GENERATIVE AI VS COPYRIGHT

Mrs. Garima Juneja, Assistant Professor, Gitarattan International Business School, IPU

Ms. Mendu Sri Soumya, Gitarattan International Business School, IPU

ABSTRACT

Generative Artificial Intelligence (AI) is swiftly reshaping various sectors through its capacity to enhance human cognition and streamline research and data analysis processes. Despite its transformative potential, the deployment of such technology gives rise to intricate legal challenges most notably, the incorporation of copyrighted and proprietary content in the training of large language models (LLMs)**.** A key issue concerns the extent to which the unlicensed use of this data may be legitimized under the doctrines of fair use (as recognized in the United States) or fair dealing (as applied in India), particularly in light of the current legislative vacuum and the absence of definitive judicial guidance on the matter.

This research examines the evolving jurisprudence on fair use in the context of AI training, with a focus on two recent U.S. cases: *Andrea Bartz v Anthropic* & *Richard Kadrey v. Meta*. While both decisions recognized the "highly transformative" nature of AI training, they diverged on the requirement of a "lawfully acquired first copy," a critical factor with direct implications for Indian copyright law. In India, the ongoing case of *ANI v OpenAI* has brought this debate to the forefront. ANI alleges unauthorized use of its news content for training OpenAI's models, while OpenAI contends its use qualifies as transformative and falls within the scope of fair dealing under Section 52 of the Indian Copyright Act, 1957.

The paper critically analyses how Indian courts might interpret the fair dealing provision in the context of AI training, considering the limited scope of Section 52 and the absence of a transformative use standard. It also explores how real-world instances such as the Ghiblistyle image generation trend, demonstrate that generative AI can replicate the creative essence of original works without directly copying them, thereby constituting a form of copyright infringement.

Through comparative legal analysis, this paper argues for the urgent need to develop a nuanced, jurisdiction-specific legal framework that addresses the intersection of intellectual property and artificial intelligence. It concludes that courts must balance the transformative potential of AI with the rights of

original creators, ensuring that innovation does not override the foundational principles of copyright law.

INTRODUCTION:

Generative AI is advancing at an unprecedented pace, with emerging trends centred on AIaugmented applications that enhance human intelligence and drive increased adoption across workplaces. One of its key advantages lies in its ability to significantly reduce the time spent on research and data processing tasks. However, alongside these benefits arise serious concerns regarding the legal and ethical dimensions of AI development particularly the use of copyrighted or proprietary content in training large language models (LLMs).

Training LLMs requires vast and diverse datasets, which may include freely accessible content as well as subscription-based or proprietary materials developed through extensive original research. The legality of acquiring and utilizing such content for AI training remains contested, especially in light of the ongoing **ANI v. OpenAI** case before the Delhi High Court. While India has yet to establish a landmark judicial precedent on this issue, recent

U.S. decisions have begun to shape the global discourse. Two significant cases: **Andrea Bartz v. Anthropic** and **Richard Kadrey v. Meta** have both examined whether the use of copyrighted materials for AI training qualifies as "fair use," particularly focusing on whether such use is transformative in nature.

Although both courts recognized that AI training constitutes a "highly transformative" activity thus satisfying the first factor of the fair use analysis but they diverged sharply on the requirement of possessing a "lawfully acquired first copy" of the material. This divergence carries significant implications for how Indian courts might interpret similar cases in the future, especially in light of the fair dealing provisions under Section 52 of the Indian Copyright Act and the ongoing arguments in ANI v. OpenAI.

Fair Use / Fair Dealing in the Context of Generative AI:

The development and refinement of large language models (LLMs) and other generative AI technologies necessitate access to extensive corpora of textual and visual material, often drawn from both open-source and proprietary datasets. This practice has sparked international legal scrutiny, particularly concerning the **unauthorized use of copyrighted works** for training

purposes. Central to this debate is the interpretation and scope of **fair use** in jurisdictions such as the United States, and **fair dealing** in regions like India and the United Kingdom. The use of protected content without explicit permission for commercial or algorithmic training challenges the traditional boundaries of copyright law and has prompted critical discourse on how existing legal doctrines should evolve in response to emerging AI technologies.

1. Fair Use under U.S. Law

Under U.S. copyright law, the doctrine of fair use is codified in Section 107 of the Copyright Act of 1976. It allows for limited use of copyrighted material without permission from the rights holder for purposes such as criticism, comment, news reporting, teaching, scholarship, or research. Courts assess fair use using a four-factor test:

i. *Purpose and character of the use*, including whether such use is of a commercial nature or is for non-profit educational purposes.

ii. *Nature of the copyrighted work.*

iii. *Amount and substantiality* of the portion used in relation to the copyrighted work as a whole.

iv. *Effect of the use upon the potential market* for or value of the copyrighted work.

In recent high-profile U.S. cases involving generative AI, Andrea Bartz v. Anthropic and Richard Kadrey v. Meta courts agreed that AI training is "highly transformative" in nature. This aligns with the first factor of the fair use analysis. However, the judgments diverged significantly on the necessity of acquiring a **lawfully obtained first copy** of the material used in training datasets.

RECENT U.S. CASES :

a) ANDREA BARTZ VS ANTHROPIC (Judge Alsup)

Anthropic(American AI Start Up Company) scraped LibGen for Book3 (a dataset containing 196,640 books in plain text format which is used for training language models) and similar datasets. These books were before acquired without lawful purchase, since books on LibGen are made available without authorization. the company reportedly created a “general purpose

library" using these books. But later after getting legal advice , Anthropic used digitized versions of legally purchased second-hand copies for training its LLM'S(ai models). It continued with maintaining its "general purpose library", though, for future use. Judge Alsup holds that creating copies of copyrighted works and storing them as a "general purpose library" from a shadow library cannot constitute fair use, regardless of whether the copyrighted work is ever exposed to any human being by the user. The Judge reasons that claiming a research purpose to copy textbooks from shadow libraries would destroy academic publishing markets.

b) Richard Kadrey V Meta (Judge Chabria)

In this case , Meta scraped LibGen and maintained library same like Anthropic but difference is that Meta didn't purchase any second hand copies for its LLM'S Training, it kept using the digitalized copies from the library in every step for the LLM'S training(without any lawful purchase) . Judge Chhabria states that "*to say that Meta's downloading was 'piracy' and thus cannot be fair use begs the question because the whole point of fair use analysis is to determine whether a given act of copying was unlawful*." Moreover, as Meta's use of shadow libraries didn't provide any advertising revenue, this factor wouldn't impact the transformative nature of the use. In direct contrast to the *Anthropic* decision, the *Meta* Court holds that downloading works from shadow libraries must be considered in light of its transformative purpose: training an AI model. The Court found that because Meta's use of the books to train Generative AI was transformative (further purpose and different character of the use), so too was its act of downloading them.

This judicial divergence illustrates the unsettled nature of fair use in AI-related copyright disputes, and highlights the importance of the "lawful first copy" criterion in shaping legal outcomes.

2. FAIR USE UNDER INDIAN LAW:

Section 52(1)(a) of the Copyright Act permits "fair dealing" for research, personal use, news reporting, and certain other purposes. The explanation to this section states that storage in an electronic medium of a copy in the process of such uses including incidental storage of a computer programme that is not itself an infringing copy for the purpose of such uses, is noninfringing. Unlike the broader and more flexible U.S. fair use doctrine, fair dealing in India is **enumerated and purpose-specific**, covering uses such as:

- Private or personal use, including research
- Criticism or review
- Reporting of current events
- Use by educational institutions
- Judicial proceedings

However, Indian law does not yet address the question of whether training an AI system on large volumes of copyrighted data qualifies as "research" or whether such training constitutes reproduction or transformation. An Indian ongoing current case regarding training of AI modules with copyrighted/non copyrighted content-

ANI VS OPEN AI case:

The case is presently under consideration before the Delhi High Court, which is deliberating on the intricate legal questions pertaining to **copyright protection**, the applicability of fair use principles, and the legitimacy of employing copyrighted materials for AI model training**.**

- **ANI's Claim:**

ANI Media Pvt Ltd filed a copyright infringement lawsuit against OpenAI, alleging that OpenAI's ChatGPT used its news content to train its AI models without permission.

- **OpenAI's Defence:**

OpenAI has argued that its use of the data falls under fair use or does not constitute copyright infringement, as the data is transformed during the training process and the output is not a direct copy of the original work.

VIEWS REGARDING THIS CASE:

In *ANI v. OpenAI*, ANI Media Pvt Ltd. contends that the requirement of possessing a “lawfully acquired first copy” is implicitly embedded within the Explanation to Section 52(1)(a) of the Indian Copyright Act. According to this interpretation, the fair dealing exception applies only when the material used for purposes such as research, news reporting, review, or personal use

originates from a legitimately obtained source. ANI's position echoes the reasoning adopted by Judge Alsup in *Bartz v. Anthropic*, where it was held that AI training constitutes fair use only if the initial source material was lawfully acquired.

ANI further argues that this lawful acquisition criterion should become especially relevant when the downstream use of such data results in commercial gain. Without this safeguard, ANI asserts, the rights and interests of original creators and copyright holders risk being unfairly undermined.

In response, OpenAI maintains that the materials in question were utilized solely for training purposes and that the output generated by its models does not constitute a reproduction of the original works. This defence draws parallels with the judicial reasoning in *Kadrey v. Meta*, where the court held that the use of unauthorized copies from shadow libraries for training generative AI systems could still be considered transformative, particularly in the absence of direct monetization at the storage stage.

However, despite the lack of immediate revenue generation during data acquisition, it is widely acknowledged that the commercial potential of trained AI models is substantial. Therefore, the legality of how training data is sourced lawfully or unlawfully must be carefully examined in evaluating the validity of fair use claims.

This case raises pivotal questions for Indian copyright jurisprudence in the era of generative AI, including:

- The extent of copyright protection afforded to journalistic content in the context of AI training;
- Whether such training activities amount to infringement; and
- The applicability and boundaries of the fair dealing doctrine when AI systems ingest and learn from copyrighted material.

The outcome of this case could have significant implications for the future of copyright law in India, particularly in relation to the use of copyrighted material for AI training.

Use of Celebrity Content in AI Training: A Threat to Personality & Copyright rights -

An emerging concern in the development of generative AI systems is the **unauthorized use of public figures' images, voices, and likenesses** to train AI models. These training datasets often include vast quantities of publicly available visual and audio content, including interviews, films, advertisements, and social media footage of celebrities. While such content is accessible, its use without consent for commercial AI development raises serious legal and ethical issues.

This unauthorized data harvesting not only violates personality rights but also potentially amounts to copyright infringement, especially when the AI reproduces outputs that mimic the celebrity's distinctive voice, facial expressions, or mannerisms. The lack of transparency in dataset composition further complicates enforcement and accountability.

Several Indian celebrities including Akkineni Nagarjuna and members of the Bachchan family have approached courts to seek protection of their persona from being used in AIgenerated deep fakes and training datasets. These cases reflect growing awareness of how AI systems, if left unregulated, could exploit a celebrity's identity for profit without authorization or remuneration.

The misuse of such personal and copyrighted data not only affects the reputation and commercial value of the individual but also creates a precedent for mass data exploitation under the guise of technological innovation. As a result, courts are increasingly being urged to recognise personality rights violations within the broader debate on AI and copyright law, signalling the need for clear legal frameworks and consent-based data use policies in AI training.

The Legal Dilemma in AI Training

The core issue in both legal systems is whether training an AI model on copyrighted content regardless of whether that content is ever reproduced in its original form constitutes infringement. The growing reliance on shadow libraries, web scraping, and automated data collection has blurred the lines between research and commercial exploitation.

The "transformative use" argument becomes particularly contentious in AI contexts, where the outputs are generated algorithmically but are undeniably shaped by the inputs, which often carry the essence, style, or expression of original authors. This concern is not just theoretical:

generative tools have already been shown to mimic artistic aesthetics, such as in the recent Ghibli-style image generation trend, raising ethical and legal alarms over the unauthorized use of creative expressions.

Implications for Indian Jurisprudence

Given India's lack of precedent on this issue, the resolution of the **ANI v. OpenAI** case may set a crucial benchmark. The courts will need to assess:

- Whether AI training qualifies as "research" under Section 52;
- Whether the training is "fair" if the original content was not lawfully obtained;
- The extent to which the output of an AI system must be dissimilar to the input to be deemed transformative;
- And whether the absence of direct monetization from data storage absolves liability when eventual commercial gain is anticipated.

These questions will shape the contours of copyright law in India in the age of artificial intelligence, and determine the future balance between innovation and protection of original authorship.

Conclusion

The growing discourse around generative AI underscores increasing concerns about the potential misuse of fair use and fair dealing doctrines, particularly in the absence of clear legal guidelines. While AI companies often assert that their models do not reproduce content directly from original sources, real-world instances such as the recent **Ghibli-style image generation trend where photos were generated through ChatGPT with the essence of the Ghibli art**, challenge this assertion. In this case, AI-generated visuals closely mirrored the unique aesthetic and emotional depth characteristic of **Hayao Miyazaki**, the acclaimed Japanese animator whose hand-drawn illustrations often require months of creative labour.

Although the AI-generated Ghibli images did not copy specific characters or scenes, they clearly replicated the artistic essence and distinctive style of Miyazaki's original works. This trend serves as compelling evidence that AI can absorb and reproduce the creative identity of

an artist, even without duplicating content word-for-word. Such replication of a creator's stylistic and expressive core without consent amounts to a form of copyright infringement, as it violates the original author's moral and intellectual property rights.

This phenomenon raises a critical issue in AI ethics and copyright law: the line between copying a work's literal expression and replicating its creative essence is becoming increasingly blurred. For example, when training data is sourced from books or scholarly articles, the AI model is exposed not just to information, but to the author's unique voice, reasoning, and intellectual framework. As a result, the AI may generate outputs that subtly reflect the author's original thought processes, thereby encroaching upon the domain of protected authorship.

Courts, particularly in jurisdictions like India where legal precedent in this domain is still evolving, must adopt a balanced and contextual approach. Each case should be evaluated on its individual facts, with careful consideration of both the transformative nature of the AI system and the degree to which original creative content is utilized. Legal frameworks must ensure that innovation does not come at the expense of undermining the creative rights and economic interests of original authors and artists.

Establishing boundaries for what constitutes acceptable use of creative content in AI training is not just a legal necessity it is essential to uphold the integrity of intellectual property in the digital age.

References:

https://roninlegalconsulting.com/bartz-v-anthropic-generative-ai-and-the-fair-usedoctrine/

https://mashable.com/article/kadrey-v-meta-judge-rules-against-plaintiffs

https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.231.0_3.pdf

https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.598.0_1.pdf

https://theleaflet.in/digital-rights/ani-v-openai-in-the-delhi-hc-everything-so-far-and-allthat-is-at-stake

https://articles.manupatra.com/article-details/FAIR-DEALINGS-AND-FAIR-USECRITICALLY-ANALYSING-THE-COPYRIGHT-EXEMPTION-DOCTRINES-IN-PLACE-IN-INDIA-AND-THE-UNITED-STATES