The Rise of AI: New Frontiers in Fair Use

By: Inaara Bhaidani

Abstract: The rapid rise of generative artificial intelligence (AI) and various large language models (LLMs) has raised pressing questions of legality and ethics when it comes to using copyrighted material to train such AI models. These companies rely on various datasets of text, images, and other media—much of which originates from legally copyrighted works—in order to develop systems that can generate new content based on those inputs. This piece examines how various lawsuits have challenged the long-standing principles of copyright doctrine within the context of the litigations that have arisen due to this practice. Does training an algorithm using existing creative works without a license constitute enough of a transformative use to stay within the bounds of fair use, or would it be considered a mass, unlawful reproduction of protected material? While AI training data *can* appear transformative and be used to generate original results that provide immense technological and social benefits, courts must narrow the doctrine of fair use to preserve the balance between innovation and rights.

Article

"The fair use doctrine is not a license to appropriate the expression of others at will."
- U.S. Supreme Court[1]

I. Introduction

In recent years, AI has been placed at the forefront of many renowned companies such as Microsoft, Google, IBM, and more. It is changing the way that people interact with art, writing, and other forms of media by essentially serving as the primary "creative mindset" behind many human innovations. Tools like OpenAI's *ChatGPT* and Stability AI's *Stable Diffusion* can produce media that is entirely original; however, this process relies on training data derived from millions of existing works—many of which are copyrighted. For example, an AI algorithm designed to produce an original photograph of a cat can only do so after being trained by applying a dataset of existing cat photographs taken by humans. It repeatedly analyzes the properties of each image in that dataset until it can intelligently use those properties to successfully generate an original cat photograph. It is important to note, however, that some, if not all, of the cat photographs that were used to train the AI model may be copyrighted. Do the

[1] Harper & Row, Publishers, Inc. v. Nation Enterprises, 471 U.S. 539, 556 (1985).

photographers know that their work is being used to create similar replications? Moreover, would they allow it?

Several recent lawsuits, including *The New York Times v. Open AI (2023)*, question whether using copyrighted works to train AI systems qualifies as fair use or constitutes unauthorized copying.[2] The outcome carries significant consequences for both the tech industry and creators. If courts rule that AI training falls under fair use, companies can continue developing these tools with minimal legal risk and will face broad liability otherwise.

II. Background

Under Title 17 § 107 of the U.S. Code describing copyright law, courts determine fair use by weighing four statutory factors.[3] The first factor, the purpose and character of the use, considers whether the use is for commercial or nonprofit educational purposes. This can be interpreted as an evaluation of whether the use adds new meaning or purpose to the original work. The second factor, the nature of the copyrighted work, refers to the creativity and objectivity of the material. Generally, more creativity results in stronger protection and rights. The third factor, the amount and substantiality of the portion used in relation to the copyrighted work as a whole, examines how much of the original work was copied and whether the "heart"—or main idea—of the work was taken. The fourth and final statutory factor in the determination of fair use is the effect of the use upon the potential market, which evaluates whether the use harms the original work's value or can substitute it in the market.

The four-factor test emerged because Congress recognized that no single rule could capture the complexity of creative reuse. Fair use was intentionally designed as a flexible doctrine that allowed courts to weigh context rather than apply a rigid formula. However, this flexibility also produces uncertainty when applied to AI, especially when it comes to generative models. The doctrine assumes a relatively limited number of works are being used in relatively identifiable ways, whereas AI training occurs across millions of works simultaneously and does not come with the ability to track which materials produce which outputs. Courts must use the four factors of fair use to decide whether or not such forms of automated mass copying can fit into existing legal frameworks.

Many court cases have been landmarks in shaping how these factors are applied. In *Sony Corporation v. Universal City Studios (1984)*, the Supreme Court found that the recording of television programs at home for later viewing was considered fair use as it served a noncommercial, personal purpose.[4] This case expounded the first factor of the Copyright Act regarding purpose and character of use as it made the distinguishment between a broadcasted television program and a recorded one, highlighting the fact that the recording was for harmless, personal purposes. A decade later, *Campbell v. Acuff-Rose Music (1994)* introduced the idea of

---

[2] The New York Times Co. v. OpenAI, Inc., No. 1:23-cv-11195, (S.D.N.Y. Dec. 27, 2023).
[3] 17 U.S.C. § 107 (2018).
[4] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984).

transformative use in the context of copyright, holding that 2 Live Crew's parody of Roy Orbinson's song titled "Oh, Pretty Woman" could be considered fair use because it added new expression and meaning to the original song.[5] This holding endorsed the first and third factors of the Copyright Act as the court deemed the parodic work transformative enough since its new meaning did not entirely steal from the original song's main idea but rather changed and added a new message to it, and the same can be applied to other forms of parody.[6] More recently, *Authors Guild v. Google, Inc. (2015)* endorsed Google's book-scanning project, finding that digitizing books to create a searchable database was transformative enough and did not replace the market for the originals.[7] This case primarily supported the fourth factor of the Copyright Act regarding market harm, proving that simply creating a database to search for books and not purchase them did not substitute the positions of the original books in the market. Although these precedents came before the rise of machine learning, they can still be applicable in the context of stare decisis or sole guidance. Courts must now apply the principles of purpose, nature, amount, and market effect to AI, where copying occurs on a massive scale and for the purpose of teaching digital algorithms as opposed to humans. Whether such uses are transformative enough to qualify as fair use remains one of the most pressing legal questions in the era of AI.

III. Analysis & Argument

The first factor of the Copyright Act characterizes the purpose and character of the use. In other words, it breaks down the intention of the use. One may ask: What it is for? Who will benefit from it? This factor favors uses that are transformative, noncommercial, or for public benefit. AI developers claim that training AI models with copyrighted data transforms them because the ultimate goal is to teach patterns and not reproduce the works themselves. In *Authors Guild v. Google Inc. (2015)*, the court accepted a similar argument when Google scanned millions of books to create a searchable database. AI is not solely analytical, though. It can fuel commercial products that can then generate outputs that closely resemble the originals. In *The New York Times v. OpenAI (2023)*, the Times argued that OpenAI's models could produce summaries of its articles that were nearly identical to their original versions. In such cases where the output practically mirrors existing copyrighted works, the claim of "transformative use" loses its value. The principle of purpose is no longer informational; rather, it comes substitutive. Transformative use adds new meaning; substitution replaces the original.

The second factor embedded in the U.S. copyright law regards the nature of copyrighted works and examines its weight against fair use in proportion with how creative the original work is. It is important to note that nearly all of the datasets used to train AI consist of expressive media, whether that be art, literature, journalism, or photography. In *Getty Images Inc. v. Stability AI (2023)*, countless professional photographs from the site were used without obtaining permission beforehand, many of which sported visible Getty watermarks.[8] This is a prime

---

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

[6] Rebecca Tushnet, *Economies of Desire: Fair Use and Marketplace Assumptions*, 51 WM. & MARY L. REV. 513, 528 (2009).

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

[8] Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135-GBW, (D. Del. Feb. 3, 2023).

example of how strongly this factor disfavors AI because these works were original and took creative effort to produce, therefore increasing the works' copyright protection.

The third factor does not support AI or endorse the training of LLMs, either. It monitors the amount of the original copyrighted work that is used, and training AI almost completely relies on ingesting entire works to maintain accuracy in outputs. While copying something entirely can still be fair use if it is necessary for a transformative purpose, the vast absorption of creative content generally goes far beyond what is reasonably required for innovation. The magnitude of copying in this case erodes the fairness of the act itself, making the training of LLMs seem extreme, unfair, or unethical.

Perhaps the most significant aspect of determining fair use is the fourth factor, which covers the harm done to the market. AI-generated content competes directly with the creators of the works that "inspired" it, whether that be writers, journalists, or illustrators. Those creators, some of whom are plaintiffs in current lawsuits, argue that these AI algorithms and outputs diminish the demand and value of original works. For example, in *The New York Times*' ongoing lawsuit against OpenAI and Microsoft, *The New York Times* claims that AI-generated summaries and rewrites of its articles undermine and damage their subscription, licensing, advertising, and affiliate-revenue models. Meanwhile, defendants claim that AI illuminates new opportunities and tools; however, when AI can replicate the distinctive style of a human's work or summarize a paid article for free, it undercuts the market for the original and can discourage those creators from even trying to make their own profit. Courts have long held that fair use does not extend to practices that harm a work's potential market and profits.

IV. Counterargument

AI proponents argue that training data is analogous to material used in classrooms for educational purposes and claim that restricting their access would stifle innovation and slow technological progress.

Critics, however, focus on the fact that humans' ideas come from their own minds. They do not make digital copies of a vast majority of content that can easily be used for commercial purposes. The scale and precision of AI make the two comparisons fundamentally different. Moreover, fair use has never granted complete immunity for technology solely because it is innovative. Precedents such as the one in *Authors Guild v. Google Inc. (2015)* only occurred because Google's intentions were transformative and noncommercial. It was not a profit-driven system that produced derivative content. Innovation does not instantly justify appropriation.

Proponents also argue that restricting AI training could hinder scientific research and prevent startups from competing with larger companies. These supporters frame copyright enforcement as a barrier that protects incumbents at the expense of innovation, but this argument overlooks the fact that innovation can be incentivized via licensing schemes or compensation. Technological progress and copyright protection are not exclusive and can coexist if companies acknowledge the value of the works that fuel their models.

V. Conclusion

Evaluated together, these factors lean away from justifying AI training in the context of U.S. copyright law. The purpose can always be commercial depending on how the user takes

advantage of the output, the contents of the datasets are creative, the copying is nearly total, and the harm to the market is evident. Courts should narrow the fair use doctrine in order to prevent the exploitation of creative works when it comes to AI.

Given the state of today's technological world, gaps have been exposed in the U.S. copyright law to the point where courts are forced to consider whether copying vast amounts of creative works to develop AI models that generate derivative content qualifies as fair use. There is a great deal of tension between machine learning engineers fostering innovation and the protection of creators' rights. Ultimately, courts must carefully adapt to the digital era, ensuring that progress does not come at the expense of human creativity.