The Scrape and the Standard: Evaluating "Fair Use" in AI Training Datasets under Modern Copyright Law

[1]shriya rajput, [2]Dr. Anuj Sharma
[12]Amity University , Lucknow

I. Introduction

The rapid rise of generative artificial intelligence often feels like a digital magic trick. With a single text prompt, a user can quickly create a complex legal brief, a functional Python script, or a high-quality digital painting in the style of a master artist. However, this digital magic does not occur in a vacuum, nor does it produce new material from thin air. Beneath the sleek user interfaces and corporate marketing lies a vast, hidden system powered by automated internet scraping. Large Language Models (LLMs) and diffusion-based image generators do not truly grasp the nuances of human emotion, grammar, or visual design. Instead, they need training on billions of existing data points. For years, organizations like Common Crawl, LAION, and various tech companies have quietly scoured the open web, collecting books, academic articles, personal photos, and digital art into enormous datasets without explicit permission.

This sudden technological shift has created a significant legal, economic, and cultural clash. On one side, AI developers and venture capitalists see mass data scraping as essential for societal innovation. They argue that limiting access to public data would halt technological progress, slow economic growth, and give international competitors with less strict regulations an edge. On the other side, human creators—including freelance writers, visual artists, journalists, novelists, and photographers—see this practice as systemic, unauthorized theft. They believe their life's work is being used without consent to create commercial software that competes with and ultimately replaces them in the market.

At the heart of this multi-billion-dollar conflict is American copyright law's fair use doctrine. AI developers argue that mass data collection is a transformative process protected by this doctrine, claiming that the copies made during training are only intermediate and for analysis. However, a careful application of modern copyright law shows that unauthorized scraping does not meet the requirements under 17 U.S.C. § 107. The commercial nature of these models, along with a tightening judicial perspective on what qualifies as transformative and clear market substitution, undermines this defence. To prevent either choking off technological progress or devastating creative industries, copyright law needs to change. Instead of pushing for a blunt, all-or-nothing fair use decision, the legal system should move toward structured compulsory licensing systems that ensure fair compensation for human authors.

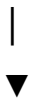
II. The Mechanics of Data Ingestion: Copying vs. "Learning"

To properly evaluate the legality of artificial intelligence training, one must look beyond marketing terms like "machine learning," "neural network ingestion," and "data digestion" to examine the actual engineering reality. Silicon Valley executives often use a common defence before congressional committees and public forums. They claim that AI models read, analyse, and learn from books or art just like a human student in a library or a museum. This analogy may comfort investors, but it is legally flawed and technically inaccurate. A human student uses their brain to understand concepts, combine ideas, and find inspiration while leaving the original work intact. In contrast, an AI model needs powerful computer servers to make explicit, unauthorized digital copies of copyrighted material as a necessary step for its algorithmic analysis.

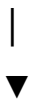
The training cycle of an AI model starts with the ingestion phase. Software crawlers copy selected creative works from the internet and store them in server RAM or local high-speed storage. Once stored, the training software breaks these works down into high-dimensional mathematical vectors. For text, words turn into numerical tokens. For images, pixels change into numerical coordinates representing colour, contrast, and spatial relationships. The model then runs training loops across these vectors to calculate statistical probabilities—for example, the chance that a specific word follows another or the mathematical pattern that defines a "gothic oil painting."

AI companies assert that since the original raw files are often deleted or heavily altered after extracting these mathematical patterns, no permanent copyright violation has taken place. They argue that the model itself does not hold a list of JPEG images or text files; instead, it contains weights, biases, and parameters. However, copyright law does not care whether the final product has the original file nor does it excuse infringement based on the short life of a copy. [Original Copyrighted Work]

[Ingestion & Scraping Process]



[Temporary Copy Created in RAM / Local Storage] <── Prima Facie Infringement (17 U.S.C. § 106)



[Vectorization & Mathematical Analysis]

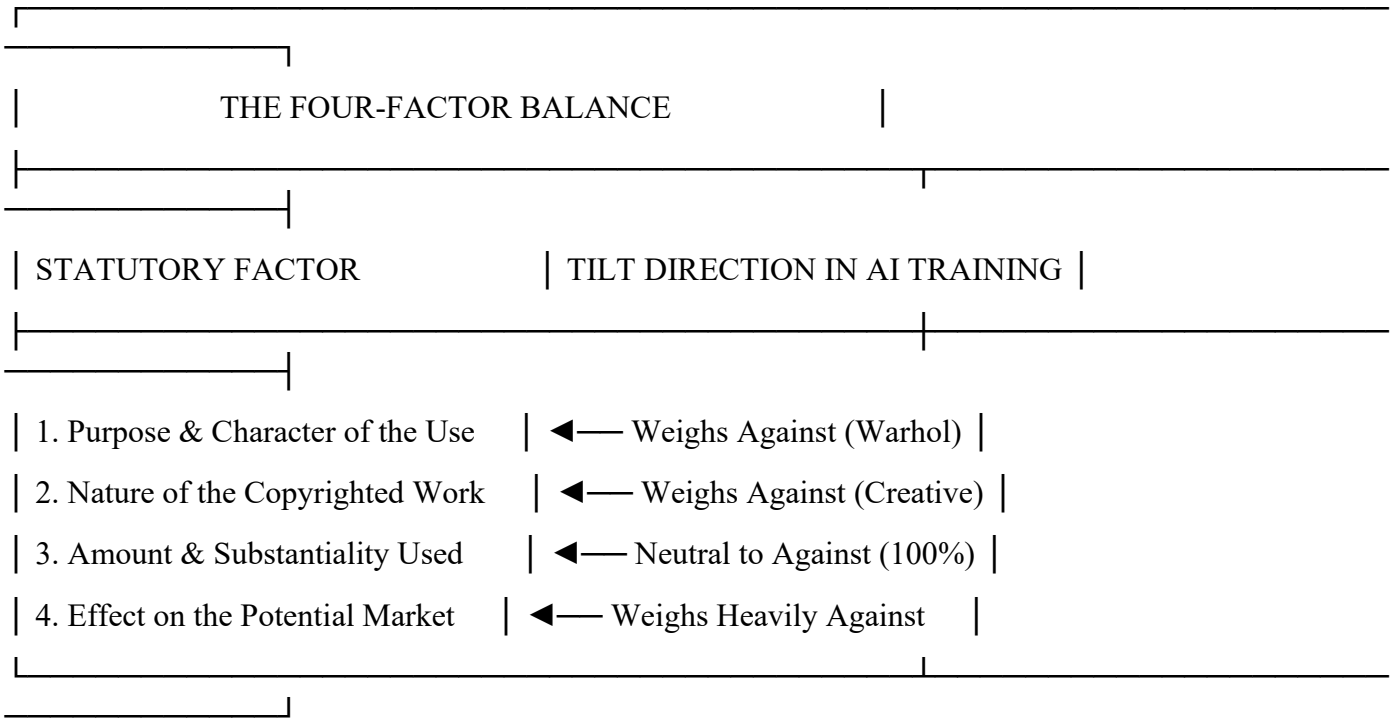
[Final Weights / Algorithmic Model]

The language of the Copyright Act of 1976 is clear. Under 17 U.S.C. § 106, a copyright owner has the exclusive right "to reproduce the copyrighted work in copies." A copy is defined as any material object that contains a work fixed by any method known now or in the future. This includes works that can be perceived, reproduced, or communicated, either directly or with the help of a machine or device.

The main legal case on this issue is MAI Systems Corp. v. Peak Computer, Inc. (1993). In that case, the Ninth Circuit decided that moving software or data into a computer's temporary RAM creates a temporary copy that lasts long enough to count as copyright infringement. Courts have consistently upheld the MAI Systems decision over decades of digital copyright cases. Because an AI model cannot calculate its parameters without first making complete, unauthorized digital copies of millions of copyrighted files in its memory, the training process clearly infringes on the reproduction right. The key legal question is not whether copying took place—the technology itself shows that it did. The only real question is whether that systematic copying can be justified under fair use.

III. The Fair Use Battlefield: Deconstructing the Four Factors

When an AI developer faces a lawsuit for unauthorized data scraping, their defence relies almost entirely on the affirmative defence of fair use, codified in 17 U.S.C. § 107. Fair use is not a blank check; it is a highly structured, equitable doctrine that requires courts to balance four specific statutory factors on a case-by-case basis. Applying these factors to the massive scale of AI training reveals structural vulnerabilities in the tech industry's defence.



Factor 1: The Purpose and Character of the Use

The first factor addresses the purpose and character of the infringing use-whether the use is of a commercial nature or for non-profit educational purposes. For the past three decades, tech platforms have relied heavily on the judicial doctrine of "transformative Ness"-coined by Judge Pierre Level and later adopted by the Supreme Court in Campbell v. Acuff-Rose Music, Inc. (1994). A work is transformative if it "adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message."

AI developers claim that using copyrighted text or imagery merely as statistical input to create a tool that performs many functions would be wholly transformative. However, the doctrine of transformativeness was dramatically refocused in the Supreme Court's ruling in Andy Warhol Foundation for the Visual Arts, Inc. V. Goldsmith (2023). The Court essentially brought the first factor back to centre stage and made it clear that "a commercial use may not be considered transformative where the purpose and character of the secondary work is to supersede or substitute for the original." So, if there's commercial overlap in purpose between the secondary work and the original creation, the use falls apart as fair.

Commercial generative AI is not being trained for non-profit research; rather, it is a product that is sold at a subscription price to customers. Furthermore, AI tools train output in a manner and with an expressive/commercial intent that is identical to that of the original works used to train them. When an AI model trains on a novelist's works for commercial fiction writing or a commercial photographer's portfolio to produce commercially useful lifestyle images for advertisements, the purpose of the ingestion directly corresponds with the commercial purpose of the original creation, strongly weighing against fair use under the post-Warhol standard.

Factor 2: The Nature of the Copyrighted Work

The second statutory factor guides courts to analyse the nature of the copyrighted work. Certain creative works, such as fictional stories, poetry, movie scripts, music and conceptual illustrations, have a more central purpose in copyright than works that contain purely factual or technical information like phone books, scientific reports, or historical timelines.

AI training does not make this distinction and indiscriminately scrapes both facts and fiction. Contemporary generative AI training data are comprised primarily of highly creative works of literature, commercial digital art archives, independent journalistic works, and copyrighted lyrics. Because these works sit at the absolute heart of the space copyright was intended to protect and were created as such, the copyright protection awarded for them is quite strong and is sharply against the AI developer when these highly creative works are taken and used commercially without licensing.

Factor 3: The Amount and Substantiality of the Portion Used

The third factor examines how much of the copyrighted work was used and whether the portion taken was substantial in relation to the work as a whole. In copyright disputes, courts generally view copying an entire work as weighing against a finding of fair use, while the use of a small or insignificant excerpt is more likely to be excused.

In the machine-learning context, however, AI systems typically cannot operate effectively on isolated excerpts, random paragraphs, or heavily cropped images. To identify patterns, relationships, and stylistic features, developers often argue that the model must be trained on the complete work. In practice, that means scanning every page of a book or analysing every pixel of an image.

Supporters of AI training practices frequently cite earlier technology cases to argue that copying an entire work does not automatically defeat a fair-use defence when the copying is necessary for a technical purpose. Courts have recognized this principle in certain circumstances, particularly where the copying served a functional, non-expressive purpose, such as reverse-engineering software for interoperability or creating searchable indexes.

The challenge for AI companies is that training models on vast collections of creative works may not fit neatly within those precedents. While earlier cases involved tools that helped users access or organize information, generative AI systems can produce new content that may compete with the original works. As a result, courts may view the wholesale copying of creative portfolios less as a technical necessity and more as a use that directly implicates the interests of copyright owners, making the third fair-use factor harder to justify.

Factor 4: Effect on the Potential Market

The fourth factor in the fair use test is usually considered the most important because it looks at the financial impact of using someone else's work. Judges do not just look at whether one person’s actions caused harm. Instead, they think about what would happen if everyone started doing it. They ask whether this behavior would significantly hurt the market for the original work or take away future licensing opportunities.

Many people who are concerned about how AI is trained see this as their strongest argument. Unlike older tools that just help people search or organize information, generative AI can create things that compete directly with the original content. Because of this, the debate has shifted. It is no longer just about whether the AI copied data during training, but whether the technology itself devalues the original work.

We can see this in digital media today. If an AI is trained on a journalist's reporting or articles behind a paywall, users might just get their summaries and explanations from the AI instead of visiting the original site. This causes publishers to lose out on traffic, ad revenue, and subscriptions. While one person doing this might not matter much, critics argue that millions of AI queries could eventually make it hard for original sources to stay in business.

A similar problem is happening in the arts. AI can generate illustrations and commercial designs in seconds. Independent artists argue that since these systems were trained on their work, they allow companies to avoid paying for commissions. From their perspective, the AI is not just learning; it is competing in the same market as the artists who provided the training data.

People developing AI disagree, arguing that what the AI produces is different from the training material. Even so, the question of whether AI replaces the need for the original work is the main focus of the legal debate. Copyright law was created to protect the financial incentives that lead people to create things in the first place. Whether AI actually harms these markets will likely be the most important question for courts to decide in the future.Judicial Precedents: From Google Books to Modern AI Litigation

To understand how courts may approach contemporary AI copyright disputes, it is useful to examine how previous generations of technological innovation were treated under copyright law. Throughout history, courts have repeatedly been asked to balance the interests of creators against the benefits of emerging technologies.

One of the most frequently cited cases in discussions surrounding AI is Authors Guild v. Google, commonly known as the Google Books case. In that dispute, Google scanned millions of books from research libraries and created a searchable digital database. The Second Circuit ultimately concluded that Google's conduct constituted fair use despite the fact that entire books had been copied during the scanning process.

The court's decision rested heavily on the nature of Google's use. Although Google digitized complete books, it did not provide users with unrestricted access to those works. Instead, the service displayed only limited snippets of text and functioned primarily as a search and discovery tool. Users could locate relevant books and identify passages containing particular terms, but they could not use the system as a substitute for purchasing or borrowing the original works. According to the court, Google Books expanded public access to information about books without displacing the market for the books themselves.

This distinction has become a central point of disagreement in debates over generative AI. Supporters of AI training argue that Google Books demonstrates that large-scale copying can qualify as fair use when it serves a transformative technological purpose. They contend that machine learning systems use copyrighted works as inputs for statistical analysis rather than for direct consumption by users.

Critics, however, argue that generative AI differs in important respects. Unlike Google Books, which merely helped users locate existing works, generative AI systems produce new outputs that may compete with human-created content. An individual using an image generator or a large language model is often seeking a finished product rather than information about where an original work can be found. From this perspective, the technology functions less like a digital card catalogue and more like a substitute for the labour of writers, artists, designers, and other creative professionals.

Recent litigation has further intensified these debates. Lawsuits brought by authors, news organizations, visual artists, and publishers have raised questions about whether AI systems merely learn general patterns or whether they sometimes retain and reproduce protected expression. Plaintiffs have presented examples in which AI models allegedly generated passages of text, images, or other content that closely resembled copyrighted works contained within their training data.

These allegations remain the subject of ongoing legal proceedings, and courts have not yet established a definitive framework for resolving them. Nevertheless, the cases have shifted public attention toward issues of memorization, reproduction, and market impact. As courts continue to evaluate the evidence, the outcomes of these disputes may play a decisive role in shaping the future relationship between copyright law and artificial intelligence.

IV. Human Inspiration and Machine Learning: A Difficult Comparison

One of the most common defences of AI training is the argument that human beings also learn by observing the work of others. Writers read books, artists study paintings, and musicians listen to existing compositions before developing their own creative voices. According to this view, machine learning represents a technological extension of a process that has always existed within human culture.

Although the comparison has intuitive appeal, many creators argue that it overlooks important differences between human learning and machine learning. Human artists develop their skills gradually over many years. They absorb influences from a limited number of sources, combine those influences with personal experiences, and ultimately produce works shaped by individual judgment and creativity. Human learning is constrained by time, memory, effort, and biological limitations.

AI systems operate on an entirely different scale. Modern models can process billions of images, books, articles, recordings, and other works within relatively short periods of time. Rather than learning through lived experience, they identify statistical relationships within vast datasets and use those relationships to generate outputs. While this process may share certain conceptual similarities with human learning, the scale and speed involved are fundamentally different.

For many artists and writers, the primary concern is not simply that AI systems learn from existing works. Rather, it is that commercial entities may profit from the use of enormous quantities of creative material without compensating the individuals who produced it. Critics argue that this creates an imbalance in which the economic value generated by creative labour is increasingly captured by technology companies rather than by creators themselves.

Supporters of AI development respond that innovation has often disrupted existing industries and that broad access to information has historically contributed to technological progress. Nonetheless, the question remains unresolved. Whether machine learning should be treated as legally analogous to human learning is likely to remain one of the most significant issues in future copyright debates.

V. Beyond Fair Use: The Case for Compulsory Licensing

The ongoing conflict between AI developers and copyright holders illustrates the limitations of relying exclusively on fair-use litigation to resolve these complex issues. Fair use was designed as a flexible legal doctrine, but it may not be sufficient on its own to address the unprecedented scale of modern AI training practices.

If courts adopt a highly restrictive approach, AI companies could face extensive liability and significant barriers to future development. Such outcomes might discourage investment, limit innovation, and slow the advancement of potentially beneficial technologies. On the other hand, if courts interpret fair use too broadly, creators may lose meaningful control over how their works are used and monetized. Such a result could undermine the economic foundations of creative industries and reduce incentives for future artistic and literary production.

For this reason, scholars, policymakers, and industry participants have increasingly explored alternatives that move beyond the traditional fair-use framework. One of the most widely discussed proposals is the creation of a compulsory licensing system for AI training.

The United States has adopted similar solutions during previous periods of technological disruption. In the early twentieth century, inventions such as player pianos and commercial radio broadcasting created significant tensions between technological innovation and copyright protection. Rather than banning these technologies or allowing unrestricted use of copyrighted works, Congress established licensing mechanisms that enabled creators to receive compensation while allowing new industries to flourish.

A comparable approach could potentially be adapted for generative AI. Under such a framework, developers would be permitted to use copyrighted works for training purposes under specified legal conditions. In exchange, they would contribute to a collective licensing fund that compensates authors, artists, journalists, publishers, and other rights holders. Compensation could be distributed through transparent mechanisms designed to reflect the use and value of creative works within training datasets.

Technological tools may also assist in implementing such a system. Standardized metadata, machine-readable licensing preferences, digital watermarking technologies, and other tracking mechanisms could help creators indicate how their works may be used while facilitating compensation where appropriate. Although significant practical and technical challenges remain, these proposals demonstrate that alternatives exist beyond the binary choice of unrestricted use or complete prohibition.

Compulsory licensing is not a perfect solution, but it offers a possible middle ground. By balancing the interests of innovation and creator compensation, such a framework could help reduce legal uncertainty while preserving the incentives necessary for both technological and cultural development.

VI. Conclusion

The fair-use doctrine was developed to provide flexibility within copyright law and to ensure that legal protections did not unnecessarily hinder education, research, commentary, criticism, and creative expression. The emergence of generative artificial intelligence has placed unprecedented pressure on those principles and has raised questions that lawmakers and judges could scarcely have anticipated when the doctrine first evolved.

At its core, the debate concerns the relationship between technological innovation and human creativity. AI systems derive value from enormous quantities of human-generated content, yet many creators fear that those same systems may diminish demand for their work or reduce opportunities for compensation. At the same time, AI developers argue that access to information and data is essential for continued innovation and economic growth.

The challenge facing policymakers is therefore not simply legal but also economic, technological, and cultural. Any sustainable solution must acknowledge both the transformative potential of artificial intelligence and the contributions of the writers, artists, journalists, musicians, researchers, and other creators whose work forms the foundation of modern training datasets.

Whether the future lies in judicial decisions, legislative reforms, licensing frameworks, technological safeguards, or some combination of these approaches, the ultimate objective should be to create a system that promotes innovation without undermining the creative ecosystem on which that innovation depends. A balanced and forward-looking approach offers the best opportunity to ensure that technological progress and human creativity continue to develop together rather than at each other's expense.

References

[1] *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258 (2023).

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

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

[4] *MAI Systems Corp. v. Peak Computer, Inc.*, 991 F.2d 511 (9th Cir. 1993).

[5] Copyright Act of 1976, 17 U.S.C. §§ 106, 107 (2012).

[6] Level, Pierre N. *Toward a Fair Use Standard*, 103 Harv. L. Rev. 1105 (1990).