THE LEGALITES LEXSCRIPTA
ISSN 3108-2416 (ONLINE)
Editor-in-Chief: - Prof. (Dr.) Aryendu Dwivedi
Volume II Issue II (April-June) Page No.: - 163 to 172

Beyond Fair Use: Why AI Training Needs Mandatory Licensing

Shefali Singh

Abstract: This chapter talks about the idea of "ingestion crisis" in generative intelligence and what are its impacts on copyright law. This chapter mainly argues that the earlier assumption, which says that the AI developers could freely use the publicly available data to train AI under the disguise of the doctrine of fair use, is no longer convincing. In cases such as *Authors Guild v. HathiTrust*, *Harper & Row v. Nation Enterprises*, and *Andy Warhol Foundation v. Goldsmith*, and also involving the recent lawsuits involving ai companies, the chapter shows that courts are now paying closer attention to the economic impact of AI. In particular, when AI-generated outputs begin to compete with or replace original human works, the defence of fair use becomes weaker.

The chapter further highlights the practical difficulty of relying on individual licensing agreements, since it is nearly impossible for AI companies to negotiate with thousands or even millions of copyright holders. To address this issue, the chapter suggests Mandatory Collective Licensing as a more practical solution. This system, similar to those used in the music industry, enables a central organisation to oversee licences and allocate royalties to creators. It also examines India's proposed "One Nation, One Licence" scheme and compares it with the European Union's framework, which emphasises transparency and disclosure.

At the same time, the chapter acknowledges that collective licensing raises important concerns, especially regarding moral rights, consent, and the ability of creators to opt out. It suggests that these concerns can be addressed by introducing safeguards such as tiered licensing, strong transparency standards, and meaningful opt-out provisions. In conclusion, the chapter argues that a balanced and structured licensing framework is necessary to support the growth of AI while ensuring that creators are fairly protected and compensated.

Keywords: Generative Artificial Intelligence; Copyright Law; Fair Use Doctrine; Mandatory

Collective Licensing.

1. INTRODUCTION: THE INGESTION CRISIS

The era of free data to train AI is soon coming to a close. During the past decades developers had assumed that the huge amount of publicly available internet data to be used in AI learning constituted a non-expressive transformation, and thus should be excluded under the fair use exemptions as the processing of books by indexing and digitising them qualifies as transformative use.[1] Similarly, Authors Guild v. HathiTrust[2] was a case that created a searchable database to be used in research that was founded on scanning books and this was held to be non-expressive transformative use which constituted fair use of the work.[3] However, the cases mentioned above assumed explicit and restricted purposes of massive processing of copyrighted content, whereas the modern technological systems have to process enormous amounts of different types of content in order to learn, and this practice can be commercialised and not authorised by the creators of the content.

Conversely, due to the fast growth of the Generative AI technology, the available resources of high-quality data have been exhausted, which has led to what legal theorists have called the Ingestion Crisis.[4] It is not only a data supply problem, but also one of access to quality, reliable and legally appropriate data. As numerous sources are either put behind paywalls or licensed to third parties, the amount of sources that are freely available is declining. In the meantime, the demand of data is also increasing as the newer models are bigger in terms of data they need to operate. This poses a dilemma between the necessity to have technological advancement and the necessity to have legality. Also, the wide application of artificial intelligence products has led to a situation where most information being consumed is synthetic and it is hard to differentiate between authentic and counterfeit information. It has now become feared that there might be contamination of the data, where the new model is modeled on the results of the older model, thereby undermining the validity of the results.

Nevertheless, the principal problem is that it is inherent in the nature of AI outputs. Compared to the past transformative uses of the copyrighted works directed at the production of a completely new product that utilizes the old work, the current production of outputs that directly replace original works created by humans by the AI models is problematic in both the

[1] Samuelson, P. (2017). *Unbundling fair uses*. Fordham Law Review, 77(5), 2537–2621.
[2] *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014).
[3] *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015)
[4] Villalobos, P., et al. (2022). *Will we run out of data? Limits of LLM scaling based on human-generated data*. Epoch AI.

application of the principle of fair use/dealing and market substitution specifically. The courts have generally taken a favorable view on the issue of market substitution, as is apparent, in the case of Harper and Row v. Nation Enterprises[5] where the court considered unauthorized use of copyrighted work to be a violation because of its impact on the market of such works. Currently, the outputs generated by AI also can unseat works created by humans in terms of market share, thereby losing their economic rights to creators. The threat has become a reality with recent lawsuits filed by Getty images against Stability AI and The New York Times against open AI.

Therefore, the concept of the doctrine of Fair Use is no longer relevant in protecting the developers against meritless claims. The increasing number of these litigations indicate a shift in the judicial and regulatory bodies attitude, which are now more cautious about the use of traditional exemptions to new technological situations. The previous idea that data consumption is only an analysis task is being questioned, particularly when the output obtained is directly correlated with the input data. Also, these technologies have a business side, thereby complicating their fair use defense during court proceedings. In India, though fair dealing is provided under the Copyright Act, 1957, in Section 52[6], the scope is rather limited and mass consumption of data is not explicitly addressed.

In our research paper, we will consider how Mandatory Collective Licensing can be implemented in order to resolve the tension that currently exists between the AI industry, which requires data and the creative economy, which requires financial incentives. Introduction of collective licensing to address the ingestion crisis is necessary as it is one of the ways through which the issue can be resolved amicably by ensuring that it is feasible to access huge data sets yet still generate revenues on the creative individuals. The success of such a mechanism will to a large extent hinge on the way it is implemented. With increasing less accessibility of data and potential legal threats, there is a necessity to shift towards licensing arrangements.

2. WHY THE FAIR USE SHIELD HAS FAILED.

2.1. Transformation vs. Substitution

An idea that the Fair Use principle of copyright laws has is the idea of transformation, meaning the use of copyrighted content in a new purpose that has a different purpose, character or

[5] *Harper & Row, Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539 (1985).
[6] The Copyright Act, 1957, § 52 (India).

meaning but does not replicate the original.[7] Within previous rulings of the American courts, transformation has been considered to embrace technological advancements whereby copying was permissible in cases where the copied material assisted in facilitating a new functionality such as searching, indexing or analysis. This interpretation of transformation has changed recently, especially regarding generative technologies, in which courts have held that the use of AI algorithms trained on an artist's works and generating art similar to those will no longer constitute a mere transformation since it replaces the artist's place in the market.[8]

This transformation/substitution dichotomy has also been significant in the recent debates about copyright. Transformation is a process whereby an addition has taken place that does not result in any harm to the market of the original product. Substitution on the other hand, occurs when the product that has been produced serves a similar purpose to the original product and serves as its substitute. The results of an AI system can serve as direct replacements of commissioned works by a human being when the results are reflective of recognizable artistic styles and moods or other creative elements of the work.

The recent changes in judicial precedent have also been a positive move towards this strategy. In Andy Warhol Foundation of the Visual Arts, Inc. v. Goldsmith[9], the Court narrowed down the scope of transformative use by emphasizing that although an aspect of transformation might be present in the production of a new art object, the same would still be classified as infringement in case the economic purpose was identical to the original one. This argument is particularly relevant when it comes to AI-generated products as they can differ in shape, yet the purpose of either cannot be regarded as that different.

This growth implies that the difference between what is permissible change and what happens to be the forbidden replacement will be blurred shortly. As technologies have been developed that could create work in a style that is identifiably similar, there is a risk that these will create unfair competition and attack the value of creative work. It challenges the concept that learning and creation of content is purely an analytical activity, but instead has practical economic implications on the creator of the original.

2.2. The Court Case: Bartz v. Anthropic (2025).

[7] Leval, P. N. (1990). *Toward a fair use standard*. Harvard Law Review, 103(5), 1105–1136.
[8] Gervais, D. (2023). *AI and copyright: The training problem*. Journal of Intellectual Property Law.
[9] *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508 (2023).

This pattern reversed to the side of the owners of copyright in the 2025 cases, such as Bartz v. Anthropic.[10] In this case, the court decided in favor of the owner of the copyright by stating that free training was not applicable in cases whereby the degree of ingestion has resulted in perceivable damage to the major market of the copyright holder. It is crucial to mention that there had been a hope that a training on massive scale with the use of copyrighted materials would be regarded under the fair use. The court instead said that this act cannot be considered as legal or not depending on the activity being transformative and non-expressive but instead by its economic connotations. In the case where the scope of consumption can be used to generate products that can be used to replace the original, the validity of free training is doubtful.

Another important factor highlighted in this decision was the importance of market impact to the determination of the outcome of the copyright test, which helped even more to underline the fact that even the most advanced applications should not ignore the economic interests of the creators. By not accepting the notion that technological progress could be used in defence of any form of infringement, the court implied that the court would not accept the notion that the negative impacts of massive unlicensed exploitation could be used as a defence against any form of infringement. This is in line with the current trend in the judiciary in determining the economic value of information technologies.

Thus, the case law that has been established will definitely send the right signal to the software developers, and the policy makers, that whatever degree of legal acceptance of any form of data scraping that we have currently is shrinking by the day. The shift towards the former Wild West scenario when masses of Internet content could be gathered and utilized freely and without much difficulty is evident. Instead, it appears that the time has now come when there needs to be some structure to the availability of any copyrighted material.

2.3. The Logistical Impasse

The practical standpoint of an individual licensing strategy, requiring the AI firm to conclude agreement after agreement with individual artists, writers or rights owners, is viewed as unrealistic. It is hard to envisage how a company would be able to find out information about potentially millions of owners and enter into negotiations with them due to the sheer volume

[10] Bartz v. Anthropic PBC, 787 F.Supp.3d 1007 (N.D. Cal. 2025)

of information that needs to be processed while training LLMs.[11] In addition, such an endeavor would cost significantly more than the licenses themselves, making the entire idea economically impractical.

Making the issue even more problematic is the disunified nature of the ownership of the information. The owners of some data may be a number of people, such as publishers or authors. Besides, when it comes to the orphan works, there would be no means of the owner of the AI model to locate the copyright holder.[12] Even in the event that the company resorted to such kind of process it would be very inefficient and would also be highly questionable in the light of law.

This implies that the conventional bilateral contracting takes place, making it a significant entry barrier to any organization in the industry whether big or small. This type of arrangement has a process that limits the access to the data depending on the administrative processes, as opposed to effective legal mechanisms. It is now clear that a new solution to this issue is required, possibly in the form of a collective licensing scheme, in order to have legal access to big data in a fair way.

3. COLLECTIVE LICENSING SOLUTION

3.1. Spotify Model of AI.

Mandatory Collective Licensing (MCL) scheme proposes a so-called blanket-mechanism in addressing the growing complexities that accompany the use of big data. The AI companies will have to pay a fee or a specified percentage of the revenue of the firm directly into the cash kitty of a central governing authority that will have the mandate to guarantee the allocation of the revenue generated by the licensing operations to the creators. Unlike the traditional license agreements, MCL will make it easier for the developers of data technology to have access to a wide repertoire of protected works via a standardized scheme.

The proposed solution will be similar to the currently used system in the entertainment sector specifically the PRS to Music and ASCAP.[13] Similarly to the PROs collecting license fees on

[11] Landes, W. M., & Posner, R. A. (2003). *The economic structure of intellectual property law*. Harvard University Press.

[12] U.S. Copyright Office. (2015). *Orphan works and mass digitization report*.

[13] Towse, R. (2016). *Copyright, collecting societies and music industry*. Journal of Cultural Economics.

various users of the copyrighted work of their members and giving the revenue back to the respective rights holders, the proposed scheme will be based on the principle of centralized administration and collective management in which each creator or right holder does not enter into a negotiation.[14]

A significant advantage of the above strategy is that, it guarantees a consistent and predictable process of acquiring access to data and receiving revenue out of such access. Programmers can safely process such large quantities of data legally without fear of the imminent exposure to legal proceedings, but content creators can be sure that they will get paid to use their work. But the success of such strategy demands that actions of the central agency that is going to control the process should be heavily controlled as otherwise one can hardly guarantee the correct distribution of revenues.

3.2. One Nation, One licence Model in India.

In fact, India has been a significant player in this new policy towards the regimes of structured licensing. In particular, the Department of Promotion of Industry and Internal Trade (DPIIT) Working Papers 202526 suggested the creation of Copyright Royalty Centralized Administrative Trust (CRCAT).[15] The CRCAT will be an authoritative body that will oversee the licensing procedure of big data business, particularly with respect to AI training operations. Based on this structure, AI firms will have to make contributions to the central fund either through a set sum of fee or revenue mode without necessarily having to enter into individual deals with the rights holders.

The idea behind such an approach is quite simple - to be capable of having the rights of Indian authors followed and protected in the cases when their pieces are used in training. This solution will not only create equitable remuneration of creators, but will also provide the much needed clarity of regulations to the developers who will be forced to comply with a set of regulations without necessarily having to enter into complicated individual agreements.

Indeed, India holds a major role in this new policy in policy towards regimes of structured licensing. In particular, the Department of Promotion of Industry and Internal Trade (DPIIT)

[14] WIPO. (2017). *Collective management of copyright and related rights*.

[15] Department for Promotion of Industry and Internal Trade (DPIIT). (2023). *Discussion paper on AI and copyright policy*.

Working Papers 202526 suggested the creation of Copyright Royalty Centralized Administrative Trust (CRCAT).[16] The CRCAT will be a regulatory entity that will undertake the process of licensing of big data business, especially on the operations of AI training. In accordance with this framework, AI companies will be required to contribute to the central fund either by a fixed amount of fee or revenue mode without necessarily requiring to enter to individual agreements with the rights holders.

The concept of this strategy is quite straightforward - to safeguard and guarantee the rights of Indian authors when their work is used to train people. Such a solution will not only ensure that the creators are fairly paid but it will also offer regulatory clarity to the developers that will have to abide by a set of rules, without necessarily making complicated individual arrangements.

3.3. Global Comparisons

The European Union has adopted a revenue-based and disclosure-based system as India moves towards a statutory system of one nation, with a centralized licensing system. The EU AI Act Compliance Guidelines (2026) designed in the framework of the overall context of the EU AI Act requires the authors of high-impact models to maintain and distribute detailed summaries of the copyrighted information used to train. The emphasis on disclosure can be explained by the overall principle of regulation pursued by the European Union, in which transparency is a crucial aspect. In addition to the mandatory disclosure, it has been argued that a revival towards a revenue-sharing system is possible where it is recommended to withhold 5-7 per cent of the total revenue to compensate the rights holders.[17]

The mechanism of the EU licensing system is rather dissimilar to the suggested one in India as the system is not based on the mechanism of a single statutory licensing body. In other words, it includes both regulation and markets, where there is the possibility to reach negotiated agreements and develop sectoral systems of licensing, while at times collective licensing can be applied. On one hand, this system will give a certain freedom regarding the manner in which

[16] Department for Promotion of Industry and Internal Trade (DPIIT). (2023). *Discussion paper on AI and copyright policy*.

[17] European Parliament. (2024). *Artificial Intelligence Act (EU AI Act)*.

the compensation is made but on the other hand, this can bring about fragmentation as different industries will have different standards of compensation.

Furthermore, the EU regime of data usage equally relies on the existence of different copyright directives that have been previously issued within the EU to govern some of the activities that concern data usage. Text and data mining can only be used under some conditions such as legal access or capability of the owners of rights to refuse availing their permission.[18] However, these rules need to be changed since within the rapidly developing generative technologies, they do not govern the instances of commercial application.

The mandatory licensing system would override that choice, it would create severe issues with the slow erosion of the moral and autonomy-related copyright rights. To begin with, beside the economic right to copyright to be paid, there is a right of the creator to make choices with regard to the way of use, association and transformation of the work. By automatically deciding which copyrighted material to use, MCL inherently denies these personal rights and asserts its decision over the right to refuse.

Things get particularly tricky when the objection is based on ethical, reputational, and other issues. To illustrate, when certain works may be utilized to teach an AI to create the type of derivative content that will be opposite to the values or art of the creator, any kind of financial rewards will not be enough. In this case, the right involved is obviously related to the concept of moral rights of integrity and attribution that is internationally accepted and explicitly defined in Section 57 of the Indian Copyright Act of 1957.[19] By limiting the right to refuse in this way MCL will be straining such fundamental protections and reducing them to the demands of the industrialized world.

Moreover, any such plans will probably place individual choice secondly to considerations of industry convenience since they will make it necessary to engage in order to sacrifice the consent of creators at the altar of data access to AI creators. This is not only against the spirit of copyright laws, but this will also pose challenges that may lead to resistance and even challenges by the very creators. Finally, due to this contradiction, the opt-out problem is one of the critical ones in MCL.

[18] Directive (EU) 2019/790 on Copyright in the Digital Single Market, Arts. 3–4.

[19] The Copyright Act, 1957, § 57 (India).

4. RECOMMENDATIONS & CONCLUSION

When creating a sustainable ecosystem of AI, the shift towards the Mandatory Collective Licensing regime must be accompanied by adequate protections that guarantee creativity and rights of creators are not violated. The initial protection that is paramount towards the desired result is the implementation of tiered licensing scheme whereby the lower licensing costs are offered to institutions like the universities, independent researchers, and startups. This will assist in making sure that the ecosystem of AI is diversified in terms of its innovation work rather than the leading companies, including Google and Microsoft, enjoying the shift towards Mandatory Collective Licensing. The second protection, which must be implemented, is the introduction of strict transparency, in particular, the implementation of standards, like C2PA (Content Provenance and Authenticity).[20] Additionally, there must be viable opt-out provisions in existence which would guarantee the protection of moral rights by allowing the creators an opportunity not to have their work utilized in some applications such as style mimicry and identity creation whilst still having the opportunity to join larger sets of data when they wish; a far more balanced method of establishing a good compromise between the convenience of collective licensing and individual freedom. All these steps can assist in transforming MCL into an equal system that will reflect the interests of all stakeholders and not just big companies.

Although, admittedly, there are certain blemishes, the concept of compulsory collective licensing as such appears to introduce itself as a sensible and procedural measure towards resolving the growing tension between the copyright laws and the development of AI in that it will introduce some much-needed legal clarity into the developers and, in the process, provide the means of something to compensate the creators. Since the current trends are moving towards opposition to the use of broad perspectives of fair use, such a solution will allow us to leave what might turn out to be a significantly unsustainable standpoint to one that will consider our technical innovations alongside the interests of the creators. With this said, its validity will be quite dependent on the effectiveness with which it will be conducted. Instead of having endless arguments on the necessity to compensate creators, it is necessary to redirect the debate towards developing ways of compensating that will consider the interests of all the concerned parties.

[20] Coalition for Content Provenance and Authenticity (C2PA). (2023). *Technical specification*.