# COPYRIGHT IN THE AGE OF GENERATIVE AI: LEGAL CHALLENGES AND THE FAIR USE DOCTRINE IN THE CONTEXT OF TRAINING DATA

Prof. (Dr.) Asmita Vaidya, Principal, Government Law College, Mumbai

M.E.S.V. Krupakar, Research Scholar, Dept. of Law, University of Mumbai

**ABSTRACT**

The rapid proliferation of generative artificial intelligence (AI) systems has intensified legal and ethical debates concerning the use of copyrighted material in training datasets. These systems are built on large-scale ingestion of digital content like text, images, and audiovisual works which are sourced from publicly accessible platforms without express authorization. Such practices raise complex questions regarding unauthorized reproduction, derivative works, and the limits of permissible use under copyright law.

This paper examines the legal ramifications of incorporating copyrighted content into AI training, with particular emphasis on the fair use doctrine in the United States, while also considering parallel fair dealing frameworks in other jurisdictions. It traces the historical development of fair use and its expansion to accommodate technological innovations, especially transformative and non-expressive uses, through landmark decisions.

The paper further engages with contemporary scholarly debates and doctrinal tensions surrounding the application of fair use to AI training. It provides a detailed analysis of recent litigation involving *Bartz v. Anthropic*, *Kadrey v. Meta*, and *Thomson Reuters v. ROSS* highlighting jurisprudential inconsistencies both across these cases and in relation to established precedent on transformative use. Ultimately, the paper argues that in light of disruptive technological advances such as AI, legal systems require greater predictability and clarity and the adoption of explicit text and data mining exceptions may offer a more coherent and future-ready framework for balancing innovation with the protection of authors' rights.

**Keywords**: Generative AI, Copyright Law, Fair Use Doctrine, Training Data, Text and Data Mining (TDM)

## 1. Introduction

The advent of generative artificial intelligence (AI) technologies represents a transformative phase in the domains of machine learning and digital content creation. Advanced models such as OpenAI's GPT series, Google's Gemini, Meta's LLaMA, Anthropic's Claude, and generative image systems like DALL·E and Midjourney have demonstrated unprecedented abilities to produce coherent textual narratives, photorealistic images, functional code, and original music.[1] These developments are not solely technical achievements. They reshape paradigms of authorship, reshape labour dynamics, and challenge the legal structures surrounding intellectual property protection.[2]

A critical component driving this transformation is the manner in which large language models (LLMs) and multimodal models are trained. These systems depend on massive corpora comprising text, images, audio, and video, much of which is sourced from publicly accessible content on the internet.[3] Such data sets often include copyrighted works, including books, journalistic articles, source code, photographs, and artistic content, acquired through web scraping or automated crawling techniques. Although these AI systems typically do not retain or reproduce material verbatim, their output is generated through complex probabilistic associations learned from training data. This has raised pressing legal concerns about whether the ingestion and internal use of copyrighted works without permission amounts to reproduction, derivative creation, or infringement under prevailing copyright statutes.[4]

The legal permissibility of using copyrighted material in the training of artificial intelligence (AI) models remains an evolving and unsettled issue across jurisdictions. Proponents of generative AI, primarily developers and technology companies contend that the ingestion of large volumes of copyrighted data constitutes a *transformative* and *non-expressive* intermediate use. They argue that this process is akin to how human beings acquire language and cultural understanding i.e., by reading and absorbing patterns rather than replicating exact expressions. From this perspective, such use does not infringe the expressive core of the original works but instead serves an essential functional purpose in enabling

[1] see OpenAI, *GPT-4 Technical Report* 3 (Mar. 27, 2023), https://openai.com/research/gpt-4.
[2] See James Grimmelmann, *Copyright for Literate Robots*, 101 Iowa L. Rev. (2016)
[3] Pamela Samuelson, Generative AI. Meets Copyright, 381 SCIENCE 158 (2023)
[4] Guadamuz, Andres, A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs (February 26, 2023). GRUR International 2/2024 (Forthcoming). , Available at SSRN: https://ssrn.com/abstract=4371204

machine learning.[5]

Conversely, authors, artists, and copyright holders assert that the unauthorized appropriation of their works even indirectly through algorithmic processing violates their exclusive rights under copyright law. They contend that such use deprives them of potential licensing revenue, facilitates commercial exploitation without consent, and undermines their control over how their works are used and interpreted.[6] At the heart of this conflict lies a broader normative tension between advancing technological innovation and preserving the economic and moral rights of creative professionals that shapes the contours of contemporary copyright discourse in the age of generative AI.

Adding further complexity to this legal analysis is the global divergence in copyright doctrine. The United States relies on the flexible *fair use* framework, applying a four-factor test to determine whether use without permission is permissible.[7] In contrast, India's *fair dealing* approach is more restrictive, permitting exceptions only for specified purposes such as research, criticism, or review.[8] The European Union, through the Directive on Copyright in the Digital Single Market (DSM Directive), introduced harmonized exceptions for *text and data mining* (TDM), though implementation and scope vary by Member State.[9] Other jurisdictions including the United Kingdom, Japan, and Canada have adopted differing approaches, reflecting diverse policy priorities and levels of openness to technological experimentation.[10] These inconsistencies complicate the regulatory landscape and generate legal uncertainty for both innovators and rightsholders in the global AI ecosystem.[10]

Against this backdrop, the paper proceeds in six parts. Part I traces the historical development of the fair use doctrine, situating its origins in broader copyright theory. Part II examines the jurisprudential evolution of fair use, with particular emphasis on its extension to technological innovations and the recognition of transformative and non-expressive uses. Part III turns to

[5] Jane C. Ginsburg & Luke A. Budiardjo, *Authors and Machines*, 34 Berkeley Tech. L. J. 343 (2019), reprinted as Columbia Public Law Research Paper No. 14-597 (2018), available at https://scholarship.law.columbia.edu/faculty_scholarship/2323

[6] See Letter from Authors Guild et al. to Congress on Generative AI https://authorsguild.org/app/uploads/2024/09/Authors-Guild-Open-Letter-to-Generative-AI-Leaders.pdf (last accessed on July 25th 2025)

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

[8] See Indian Copyright Act, No. 14 of 1957, § 52(1);

[9] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market, art. 3–4, 2019

[10] Olga Kacprzak, *UK Consultation on Copyright and Artificial Intelligence: Walking a Fine Line*, Reed Smith LLP (Jan. 17, 2025), available at https://www.reedsmith.com/en/perspectives/2025/01/uk-consultation-on-copyright-and-artificial-intelligence

contemporary scholarly debates and doctrinal tensions that complicate the application of fair use in the context of AI training. Part IV analyzes recent litigation like *Bartz v. Anthropic*, *Kadrey v. Meta*, and *Thomson Reuters v. ROSS* and highlights the inconsistencies these decisions reveal both among themselves and when measured against established precedent. Part V advances the central argument of the paper: that disruptive technological advances such as generative AI necessitate legal systems that are more predictable, coherent, and forward-looking, and that the introduction of explicit text and data mining exceptions offers a promising model for balancing innovation with the protection of authors' rights. Finally, Part VI concludes by synthesizing these insights and considering their implications for the future of copyright in the age of generative AI.

## 2. Copyright, Fair Use, and the Evolution of Limitations

### 2.1. Copyright as a Concept and Its Public Purpose

Copyright is a legal mechanism designed to reward creators by granting them exclusive rights over the use and distribution of their original works of authorship (e.g., literary, dramatic, musical, and artistic works) for a limited time. However, the ultimate constitutional and philosophical goal of copyright in jurisdictions like the United States is explicitly stated: "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries"[11]

This framing reveals a crucial principle: the grant of a limited monopoly to authors is a means to an end, not the end itself[12]. This principle has profound implications in the context of AI training. The vast corpora of text and data that serve as inputs to generative AI systems are not simply the raw material of private ownership; they are also part of the collective pool of knowledge, whose access is necessary for fostering innovation and advancing science. The core objective is not merely to enrich authors but to stimulate the creation and dissemination of new works, thereby enriching the public domain and fostering a vibrant culture of learning and knowledge. The public is the primary intended beneficiary, gaining access to a wider range of artistic and scientific knowledge[13]. This necessary tension between the private right of the author and the public's right to access and build upon existing knowledge forms the theoretical

[11] U.S. CONST. art. I, § 8, cl. 8.
[12] William M. Landes & Richard A. Posner, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 3 (2003). Neil Weinstock Netanel, COPYRIGHT'S PARADOX 2 (2011).
[13] David Vaver, INTELLECTUAL PROPERTY LAW: COPYRIGHT, PATENTS, TRADE-MARKS 30 (2011).

and practical foundation for all copyright exceptions and limitations, including the doctrine of fair use.

### 2.2 The Doctrine of Fair Abridgement

Before 1710, copyright in England was regulated under royal privilege through the Stationers' Company, which functioned as both a monopolistic guild and a mechanism for censorship. The Statute of Anne, 1710, marked a revolutionary departure by vesting statutory rights directly in authors rather than publishers, for an initial 14-year term, renewable once.[14] The statute, however, was largely silent on limitations and exceptions. It was primarily concerned with granting exclusive rights to "print and reprint" works. Consequently, the judiciary was compelled to develop equitable doctrines to reconcile the private rights of authors with the public interest in access to learning.

The earliest judicial limitation was the doctrine of "fair abridgement." In *Gyles v. Wilcox*, the Court of Chancery held that not all abridgements constituted infringement; a "true abridgement," involving "labour and judgment" that reshaped the original into something new, could be lawful.[15] A mere "colourable abridgement," if it is essentially a disguised copy will be impermissible. This principle represented the first recognition that copyright was not absolute and that productive reworking of existing material could serve public benefit without undermining the author's legitimate rights.

Following *Gyles*, the doctrine was extended in cases such as *Dodsley v. Kinnersley* , which allowed the fair abridgement of sermons. Thus, the early jurisprudence already reflected an awareness that copyright's purpose was not to lock up knowledge indefinitely, but to permit transformative uses that added social value.[16]

### 2.3 The Key Turning Point: From "Fair Abridgement" to "Fair Use"

The English doctrine of "fair abridgement" was the direct ancestor of the modern American doctrine of "fair use," crystallized by Justice Story in the 1841 U.S. case of *Folsom v. Marsh*. Justice Story explicitly moved beyond the abridgement paradigm, asking instead whether the defendant's use was "fair and bona fide." He laid down the now-familiar considerations:

[14] John Feather, PUBLISHING, PIRACY AND POLITICS: AN HISTORICAL STUDY OF COPYRIGHT IN BRITAIN 78 (1994).
[15] *Gyles v Wilcox* (1740) 26 ER 489
[16] *Dodsley v. Kinnersley*, 27 Eng. Rep. 270 (1761)

> "*In short, we must often, in deciding questions of this sort, look to the nature and objects of the selections made, the quantity and value of the materials used, and the degree in which the use may prejudice the sale, or diminish the profits, or supersede the objects of the original work."*[17]

This formulation introduced the four fundamental factors of analysis, nearly identical to those later codified in U.S. law:

1. Purpose and character of the use (Factor One : *nature and objects of selections*).
2. Nature of the copyrighted work (Factor Two: implicitly covered by *nature of the materials*).
3. Amount and substantiality of the portion used (Factor Three: *quantity and value of the materials*).
4. Market effect (Factor Four: *degree in which the use may prejudice the sale... or supersede the objects of the original work*).

Justice Story explicitly transformed the narrow, creative-work-focused "fair abridgement" into a broad, general doctrine of "fair use." He shifted the focus from whether a work was a "new creation" (though that remains relevant) to whether the use, given its purpose and market impact, was *fair* to the copyright holder.[18]

### 2.4 The Contrast: Fair Use vs. Fair Dealing

Modern copyright jurisdictions typically employ one of two primary approaches for exceptions: Fair Use or Fair Dealing.

- Fair Dealing (e.g., India, UK, Canada, Australia): This approach operates on a limited enumeration list (statutory list). For a use to qualify, the defendant must first demonstrate that their use falls within one of the specifically enumerated purposes (e.g., research, private study, criticism, review, or news reporting). If the use falls within a category, the court then proceeds to examine whether the dealing was fair in that context.
- Fair Use (e.g., U.S.): This is an open-ended, flexible doctrine. It does not require a use to fit into a closed list but instead mandates an ad hoc, case-by-case balancing test using

[17] *Folsom v. Marsh*, 9 F. Cas. 342, 348 (C.C.D. Mass. 1841).
[18] Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1105, 1111 (1990).

the four factor[19].

This structure provides less flexibility. If a new technological use, such as AI training, does not fall within the statutory purposes, the defense of fair dealing cannot be invoked, regardless of its transformative value. Thus, fair use offers a more robust doctrinal framework to accommodate rapid technological innovations, while fair dealing may lag behind, requiring frequent statutory amendments to remain relevant.

## 3. The Evolution of Fair Use in Transformative and Technological Contexts

The fair use doctrine, codified in Section 107 of the Copyright Act of 1976, serves as a vital safeguard, balancing the rights of copyright holders with the public's interest in the creation and dissemination of new knowledge and creative works. This statute mandates consideration of four non-exclusive factors: (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used; and (4) the effect of the use upon the potential market for or value of the copyrighted work. The evolution of the first factor, the "purpose and character of the use," has proven critical in addressing challenges posed by rapidly advancing technology and complex creative appropriation.

### 3.1 Judge Pierre N. Leval and the Genesis of the Transformative Use Concept

The doctrinal mechanism required to analyze secondary uses that build upon, rather than merely replicate, the original work emerged in 1990. Judge Pierre N. Leval, who published his seminal article, *Toward a Fair Use Standar.d*. Leval posited that the core inquiry for the first fair use factor must be whether the secondary user merely supersedes the objectives of the original work or rather adds "something new, with a further purpose or different character, altering the original with new expression, meaning, or message". [20]

Leval's framework explicitly shifted the judicial focus within Factor One. Historically, commercial uses were often viewed as presumptively unfair. Leval argued that commercial character should be secondary to the transformative purpose. This conceptualization provided courts with a powerful tool to rationalize fair use in complex situations beyond traditional uses like criticism or scholarship. By centering the analysis on transformation, the doctrine directly linked the fair use exception to copyright's constitutional goal: promoting the progress of

[19] 17 U.S.C. § 107
[20] Level, *supra* Note 18.

science and the useful arts. The more transformative a new work is, the more likely it is to be considered fair use. This idea fundamentally influenced subsequent Supreme Court jurisprudence, enabling courts to accept wholesale commercial copying later in the digital age, provided the underlying purpose was functional or informational, rather than aesthetic competition.

### 3.2 Functional Necessity: Reverse Engineering in Sega Enters., Ltd. v. Accolade, Inc.

The Ninth Circuit in *Sega v. Accolade*[21] extended fair use to reverse engineering in the software context. Accolade disassembled Sega's code in order to understand the interface specifications necessary to make its video games compatible with Sega's console. The court held that intermediate copying for the purpose of achieving interoperability was a legitimate fair use, even though the copied code was verbatim and highly protected. he court stressed that fair use must permit copying when it is the only means to access the underlying, unprotected ideas or functional elements of the program. Here, the transformative element was functional: the copying was necessary not to supplant Sega's games but to foster competition and innovation in the market.

*Sega* created a specific pathway for technological uses which often involve the wholesale copying of the entire copyrighted work to be deemed transformative because the secondary use was non-expressive and aimed solely at engineering interoperability, thereby promoting competition and innovation.

### 3.3 Expressive Transformation: Parody and the Commerciality Factor in Campbell v. Acuff-Rose Music, Inc.

The U.S. Supreme Court in *Campbell v. Acuff-Rose*[22] provided the definitive modern articulation of transformative use. In addressing whether 2 Live Crew's parody of Roy Orbison's "Oh, Pretty Woman" was fair use, the Court emphasized that the first factor, purpose and character of the use, turns significantly on whether the work is "transformative," that is, whether it "adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message."[23] The Court rejected presumptions against commercial uses, holding instead that parody, as a transformative purpose, could outweigh commerciality under the fair use framework.

---

[21] *Sega Enters. Ltd. v. Accolade, Inc.,* 977 F.2d 1510 (9th Cir. 1992).
[22] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994)
[23] *Id*. at 579.

*Campbell* thus provided the doctrinal foundation for evaluating uses that repurpose copyrighted works for new, socially beneficial purposes, including technological applications.

### 3.4 The Ascendancy of Non-Expressive Transformation: Data Processing and Archival Copying

The advent of the digital era forced courts to grapple with mass copying and machine reading, testing the limits of transformation. In this phase, courts consistently prioritized the "new purpose" of data analysis over the expressive content of the original works, cementing the functional pathway established by *Sega*.

The Fourth Circuit in *iParadigms* confronted the use of copyrighted student essays by the plagiarism detection service Turnitin.[24] The service stored copies of student submissions in its database to compare against future works, thereby preventing plagiarism. The court concluded this was fair use because the purpose was highly transformative: the system did not exploit the expressive value of the essays but instead used them as data for comparison. This decision illustrates the application of fair use to non-expressive, functional uses of works in digital environments.

In *Perfect 10,* the Ninth Circuit addressed whether Google's use of thumbnail images in its image search engine constituted fair use. The court held that Google's creation and display of thumbnails was transformative because it served an entirely different function from the original images: providing an indexing and search tool rather than aesthetic enjoyment. Although the original images were commercial and expressive, the court found the search engine's transformative utility outweighed these considerations.[25] This case was pivotal in legitimizing search and indexing functions as transformative uses under fair use.

In *HathiTrust*, the Second Circuit addressed the legality of a consortium of libraries digitizing their collections to create a full-text searchable database and to provide access for the print-disabled.[26] The court upheld this as fair use, recognizing the transformative purpose of creating accessibility for disabled persons and enabling new forms of scholarly research through text mining. As with Google Books, the focus was not on expressive substitution but on functional transformation, reinforcing the role of fair use in advancing technological and societal goals.

[24] A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009).
[25] *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146 (9th Cir. 2007).
[26] *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014).

The culmination of the functional transformation trend arrived in *Authors Guild v. Google*, where the Second Circuit upheld the legality of Google's digitization of millions of books to create a searchable database and provide "snippets" of text.[27] The project involved which involved Google's massive project to scan over twenty million books for its Google Books search engine, displaying short "snippets" in search results.

The court found Google's digitization and creation of a searchable index to be highly transformative, serving to "augment public knowledge" and make information about the books available. The decision emphasized that the purpose of the copying was distinct: indexing for information retrieval, not reading for expressive enjoyment.

Regarding the scope of copying (Factor Three), the court minimized the fact that Google copied entire works, reasoning that wholesale copying was necessary to achieve the transformative function (a complete, searchable index). Furthermore, the output, the limited, "cumbersome, disjointed, and incomplete" snippet view ensured that Google's service was unlikely to provide a market substitute for purchasing the original book (Factor Four).

*Google Books* represents the high watermark of the Non-Expressive Use Doctrine. It sanctions massive, commercial, wholesale copying for a functional, non-consumptive purpose, definitively establishing that machine reading and data mining are transformative uses that prioritize informational utility over expressive content. The fact that Judge Leval, the doctrine's architect, authored the *Google Books* decision closes a critical jurisprudential loop, ensuring his expansive vision of transformation was firmly applied in the context of digital technology and public access to knowledge.

## 4. The Reproduction-Transformation Nexus: Scholarly Debates and Doctrinal Challenges in Copyright Law for Generative AI Training

The complexity of applying fair use to AI training is underscored by the profound economic and cultural stakes involved. Developers warn that requiring licenses for the vast volume and diversity of content necessary for cutting-edge systems is practically impossible, potentially throttling technological innovation. Conversely, creative communities fear that unauthorized training corrodes the creative ecosystem by using artists' works against their will to produce substitutive content. This impasse suggests that fair use, which relies on an ex-post and uncertain judicial inquiry , is being utilized as a proxy to resolve fundamental disputes over

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

resource allocation and cultural policy, a function for which it is doctrinally strained. The scholarly debate is thus centered on two primary questions: whether AI training constitutes a *prima facie* infringement, and how the technical processes of machine learning influence the application of the four fair use factors.

### 4.1 Bracha's Structural Challenge: Non-Infringement Ab Initio

Oren Bracha offers a foundational challenge that rejects the necessity of relying on the fair use defense altogether. Bracha posits that non-expressive training copies do not infringe copyright from the outset (ab initio) because they do not utilize copyrightable subject matter. Bracha's argument relies on the Spillovers Principle, a structural feature of modern copyright law that strictly limits protection to the expressive forms of an information good. This principle requires that functional elements, knowledge, and informational content must "spill over" into the public domain. Since GenAI training uses works solely as a source of information and knowledge for the machine—not for human expressive consumption—the copying lacks the requisite copyrightable subject matter, making the reproduction non-infringing. Bracha criticizes the conventional view that relies on fair use as "copy-fundamentalism," arguing that it is distorted by "confused physicalism"—focusing on the mere physical fact of a copy being made without assessing the use of the protected expression within that copy. This structural approach is intended to provide ex ante legal certainty, preventing copyright from being weaponized to address competitive or cultural policy concerns that the doctrine is fundamentally ill-equipped to handle.[28]

Bracha argues that fair use is poorly suited to address AI training because of both procedural and conceptual flaws. Procedurally, fair use is an affirmative defense, placing the burden on defendants, and its four-factor test is notoriously open-ended, fact-intensive, unpredictable, and expensive to litigate, which creates chilling effects and favors large, well-resourced actors. More fundamentally, fair use is conceptually the wrong tool: it is a back-end doctrine meant to excuse otherwise infringing uses, whereas AI training copies do not infringe at all under the "spillovers principle," since they do not implicate the expressive value that defines copyrightable subject matter.[29] In Bracha's view, the exemption for training belongs in subject-matter limitations at the "front end" of copyright, not in the fair-use framework.

[28] Oren Bracha, *The Work of Copyright in the Age of Machine Production*, 38 Harv. J.L. & Tech. 173, 201 (2024)
[29] *Id*. at 205

### 4.2 The Evolution of the Non-Expressive Use Paradigm (Sag)

Matthew Sag's scholarship provides the most established framework for defending AI training through fair use, grounding it in the doctrine of non-expressive use. This doctrine was previously successfully applied to technologies like search engines and reverse engineering. Sag argues that AI training is highly transformative because the model's ingestion process involves a complex mechanism of "decomposition, abstraction, and remix". The models do not memorize and reproduce original expression; rather, they "learn" latent features and statistical associations within the data.[30]

This perspective frames the training stage as extracting unprotectable elements (ideas, facts, information, and statistical patterns) from the copyrighted works. The resulting non-expressive use has immense social value, enabling computational linguistics, automated translation, and search engines. For Sag, the crucial point is the functional difference between the original purpose (human consumption of expression) and the new purpose (machine learning).

Sag counters Bracha's claim that non-expressive copies made for AI training fall outside copyright's subject matter by arguing that history, statutory text, and practical concerns all point the other way. He recalls Congress's response to *White-Smith v. Apollo*[31], where the Court had held that player-piano rolls were not "copies" because they were unreadable by humans. Congress rejected this logic in the 1909 Act, clarifying that fixation in a machine-readable format capable of reproducing a work suffices for a "copy." The modern Copyright Act preserves this stance, making clear that non-expressive or machine-oriented reproductions still count as copies. Sag further points out that excluding non-expressive copies would be incoherent in areas like software copyright, since software is primarily functional, and piracy often involves using copies for non-expressive, utilitarian ends. If such use were not considered reproduction, then software piracy would not amount to infringement, contradicting Congress's explicit decision to protect software. Finally, he stresses that treating non-expressive copying as categorically outside copyright would not simplify matters: it would only shift the same difficult questions into definitional disputes about whether conduct counts as copying at all. Instead, Sag argues, the fair-use framework, with its flexible, context-sensitive balancing, is the better doctrinal tool to handle these complexities.[32]

[30] Matthew Sag, *Copyright Safety for Generative AI*, 61 Hous. L. Rev. 295 (2023).
[31] 28 S. Ct. 319 and 52 L. Ed. 655
[32] Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 J. Copyright Soc'y of the U.S.A. 291, 10 (2019).

### 4.3 Ard's doctrinal proposal: focus on non-authorial vs authorial value

B.J. Ard's framework in *Copyright's Latent Space* critiques the traditional Factor One approach to fair use, emphasizing that it is ill-suited for generative AI. Ard argues that the conventional transformative purpose test fails because AI systems often have indeterminate purposes at the time of training, given the multi-party "generative-AI supply chain" in which models may be repurposed by different actors. Since the ultimate use of a model cannot be predicted when it is trained, evaluating legality based on Factor One alone is fundamentally flawed, necessitating a market-oriented analysis of how AI interacts with copyrighted works.

Ard proposes examining "copyright's latent space" to distinguish between authorial and non-authorial value, thereby clarifying what constitutes legitimate competition. Authorial value derives from the unique creative expression of a work, whose unauthorized exploitation constitutes harmful market substitution. Non-authorial value, such as ideas, conventions, or tropes, has historically been open for use by others even if it disadvantages copyright holders. Ard refocuses the Factor Four market harm inquiry to determine whether AI is exploiting protected creative choices or merely drawing on unprotected sources, aligning the analysis with the economic stakes of authorship.

Ard defines authorial value as the cultural and economic worth embedded in the author's original expressive choices, exemplified in prose, melody, or artistic arrangements. When AI systems replicate these elements, they substitute for the author's expression and should weigh against fair use. Contrarily, Bracha challenges Ard's approach, asserting that both the communicative content of works and the meta-knowledge of expressive skills have always been outside copyright's protection. He emphasizes that copyright law intentionally allows knowledge spillover: learning from existing works to create new, competing works, whether by humans or machines, remains free and essential for cumulative creativity, and does not constitute infringement.

### 4.4 Mark Lemley & Bryan Casey: The Doctrine of Fair Learning (Prioritizing Factor 1)

The central premise of *Fair Learning* is that the massive copying essential for ML systems is a non-expressive use, transforming the purpose for which the content is utilized, even if the content itself remains physically unaltered. This rationale is tethered to the foundational copyright principle that law should never grant protection over the underlying ideas, facts, or functions of a work—the "idea-expression dichotomy". While Fair Learning provides a broad defense, the authors recognize that the protection cannot be absolute. The doctrine's efficacy is

contingent upon the developer's intent and the nature of the output.

The primary limitation on Fair Learning is triggered when "learning is done to copy expression". Specifically, if a developer trains an ML system "to make a song in the style of Ariana Grande," the fair use question becomes "much tougher". In such a case, the system is deemed to be copying expression for the sake of expression, thereby posing a threat of "significant substitutive competition" within the original author's expressive market.[33]

The practical difficulty in applying this limitation stems from distinguishing between non-expressive technical learning and expressive style appropriation. This distinction places a heavy burden of demonstrating non-expressive intent on the developers, particularly as advanced generative tools blur the line between extracting neutral linguistic structures and appropriating a unique creative style. The input-centric nature of Fair Learning requires policing this intent boundary, an undertaking that can become tenuous as the link between input data and complex generative output grows increasingly opaque.

### 4.5 Other scholarly opinions

Benjamin Sobel, argues that the non-expressive use doctrine, a cornerstone of fair use for data mining is fundamentally threatened by the rise of sophisticated expressive machine learning (ML), which includes fields like natural language generation. Sobel's core contention is that when ML models generate output that is itself highly expressive and functionally equivalent to copyrightable human work (e.g., writing prose or composing music), the use of the copyrighted training data can no longer be defended as purely "non-expressive" or sufficiently transformative under the first fair use factor. Furthermore, this expressive capacity presents a new, direct threat of market substitution under the fourth fair use factor because the AI's output is designed to compete directly with the original works and the markets of their creators, essentially diverting rightful earnings. Sobel warns of a dilemma: either courts reject fair use for expressive ML, thus halting innovation, or they accept it, thereby disenfranchising human creators and allowing powerful firms to capture value at the expense of individual rights holders, suggesting the fair use doctrine may no longer serve its historical purpose of fostering public expressive activity.[34]

Sag critiques Sobel's argument, maintaining that the ability of Large Language Models (LLMs) to produce quasi-expressive works does not negate the application of the non-expressive use

[33] Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 Tex. L. Rev. 743, 750 (2021).
[34] Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 53–54 (2017).

principle to the training process. Sag asserts that the true measure for fair use is not the expressiveness of the output, but whether the original expression in the training data is being communicated to a new public. Since an LLM's output typically bears no substantial similarity to any particular copyrighted work in its massive training data, it only learns unprotectable patterns and styles and the use remains transformative.[35]

Michael Carroll's work, focused on maximizing the operational scope of TDM, suggests that certain temporary, intermediate steps in TDM research may not create copies that "count" under U.S. law, arguing that it is possible to structure cloud-based TDM research to entirely avoid implicating copyright law. A related debate concerns the influence of illegal sourcing on the fair use defense. The U.S. Copyright Office has suggested that the use of pirated copies should "weigh against fair use without being determinative". Carroll, however, takes a strong stance that the transformative benefits of the research must be prioritized. He contends that even if courts consider the good faith or lack thereof in acquiring the data, TDM research conducted on infringing sources, such as shadow libraries, should still be deemed lawful. This conclusion holds because the TDM provides significant transformative benefits without causing harm to the markets that actually matter for copyright purposes.[36]

Edward Lee advocates for a "Technological Fair Use," recognizing the need to balance the legal necessity of access for innovation with the protection of authorship.[37] Lee anticipates a "fluid" period where courts must reconcile novel AI concepts with existing legal rules.[38] He stresses that the fair use analysis must consider not just the immediate intermediary purpose of training the AI but also the ultimate purpose of the creation of new works. Given the likelihood of varied judicial outcomes across the numerous pending lawsuits, Lee foresees an eventual need for legislation to unify the doctrine and control AI to serve humanity.[39]

**5. A Tale of Three Cases: Fair Use, Infringing Inputs, and the Future of Generative AI**

The evolving jurisprudence surrounding copyright law in the context of generative AI highlights the tension between the traditional scope of copyright protection and the transformative capabilities of large language models (LLMs). Central to this discourse are

[35] Matthew Sag, *Copyright Safety for Generative AI*, 61 Hous. L. Rev. 295, 308 (2023).
[36] Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 UC Davis L. Rev. 893 (2019).
[37] Edward Lee, *Technological Fair Use*, 83 S. Cal. L. Rev. 797 (2010).
[38] Maren Hendricks, *Edward Lee: Copyright in the Age of AI Acceleration*, BYU L., Apr. 10, 2024, https://law.byu.edu/news/edward-lee-copyright-in-the-age-of-ai-acceleration. (as accessed on 3rd October 2025)
[39] Ibid.

cases such as *Andrea Bartz v. Anthropic*[40], *Richard Kadrey v. Meta Platforms, Inc.*,[41] and *Thomson Reuters v. Ross Intelligence*[42], each illustrating divergent judicial approaches to the concepts of transformative use and the legality of infringing copies in AI training. Collectively, these cases underscore the doctrinal and practical challenges in reconciling fair use with the unprecedented scale of automated data ingestion.

In *Bartz v. Anthropic*, authors, including Bartz, brought a class action against Anthropic, alleging that the company's Claude LLM unlawfully reproduced substantial portions of their copyrighted books, including quotations, summaries, and character dialogues. The plaintiffs argued that the model's outputs constituted derivative works, adversely affecting the market for their original publications. Similarly, in *Kadrey v. Meta*, the plaintiffs alleged copyright infringement based on Meta's use of their works, some sourced from shadow libraries, to train its Llama LLMs. Both cases foreground the central question of whether ingestion of copyrighted material for AI training constitutes fair use, and whether the LLM outputs themselves could be considered infringing derivatives. A distinguishing feature in *Kadrey* is the explicit involvement of illegally obtained copies, raising the additional question of whether the lawfulness of the input material conditions the applicability of fair use.

In *Thomson Reuters v. Ross Intelligence* concerned the reproduction of highly structured legal materials, specifically West Publishing's headnotes and key numbers. The court found that Ross's AI product directly replicated the expressive organization of Westlaw's content, offering a competing service, and therefore did not qualify as transformative. This case underscores the principle that transformativeness is context-dependent: the mere use of an AI to ingest content does not automatically satisfy fair use, particularly when the output mirrors the structure and organization of the copyrighted work. Here, the court emphasized that the purpose and character of the use must meaningfully diverge from the original to be transformative.

A fundamental deficiency in the *Thomson Reuters v. Ross Intelligence* ruling resides in the Court's failure to recognize the inherently non-expressive and instrumental nature of the copied headnotes. The Court effectively presumed that any reproduction of copyrighted material constitutes *prima facie* infringement, thereby disregarding the well-established principle that mere copying is not unlawful when the use is functional rather than intended to replicate or

[40] *Bartz et al. v. Anthropic PBC*, No. 3:24-cv-05417-WHA (N.D. Cal. filed Aug. 19, 2024).
[41] *Kadrey et al. v. Meta Platforms, Inc*., No. 3:23-cv-03417-VC (N.D. Cal. filed July 7, 2023).
[42] *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc*., No. 1:20-cv-00613-SB (D. Del. filed Feb. 11, 2025).

supplant the original creative expression. In this instance, the headnotes were utilized exclusively as data inputs to train Ross Intelligence's AI system, facilitating pattern recognition, syntactic analysis, and the development of sophisticated search algorithms. The Court's analysis conflated expressive reproduction with utilitarian deployment, overlooking the transformative, non-substitutive character of the copying. In failing to distinguish between content consumed for aesthetic or cognitive purposes and content employed instrumentally for technological processes, the decision risks imposing undue constraints on innovative applications of copyrighted works, potentially inhibiting advancements in AI and other fields that rely on non-expressive, functional uses of pre-existing material.

In *Bartz* and *Meta*, courts have tentatively acknowledged that training an LLM constitutes a highly transformative act. Judges have reasoned that an LLM does not reproduce copyrighted works verbatim but abstracts statistical patterns, syntactic structures, and semantic relationships to generate novel content. This aligns with the Supreme Court's articulation of transformative use in *Campbell v. Acuff-Rose Music, Inc.*, where purpose and character are central to fair use analysis. The courts recognize that the AI's purpose, pattern extraction and generation of new text, differs fundamentally from the original literary intent, favouring a fair use finding under the first statutory factor.

The critical divergence arises regarding the legality of the underlying copies. In *Bartz*, Judge Alsup suggested (obiter) that unlawful acquisition of training material such as through shadow libraries could undermine a fair use defense, emphasizing the "initial lawfulness" of the input. Conversely, in *Meta*, Judge Chhabria rejected the notion that the initial illegality precludes fair use, arguing that the transformative purpose of AI training subsumes the act of copying. Here, the courts effectively prioritize the ultimate public benefit derived from the model's outputs over the technical infringement of the source acquisition. The Meta decision also mitigated economic harm concerns, noting that shadow libraries generate negligible market competition for legitimate licensing, reducing the impact under §107(4).

Theoretical underpinnings from scholars further illuminate these judicial approaches. Pierre Leval, the progenitor of the modern transformative use doctrine, emphasizes that fair use exists to advance science and the useful arts, largely independent of moral or technical considerations regarding source acquisition. Judge Chhabria's reasoning aligns closely with Leval, prioritizing the transformative end use. Michael Carroll similarly contends that highly transformative uses, especially those serving public interest, should not be defeated by the infringing nature of

copies, reflecting a utilitarian approach that undergirds the Meta judgment.

## 6. Achieving Predictability and Uniformity in Global Text and Data Mining Exceptions for Generative AI

The exponential growth of generative artificial intelligence (GenAI) technologies, intrinsically reliant on the systematic ingestion and analysis of vast, cross-border datasets (Text and Data Mining, or TDM), has exposed a profound incompatibility between global technological development and fragmented national copyright laws. The current environment of legal divergence, coupled with the inherent unpredictability of open-ended doctrines, translates directly into a high risk profile for innovation. This environment disproportionately benefits established incumbents capable of absorbing massive litigation costs, stifling the participation of small and medium-sized enterprises (SMEs) and hindering broader market competition.

### 6.1 The Policy Crisis in AI Training Data and the Regulatory Gap

Generative AI requires a foundational step, which is TDM, that utilizes large bodies of intellectual property (IP)-protected works. The legality of this activity is currently determined by local, non-harmonized national copyright laws. This fragmentation creates systemic liability, as a single, globally deployed Large Language Model (LLM) must simultaneously comply with dozens of disparate national regimes. Consequently, developers face significant legal uncertainty regarding what constitutes permitted TDM activity, leading to a climate of fear regarding potentially "massive law suits".[43]

### 6.2 Thematic Overview of Jurisdictional Divergence (The Three Regulatory Models)

A comparative analysis of leading text and data mining (TDM) regimes reveals three distinct normative models, each reflecting different underlying philosophies regarding intellectual property rights, innovation policy, and the balance between data access and protection of rightsholders. These models—statutory privilege, conditional compliance, and the licensing imperative—are not merely technical variations but rather reflect deep-seated differences in regulatory priorities and socio-economic contexts.

### 1. Statutory Privilege (Japan)

Japan exemplifies the statutory privilege model through one of the world's most permissive

[43] Juan-Carlos Fernández-Molina & Fernando Esteban de la Rosa, *Copyright and Text and Data Mining: Is the Current Legislation Sufficient and Adequate?*, 24 Portal: Librs. & Acad. 653 (2024). https://preprint.press.jhu.edu/portal/sites/default/files/12_24.3fernandez.pdf (as accessed on 3rd October 2025)

frameworks for TDM. Article 30-4 of the Japanese Copyright Act provides that copyrighted works may be exploited for data analysis purposes regardless of the purpose or commercial intent, so long as the use does not amount to an infringement of the normal exploitation of the work or unjustifiably prejudice the legitimate interests of the author.[44] This broad statutory exception maximizes legal certainty for developers and significantly reduces transactional barriers to data-driven innovation. By decoupling TDM from rightsholder consent and licensing markets, Japan has positioned itself as a jurisdiction prioritizing velocity of data access over proprietary control, thereby signaling a pro-innovation and competitiveness-oriented policy approach.

**2. Conditional Compliance (The European Union)**

The European Union has adopted a more complex and qualified model, grounded in a bifurcated regime under the Directive on Copyright in the Digital Single Market (DSM Directive).[45] Article 3 establishes a mandatory exception for TDM conducted for scientific research purposes by research organizations and cultural heritage institutions. Article 4, however, extends a broader exception to all users, including commercial entities, while simultaneously allowing rightsholders to opt out by expressing their reservation of rights in machine-readable form.[46] This dual structure creates an environment where the enforceability of TDM rights depends not only on statutory law but also on compliance with technical and metadata auditing mechanisms.[47] The EU model thereby introduces a conditional, compliance-driven framework that attempts to balance innovation incentives with the autonomy of rightsholders, but in practice imposes greater compliance burdens and legal uncertainty on AI developers compared to the Japanese model.[48]

**3. The Licensing Imperative (The United Kingdom)**

The United Kingdom reflects a licensing-centric paradigm. Following Brexit, the UK government declined to adopt the EU's broader TDM exception, maintaining instead the narrower exception under Regulation 3 of the Copyright and Rights in Databases Regulations

[44] Copyright Act of Japan, Act No. 48 of 1970, art. 30-4, amended by Act No. 30 of 2018.
[45] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 Apr. 2019 on Copyright and Related Rights in the Digital Single Market, 2019 O.J. (L 130) 92.
[46] Id. arts. 3–4.
[47] The New Copyright Directive: Text and Data Mining Articles 3 and 4, Wolters Kluwer Copyright Blog (July 24, 2019), https://legalblogs.wolterskluwer.com/copyright-blog/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ (the bifurcated system of EU CDSM Directive Articles 3 and 4, including scope, opt-out, and lawful access).
[48] https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf

2014, which permits TDM exclusively for non-commercial research purposes.[49] Consequently, commercial TDM, including AI training remains subject to licensing and remuneration requirements. This effectively operationalizes a market-based solution, privileging rightsholders' ability to monetize the use of their works in AI development while raising costs and legal barriers for startups and commercial entities seeking to scale foundational models.[50]

### 6.3 The Imperative for Legal Certainty and Uniformity

For the global AI economy to thrive sustainably, two conditions are paramount: legal certainty and uniformity across jurisdictions. Legal certainty, established through predictable statutory exceptions rather than reactive judicial defenses, is necessary to transition GenAI development from a high-stakes, litigious endeavor reserved for well-resourced incumbents to a predictable technical activity accessible to small entrepreneurs. By minimizing transactional costs and liability risk associated with data acquisition, statutory certainty functions as an instrument of industrial policy, promoting competitive market access.

Furthermore, uniformity across jurisdictions is essential to address the principle of lex loci protectionis, the law of the country where the call for protection is invoked. Without a harmonized global TDM exception, a single LLM trained on a global corpus remains exposed to perpetual infringement claims based on the most restrictive copyright laws worldwide. This situation demands multilateral policy intervention to ensure cross-border legality and adherence to international intellectual property norms.

### 6.4 The Unpredictability of Open-Ended Doctrines (The US Fair Use Conundrum)

Unlike many international jurisdictions that have legislated explicit TDM exceptions, the United States lacks such a provision, relying instead on the doctrine of Fair Use (Section 107 of the Copyright Act of 1976). While the flexibility of Fair Use has historically allowed it to adapt to unforeseen technological challenges, its application to LLM training remains highly contested and ambiguous.

The core difficulty lies in the subjective, multi-factor test of Fair Use, particularly as applied to the mass ingestion of copyrighted works for model training. The resulting legal environment

[49] https://www.hsfkramer.com/notes/ip/2023-03/uk-withdraws-plans-for-broader-text-and-data-mining-tdm-copyright-and-database-right-exception

[50] Training AI Models: UK Government Proposes EU-Style Opt-Out Copyright Exception, DLA Piper (2025), https://www.dlapiper.com/en-us/insights/blogs/mse-today/2025/training-ai-models-uk-government-proposes-eu-style-opt-out-copyright-exception (the UK CDPA Section 29A, non-commercial constraint, lawful access, and the abandoned policy proposal).

is characterized by significant litigation risk. Recent judicial decisions regarding training data, such as those concerning *Bartz v. Anthropic PBC*[51] and *Kadrey v. Meta Platforms, Inc.*,[52] underscore that Fair Use decisions are inherently "highly fact-specific". Courts meticulously analyze the specific facts of each case, including the extent to which the use is transformative and the presence of infringing outputs or market impact.

This necessity for a case-by-case assessment means that developers cannot rely on any single judicial ruling to establish broad legal certainty across their entire global operational model. The litigation risk associated with defending against fact-specific claims functions as a significant regulatory barrier, forcing developers to operate under the continuous threat of existential legal challenges. The automation and scale of GenAI training—which involves the simultaneous creation of billions of transient copies for analysis, require a pre-emptive, certain legal right (a statutory exception) rather than a reactive legal defense (Fair Use) to ensure operational continuity and stability.

## 7. Conclusions and Suggestions

Artificial intelligence represents one of the most transformative technological forces in human history, with immense potential to augment human capabilities, accelerate innovation, and expand the frontiers of knowledge. Its applications in science, medicine, education, and creative industries are already demonstrating unprecedented gains in efficiency and problem-solving capacity. Encouraging the development and progress of artificial intelligence is therefore not merely a technological imperative but a societal necessity.

At the same time, the disruptive nature of generative AI has given rise to legitimate anxieties. These include the displacement of traditional sources of income, the narrowing of opportunities to engage in expressive and creative activities, and the potential weakening of platforms that sustain creative production. These concerns cannot be dismissed; they are genuine and demand attention. However, they are best addressed through targeted policy instruments—such as labor market reforms, cultural subsidies, or revenue-sharing mechanisms—rather than through the blunt instrument of copyright law.

The rapid proliferation of generative artificial intelligence has placed existing doctrines, jurisprudence, and legal frameworks under considerable stress. As shown in the preceding analysis, traditional copyright principles—developed for an era of human-centered creation

[51] *Supra* note 40.
[52] *Supra* note 41.

and analog reproduction—are increasingly ill-suited to address the non-expressive, large-scale, and automated uses of works that underpin AI training. Courts and legislatures are struggling to reconcile the rights of authors with the technological imperatives of innovation, often producing fragmented and inconsistent approaches across jurisdictions. This legal uncertainty underscores the urgent need for a robust, globally harmonized framework that evolves in tandem with technological progress while ensuring that authors' interests and creative incentives remain protected.

Copyright, when overstretched into regulating machine learning processes, risks both stifling innovation and undermining its original goal of promoting creativity. A coherent, internationally consistent legal regime for text and data mining (TDM) is urgently needed—one that provides predictable rights of access while respecting the legitimate interests of authors. Such a framework must contain harmonized TDM exceptions and limitations, encourage lawful reuse of data, and avoid the fragmentation of rules across jurisdictions. Needless to say, authors and their creative expressions must be rewarded fairly, even as technology is allowed to flourish in ways that benefit humanity at large.

**Recommendations for a Robust International TDM Regime**

1. **Harmonized TDM Exceptions**
   Adopt a baseline international standard on TDM exceptions under copyright, ensuring legal certainty for developers and preventing jurisdictional fragmentation.
2. **Non-Expressive Use Carve-Out**
   Explicitly distinguish between expressive uses (e.g., reproductions that substitute for works) and non-expressive uses (e.g., algorithmic training) to clarify the scope of permissible TDM.
3. **Global Licensing Framework**
   Create a cross-border, interoperable licensing system for datasets to reduce transaction costs and facilitate access for developers operating in multiple jurisdictions.
4. **Compulsory Licensing Mechanisms**
   Introduce narrowly tailored compulsory licensing provisions for cases where private licensing obstructs innovation that serves broad public interests.
5. **Open Data Mandates for Publicly Funded Works**
   Require that works and datasets funded through public resources be made accessible for TDM under open licenses, subject to safeguards for privacy and national security.

6. **Data Provenance and Transparency Standards**
   Mandate traceability of datasets used in AI training to allow auditability and ensure lawful sourcing, while respecting commercial confidentiality.
7. **International TDM Registry**
   Establish a global registry of datasets used for AI training, operated by a neutral international body, to facilitate attribution, licensing, and monitoring.
8. **Time-Limited Embargoes**
   Permit rights-holders temporary embargo periods before their works are made TDM-accessible, ensuring a balance between immediate commercial exploitation and long-term public benefit.
9. **International Dispute Resolution Mechanisms**
   Establish cross-border arbitration and mediation procedures to resolve conflicts efficiently, avoiding inconsistent outcomes across national courts.
10. **Safe Harbor for Non-Profit Research and Education**
    Grant statutory protection for universities, public research institutions, and non-profit entities engaged in TDM for socially beneficial purposes.

**Recommendations for Protecting Authors' Interests**

1. **Revenue-Sharing Mechanisms**
   Establish collective rights management systems to distribute royalties generated from commercial AI systems trained on copyrighted works.
2. **Attribution Requirements**
   Require disclosure and, where feasible, attribution of authors whose works form part of training datasets.
3. **Tiered Access Models**
   Grant broader TDM access rights for research and educational institutions, while requiring commercial developers to obtain licenses or pay remuneration.
4. **Transparency Obligations for AI Developers**
   Mandate disclosure of categories of works or datasets used in training, without compelling disclosure of sensitive proprietary details.
5. **Creative Incentive Funds**
   Create global funds, financed by levies on commercial AI developers, to support authors, cultural industries, and artistic innovation.

6. **Opt-Out Mechanism**

   Provide an international "opt-out registry" for authors who do not wish their works to be used in AI training datasets.

7. **Platform Accountability**

   Impose due diligence obligations on AI companies and content platforms to ensure that datasets are lawfully sourced and licensed.

8. **Public Interest Safeguards**

   Preserve unrestricted TDM access for journalism, education, research, and other activities essential to democratic and cultural development.

9. **Monitoring and Review Committees**

   Establish multi-stakeholder oversight bodies at the international level to review the impact of TDM and AI on creative economies.

10. **Balanced Enforcement Mechanisms**

    Ensure enforcement measures are proportionate, focusing on systematic infringement by commercial actors rather than incidental or small-scale TDM activity.