Copyright Harmonization?, 17 J. INTELL. PROP. L. & PRAC. 12, 12-22 (2022).

[49] MAURIZIO BORGHI & STAVROULA KARAPAPA, COPYRIGHT AND MASS DIGITIZATION: A CROSS-JURISDICTIONAL PERSPECTIVE (Oxford Univ. Press 2020).

[50] U.N. Conf. on Trade & Dev., Technology and Innovation Report 2021: Catching Technological Waves Innovation with Equity (2021), https://unctad.org/system/files/official-document/tir2020_en.pdf.

[51] Rishabh Ghosh, AI for Development: The Role of Intellectual Property in Emerging Economies, 25 J. WORLD INTELL. PROP. 456 (2022).

4.4.1. Fair use and its equivalents in India and South Africa

Owing to historical factors, modern developing countries such as India and South Africa are in a rather difficult position in adapting copyright legislation to the challenges and requirements of AI development. India and South Africa have both taken a more fluid approach to copyright exceptions and limitations, but their legal frameworks are unclear when fair use doctrines are used for AI training datasets.

Copyright Act, 1957 of India does not contain a specific provision labelled as fair use. Instead, it recognizes the doctrine of fair dealing under Section 52. The provision specifies specific uses such as criticism, review, and research; there is no room for interpretation in any AI-related matter.[52] As regards technology and Copyright in India, a fair use case in point is *India TV Independent News Service Private Limited Vs Yashraj Films Private Limited (2012).*[53] While this case was mainly about fair use in Indian copyright law, some of the concepts raised in this paper about the training data for AI could also be generalised.

In this case, the Delhi High Court applied the four fair use factors and recognized that trivial or minimal infringements may fall within the *de minimis* doctrine, which at times overlaps with fair use defences. Although the judgment did not address AI or machine learning directly, its approach to fair use is significant for current debates. It offers guidance on whether datasets derived from copyrighted works for training artificial intelligence could be treated as fair use or dismissed as *de minimis* use.

Like many countries, South Africa's Copyright Act of 1978 has exceptions and limitations for fair dealing for the purpose of research and private study, yet has no provisions specific to modern technology. "Fair use" has been proposed under the Copyright Amendment Bill, 2019, and new exceptions and limitations for text and data mining have been proposed and added, although these reforms remain unoperationalised.[54]

4.4.2. Policy and Regulatory Barriers to Sustainable AI Innovation

Currently, there are limitations in the existing legal framework pertaining to copyright protection to support AI developers in using copyrighted content to optimise AI without violating intellectual property rights. For example, the lack of clear regulations addressing transformative use norms in the case

[52] Prashant Nair, Fair Use in the Context of AI: Legal Challenges in India, 28 *J. INTELL. PROP. RTS.* 45, 45-56 (2023).

[53] India TV Independent News Service Pvt. Ltd. v. Yashraj Films Pvt. Ltd., (2012) FAO(OS) Nos. 583 & 584 (Del. HC)

[54] Thandeka Nkosi, The Copyright Amendment Bill: A Step Towards Fair Use in South Africa, 139 S. AFR. L.J. 295, 295-312 (2022).

of AI may lead to litigation and subsequently it might hinder development of innovation.[55] Similarly, the lack of clarity in the legal framework hampers AI development in developing countries and increases the gap between developed and developing countries.

Several challenges originate from inadequate alignment and ineffective cooperation with international copyright frameworks and enforcement mechanisms. India and South Africa face challenges in negotiating domestic policies and international obligations in treaties such as the Berne Convention.[56] The uncertainty regarding application of fair use and related exceptions undermines international cooperation and deters investment in AI technologies. These gaps can be addressed only through copyright reform agendas that align with international standards and provide a supportive environment for innovation.

4.4.3. The India AI Initiative: A Policy Push to Bridge the Data Gap

The Government of India has recognized critical gaps in data access and innovation and, to strengthen the AI ecosystem, has launched the IndiaAI Initiative as a broad, enabling programme. One of the most important components of this program is the establishment of the so-called IndiaAI Datasets Platform, which will create the so-called high-quality, non-personal, and anonymized datasets in the domestic AI industry (Ministry of Electronics and Information Technology.[57] This is an initiative that directly responds to the lack of training data, which has been discussed in the first section. The initiative seeks to reduce reliance on potentially infringing copyrighted material and lower barriers to innovation by curating and supplying large-scale, legally permissible datasets to developers. The initiative however, depends on how it will overcome the very ambiguities of copyright in the current system of fair dealing. The operational policies of the platform will have to specify the legal position of data aggregation and processing, which could become a decisive point on how the copyright law in India should be adjusted to meet the requirements of AI.[58]

[55] Rishabh Kumar & Arjun Das, AI Innovation and Copyright Law in India: Bridging the Gap, 17 INDIAN J.L. & TECH. 150, 150-167 (2021).

[56] Vikram Singh, Harmonizing Copyright Frameworks in Emerging Economies, 10 GLOB. INTELL. PROP. REV. 250, 250-267 (2023).

[57] Ministry of Elec. & Info. Tech., IndiaAI: Unlocking AI's Potential for India (2024), https://www.meity.gov.in/indiaai.

[58] Rajesh Kumar & Amitava Das, AI Innovation and Copyright Law in India: Bridging the Gap, 17 INDIAN J.L. & TECH. 150 (2021).

Table 1 provides the comparative analysis between the U.S., EU and emerging economies (India and South Africa):

Aspect	United States: A Flexible Approach	European Union: A Restrictive Framework	Emerging Economies: A Developing Narrative (India and South Africa)
Legal Basis	Section 107 of the Copyright Act which set out the standard for fair use known as the four-factor test.	A combination of fair dealing and explicit fair use only as the DSM Directive allows only for limited text and data mines.	India: Fair dealing under Copyright Act, 1957; South Africa: Statutory exceptions and limitations under the Copyright Amendment Bill.
Key Factors for Applicability	Stress on the idea of use, purpose and market impact. The advantage of broad and flexible interpretation is that much is left to AI development.	TDM exceptions and limitations as provided for research purposes, but with a possibility for copyright owners to opt out. No regard for the concept of transformation or public benefit.	India: Lack of guidelines on how transformative use can be achieved. South Africa: It encompasses mere public interest as well as education interest but does not cover provisions of AI law.
AI Dataset Usage	Regularly supports fair use defences in AI training, more so under the transformative use concept.	TDM exceptions and limitations unable to accommodate large scale AI training with stock conglomerate of copyrighted works thereby stinging innovation.	Uncertainty in both countries about status of AI datasets; new discussions about whether to put within the ambit of fair use or fair dealing.
Policy Environment	Moderately innovation friendly with clear case law defining it (e.g., *Authors Guild v. Google).*	At the regulatory level there is a shift of concern towards recognizing users' rights while at the same time having robust measures to guard copyrights.	India: Gradual enlargement of the purposes of fair dealing as recognized by the judiciary. South Africa: Modernisation is the primary purpose of the Copyright Amendment Bill but it is still a subject on many controversies.
Challenges for AI Development	Lack of predictability with reference to general legal cases coming before court for determination especially involving issues of AI; use of litigation to establish legal precedents concerning issues of AI.	Many developers have licensing costs and are unsure about what is legally acceptable regarding TDM activities.	The two jurisdictions struggle with uncertainty of law as well as lack of clear guidance related to the use of AI's, thus a potential for disparate treatment of similar entities as well as lack of incentive to innovate.

Table 1

5. Implications for AI Development

5.1. Convergence and divergence in global fair use practices

The use of fair use and related exceptions for AI training information exhibits remarkable differences across jurisdictions. In the United States, fair use is generously defined within the flexible four-factor framework which gives significant importance to the nature of use: transformative, which makes it quite beneficial for AI training.[59] For example, US courts have often relied on fair use arguments whenever new information is generated for a different use, as in *Authors Guild v. Google Inc*.[60]

On the other hand, the EU has no general fair use rule. It relies on specific statutory exceptions under the DSM Directive, including limited exceptions for text and data mining. Although well intentioned, these exceptions are narrow. They permit only limited users, such as researchers and educational institutions, to use copyrighted works for training without legal uncertainty, while commercial AI developers remain exposed to potential liability if they rely on copyrighted material. Some countries rely on statutory regulations, while others rely on undefined doctrines of fair dealing. Some countries, such as India and South Africa, use both. However, there is a weak judicial doctrine to clarify these concepts for developers who wish to exploit these doctrines and exceptions.[61]

5.2. Legal uncertainty and its impact on cross-border AI projects

The inconsistency in the application of fair use doctrines and related exceptions results in legal risks to AI developers, especially when enforced in different jurisdictions. For instance, a dataset assembled in the US that is perceived as lawfully compliant with fair use laws may not be permissible for legal use in the EU contracting zone, considering its stricter framework of copyright law.[62] Innovation driven by fragmentation leads to high compliance costs and higher chances of legal battles, discouraging international partnerships.

In practice, many businesses manage legal risk by relying on licensing agreements, which can limit access to important training data. The cost of these licenses is often beyond the reach of smaller firms and developers, particularly in developing countries, thereby reinforcing existing global

[59] James Grimmelmann, Copyright for Literate Robots: TDM and Fair Use, 43 COLUM. J.L. & ARTS 1, 1-36 (2020).

[60] 804 F.3d 202 (2d Cir. 2015)

[61] Ramon Lobato, Fair Use and AI: A Global Perspective, 14 INT'L J.L. & TECH. 221, 221-240 (2022).

[62] Alain Strowel & Nathalie Ide, Copyright Challenges in AI and Data Mining: A European Perspective, 28 J. INTELL. PROP. 231, 231-251 (2021).

inequalities in AI development. Moreover, the absence of a standardised legal approach makes it harder to build AI systems that can operate across jurisdictions, ultimately slowing innovation.

The legal divide between the United States and Europe remains stark. A Europe-based AI company that trains its models using web-scraped data must comply with the EU's text and data mining exceptions and respect any opt-outs exercised by rightsholders. If the same company seeks to deploy its model or collaborate with partners in the United States, it must also assess its exposure to fair use litigation, where identical scraping practices may be treated as infringing unless and until a court rules otherwise, as illustrated by cases involving *Stability AI*.[63] This regulatory misalignment forces companies to adopt complex, jurisdiction-specific data governance frameworks, increasing costs and constraining collaboration and cross-border innovation.

The legal ambiguity that arises as a result of differences in fair use principles unfairly affects developers in growing economies. While, large global companies can absorb the costs of extensive licensing and legal disputes, but SMEs and start-ups in countries such as India and South Africa operate with far more limited capital. The threat of litigation or the initial price of licensing huge datasets can be prohibitive in their case and they would be locked out of training state-of-the-art AI models.[64] This contributes to a data divide in which inaccessibility to high quality training data strengthens global inequalities in the development of AI. Thereby, the ambiguity in the application of the law in these regions, as well as a lack of its development, is not only the legal risk but also a structural obstacle to economic and technological catch-up, which continues to consolidate the position of a limited number of tech giants of the Global North.[65]

In conclusion, this paper has demonstrated that fair use doctrines and related exceptions are practiced without uniformity in different jurisdictions, creating hassles on AI development and underscoring the need for globally harmonised policies somewhere between innovation and protection of exclusive rights.

[63] No. 3:23-cv-00201-WHO (N.D. Cal. filed Jan. 13, 2023)

[64] World Intell. Prop. Org., WIPO Conversation on IP and AI: Draft Issue Paper on Intellectual Property Policy and Artificial Intelligence, Seventh Session (2022), https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_7_ge_22/wipo_ip_ai_7_ge_22_1_rev.pdf.

[65] INDERMIT S. GILL & JAMIL ZAHID, THE GREAT DIVIDE: AI AND THE GLOBAL SOUTH (Carnegie Endowment for Int'l Peace 2023).

5.3. Ethical considerations: balancing innovation and copyright protection

Machine learning and generative AI models rely heavily on data for training, much of which consists of copyrighted datasets. Protecting and encouraging innovation while defending copyrights and authors' rights raises pressing ethical questions. On the one hand, the evolution of AI requires a variety of datasets to create balanced, impartial, and fair technologies. On the other hand, the excessive use of copyrighted material risks eroding creators' rights and weakens the incentive structure that sustains creative production.

Ethical considerations support the use of fair use doctrines and related exceptions as the foundation for an overarching concept of transformative use. This approach occupies a middle ground by enabling the creation of new and socially beneficial outputs while ensuring that the economic and expressive value of the original works is not undermined.[66] However, a certain degree of unpredictability arises due to the unclear meaning of the term transformative use in legal parlance which makes the environment of AI developers somewhat challenging. In addition, the ethical principle of explainability, to some extent, requires developers to disclose the nature and sources of the data used for training. This expectation aligns AI development with prevailing standards of accountability and trustworthiness.[67]

Moreover, the use of data from minorities should be fair and equitable. The absence of consent in the use of cultural or creative works deepens existing disparities, making this one of the most contested societal effects of AI. Thus, ethical AI development should not only observe legal requirements on the use of copyright but also take precautions to avoid using content that is demeaning to others.

5.4. Role of licensing as an alternative to fair use

Given the complexity and unpredictability of fair use and related exception analyses, copyright licence mining emerges as a reasonable mechanism for acquiring training data for AI systems. Licensing ensures that authors and rights holders are compensated for their content, while at the same time providing developers with legal certainty. For instance, some, such as Shutterstock and Getty Images, offer AI-specific licencing, allowing large datasets to be provided if the terms are clear.[68]

[66] Andres Guadamuz, Copyright and Artificial Intelligence: Navigating Tranformative Use in AI Training, 45 EUR. INTELL. PROP. REV. 15, 15-21 (2023).

[67] Reuben Binns, Transparency and Accountability in Artificial Intelligence: Ethical Challenges and Legal Frameworks, 7 AI ETHICS J. 123, 123-140 (2022).

[68] Xiaoming Feng, Sheng Liu & Yiming Zhang, Licensing Models for AI Training: Opportunities and Challenges, 5 J. AI POL'Y & REGUL. 89, 89-105 (2023).

However, licensing may also function as a constraint, particularly for resource-constrained developers who lack the capacity to secure access to large-scale datasets. Such disparities risk entrenching industry dominance among a small group of major players, thereby limiting competition and diversity within the AI ecosystem.[69] Moreover, an excessive emphasis on licensing regimes may erode the public domain and other free-use exceptions, thereby constraining inventive and transformative AI development.

Legal uncertainty has destabilised existing market arrangements, prompting a shift towards licensing mechanisms as a defensive response to infringement and regulatory risks. Although sites such as Shutterstock and Getty Images have AI licenses available, we are also witnessing strategic alliances that are not restricted to the usual licensing. For instance, OpenAI has entered into content licensing agreements with news organisations such as Axel Springer and The Associated Press, allowing it to use their content in exchange for a licensing fee. Likewise, Google has launched a 'Generative AI Updater' which publishers can use to prevent the use of content in training AI. On the one hand, these developments represent a way to a compliance. On the other hand, such arrangements risk creating a two-tier ecosystem in which well-capitalised incumbents can afford extensive licensing agreements, while start-ups are left to rely on uncertain fair use defences. This dynamic may ultimately entrench the market power of a small number of dominant firms.

Hence, if licencing is pragmatic, it should be used in conjunction with the provisions of fair use. All stakeholders should strive to create fair conditions which are legally affordable and protect authors' rights.

5.5. The Intersection of Copyright Doctrines and AI Ethics

Laws on AI training with the use of copyrighted content are not just technical legal issues, but are deeply connected with the main aspects of AI ethics. The decision between a flexible fair use system (U.S.) and a restrictive licensing-based system (E.U.) carries ethical consequences far reaching on how to come up with equitable and trustworthy AI systems.

One of the main ethical issues is prejudice and representativeness. The bias of the dataset in restrictive copyright schemes can be increased by the fact that the expense of a license is very high, and therefore the developers develop models based on a restricted sample of commercially obtained or low-cost data.[70] This dynamic can systematically marginalise the creative

[69] Pamela Samuelson, Copyright Law and AI: Reconciling Licensing with Innovation, 75 STAN. L. REV. 987, 987-1024 (2023).

[70] Daniel J. Gervais, The Contours of AI and Copyright, 36 HARV. J.L. & TECH. 397 (2023).

outputs of less privileged communities, as their content is less likely to be captured within mainstream licensing frameworks. As a result, existing societal biases risk being reproduced and reinforced in AI-generated outputs. However, a lenient application of the doctrine of fair use can also facilitate more comprehensive and diverse datasets as it allows one to use a wider range of human knowledge, which is needed to create fair and unbiased models.[71]

Moreover, the transparency and explainability principle of AI ethics is in direct conflict with the obscurity required in copyright risk reduction. In order to circumvent legal hurdles in the stringent copyright laws, developers can be encouraged to conceal the content and origin of their training datasets. However, this practice compromises the ethical requirement of transparency, making it difficult to audit AI models for bias and accuracy, or to trace the origins of their outputs.[72] A more appropriate approach to legal compliance and ethical AI development would involve a copyright framework that explicitly preserves space for auditing and accountability research, potentially under a more flexible fair use standard.

Lastly, the doctrine of transformative use operates as a legal proxy for an underlying ethical trade-off between creative freedom and the protection of original works. When the copying or use of a copyrighted work is deemed transformative under fair use, it typically indicates that the new work serves a different purpose from the original. In doing so, it contributes added value to the public and aligns with the ethical objective of promoting innovation for broader societal benefit.[73] This allows the copyright system not to be abused to shut down the creation of AI programs that do not simply replace the original works, but provide new functionalities, including a diagnostic tool in healthcare or a creative form. Thus, transformative use, as recognized in law, functions as an important mechanism for ensuring that copyright law encourages, rather than deters, ethical and socially valuable AI innovation.

[71] Nicolas P. Suzor et al., What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation, 13 INT'L J. COMMC'N 18 (2019).

[72] Heike Felzmann et al., Transparency You Can Trust: Transparency Requirements for Artificial Intelligence Between Legal Norms and Contextual Concerns, 7 BIG DATA & SOC'Y 1 (2020).

[73] Matthew Sag, The New Legal Landscape for Text Mining and Machine Learning, 66 J. COPYRIGHT SOC'Y U.S.A. 291 (2019).

6. Toward Harmonization: Bridging Global Differences

6.1. Challenges of harmonizing fair use doctrines

The consideration and integration of fair use doctrines and related exceptions across jurisdictions is a complex task, owing to inequalities among sovereign nations and the pressures of globalisation. As a legal principle, fair use and related exceptions are embedded in cultural, economic, and legal paradigms that differ across communities. For instance, although America recognizes a flexible fair-use system, the European Union has specific exceptions and limitations regarding copyright provisions. Likewise, developing countries operate within distinct socio-economic contexts and therefore tend to prioritise access to knowledge and innovation over stringent copyright protection.

National sovereignty poses a problem in developing a single and coherent idea, as nations may find these efforts an attempt to encroach on their right to legislate based on national concerns. This reluctance is due to geopolitical and economic relations that affect negotiations between two or more countries. For instance, creating economies may decline frameworks assessed as beneficial to created countries or large AI corporations.

However, AI development has occurred worldwide and requires universal cooperation. Data collected from several jurisdictions frequently in AI models pose a challenge owing to the present inconsistencies in copyright laws. This means that harmonisation must capture the interplay of these tensions, promote a fair process of integration, and design frameworks that accommodate the clash of legal cultures when facing international problems.

Therefore, coordinating national interests with the need for an international framework that promotes AI-related exceptions and limitations remains inherently challenging. In this context, tailored approaches that strategically utilize multilateral agreements become necessary. Such approaches should focus on incorporating carefully defined exceptions and limitations for AI while ensuring that any resulting harmonisation delivers tangible benefits to all key stakeholders, including creators, developers, and users. This balance is necessary in developing mechanisms for stimulating innovation to meet current needs while continuing to respect the rights of international patent holders.

This has been compounded by the differences in policy objectives that are being enshrined in the pertinent laws. The AI Act of the EU focuses on transparency and a risk-based approach, which is human-oriented, which is consistent with its conservative copyright attitude. In the meantime, the U.S is more innovation-focused, more adaptable, with the doctrine of fair use and has not enacted extensive legislation on AI. The regulations on China are set differently on the issue of content control and the socialist core values. Balancing a global structure must thus balance not only the legal doctrines, but essentially disparate regulatory philosophies and values in the society regarding technology and creativity and governmental role, in relation to technology.

6.2. Proposed strategies for a balanced global framework

To overcome the existing international differences in the fair use doctrine and to promote a more balanced system of AI development, policy-makers and business executives should stop theorizing and implement multi-faceted measures. The proposals that are to be made below aim at establishing a legal certainty without infringing on the principles underlining copyright.

6.2.1. The Need for a Multilateral Treaty on AI-Specific Exceptions

The best long-term response is the introduction of a new international treaty, which is similar to the Marrakesh Treaty, but oriented towards the era of AI. This convention would not transplant a U.S.-style fair use doctrine into other jurisdictions. Instead, it would introduce a compulsory minimum exception for non-expressive text and data mining. This arrangement would create a separation in the eyes of the law between the act of training an AI (a non-expressive and analytical act) and the output of it (which would be considered equally liable to the scrutiny of copyright). By signing such a treaty, signatory states would recognise that the use of copyrighted material for computational analysis, pattern recognition, and model training does not amount to infringement. This recognition would apply only where the underlying source material has been lawfully accessed. This would give legal predictability to cross-border AI projects and ensure that there is not a race to the bottom where only the liberal jurisdictions will be the ones that benefit innovation.

6.2.2. Mandating "Data Provenance and Transparency" as a Legal Safeguard

Ethical data practices should be clearly aligned with legal safe harbours to provide both normative guidance and regulatory certainty. We suggest that adherence to sound Transparency and Provenance Systems may act as a countervailing element in copyright cases or a requirement to enjoy AI exceptions. Developers must be requested to:

1. **Document Datasets:** Having auditable documentation of sources of training data, which emphasize the usage of licensed and public domain and open access data.
2. **Use Takedown Mechanisms:** Having effective mechanisms in place to take down particular copyrighted works in training datasets when requested to do so by the rightsholder, but along machine learning pipelines.

3. **Publish Model Cards:** Making public the wide-ranging composition and traits of training data that foster trust and permit critique without disclosing proprietary model weights.

This approach shifts the focus from whether data was used to how it was used, thereby rewarding responsible developers and creating a market advantage grounded in transparency.

6.1.3. Fostering Market-Based Solutions with Compulsory Licensing Pools

Although voluntary licensing is preferable in principle, it is unlikely to function at the scale required for AI training. An international framework could therefore support the creation of copyright collectives and licensing pools for AI related uses. Within such a framework, a regulated and, where necessary, mandatory licensing system could be introduced once rightsholders in a particular category, such as news publishers, scientific journals, or stock image providers, have organised themselves into a sufficient critical mass. It would allow creators to get a fair compensation at the same time give developers an easy one stop shop to legally clear large quantities of data. This model has a balanced approach to property rights of creators and the functional needs of innovators to avoid a market failure.

6.3. Proposed Implementation Mechanisms

Although the strategic objectives of multilateral agreements and ethical transparency are clear, their realisation depends on concrete regulatory and operational mechanisms. In the absence of the implementation structures, these proposals may only exist as theories. In this section, possible models of governance, compliance and technical implementation are outlined. In order to fill the existing gap, the following action plans are suggested:

1. **Make it an International Technical Standard:** There should be a technical standard on Ethical AI Data Sourcing, created by an international organization (e.g., the WIPO). This standard would certify datasets that meet a transparent provenance, sound rights clearance (e.g., through standardized licenses) and rational curation. Conformity might provide a safe haven against some infringement.

2. **Establish a Multilateral Clearinghouse:** It should have a centralized, multilateral centralised licensing body of AI training data. This platform would enable large scale licensing for rightsholders, maintain a database of pre cleared public domain and licensed works, and offer a streamlined dispute resolution mechanism. Together, these functions would reduce transaction costs and legal uncertainty for developers operating across borders.

Such mechanisms would translate the abstract demand for balance into a practical and enforceable framework, providing concrete guidance on how AI should be governed at the global level.

7. Conclusion

Copyright exceptions relating to AI training are not merely fragmented across jurisdictions; they pose a significant obstacle to the responsible and equitable development of artificial intelligence. This analysis highlights a fundamental divergence between the United States approach, which is flexible and open ended, fostering innovation at the cost of legal predictability, and the European Union approach, which is more prescriptive and rights oriented, prioritising certainty for rightsholders. Caught between these models, emerging economies risk being marginalised or becoming arenas for legal and regulatory conflict.

The underlying issue lies in the false dichotomy often drawn between innovation and protection. There is a need to recognise AI training as a transformative, non-expressive use that is integral to contemporary scientific and technological progress. The justifications of the balanced framework do not lie only in the legal or economic arguments but in the ethical and practical ones. In the absence of an international strategy, there is a risk of an AI data oligopoly emerging, in which only well financed companies are able to navigate complex global licensing regimes. Such concentration would entrench market power and contribute to the homogenisation of AI development.

The proposals advanced include a multilateral treaty on text and data mining, a commitment to data provenance, and the development of innovative licensing pools. These measures are not merely aspirational suggestions but essential pillars for a functional global AI ecosystem. Together, they offer a roadmap for reconciling national sovereignty with the borderless nature of data and innovation.

Future studies should monitor the resolution of high-profile cases in the U.S. and the enforcement of the transparency regulations in the EU AI Act since they will offer invaluable practical information on the sustainability of existing practices. Policy makers should consider model safe harbour provisions for non-commercial AI research, or alternatively, adopt a model specific text and data mining exemption similar to the approach taken in Japan. International treaties, such as a potential WIPO agreement on AI and copyright, should focus on establishing minimum standards of interoperability between differing legal systems rather than imposing compulsory and unrealistic uniformity.