

Copyright and Artificial Intelligence: A Regulatory Puzzle in Fair Use

Noah Yoder*

Abstract

Artificial Intelligence (AI) has the capacity to be an incredibly impactful technology that would not be able to exist without human-created works which implicate copyright law. The United States (U.S) and European Union (E.U.) face a regulatory asymmetry at the intersection of AI and copyright. This article discusses a potential solution to address both the regulatory void in the U.S., and the cumbersome nature of the proposed solution by the EU through means of a targeted amendment to copyright law and fair use.

Keywords: artificial intelligence, copyright law, fair use, adoption, regulation.

A Introduction

With the popularization of generative artificial intelligence (AI) following OpenAI's release of ChatGPT in late November of 2022, the uses (and misuses) have taken centre stage in customer service, education, content creation, business, healthcare and entertainment.¹ Indeed, ChatGPT alone sees roughly 3 billion visits monthly compared with roughly 80 million visits monthly in the case of other popular platforms, like Claude and Perplexity AI.² With the projected number of users expected to reach 700 million by 2030, many are left wondering how AI should be regulated to protect developer and user rights in the face of controversy.³

The wave of AI copyright litigation, in the U.S. and otherwise, presents a critical paradox: settlements provide immediate relief to individual plaintiffs but deprive the legal system of needed precedential guidance. This pattern forces AI companies to divert billions from innovation to risk management while leaving creators and developers alike uncertain about their rights.

In the scope and context of the U.S., without judicial resolution of fundamental questions about training data and fair use, AI developers face a choice between

* Noah Yoder is Senior Associate, Legal Technology Analyst. Graduate of Indiana University McKinney School of Law (2025).

1 Bernard Marr, *A Short History of ChatGPT: How We Got to Where We Are Today*, Forbes (19 May 2023), <https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today>.

2 *Most Popular AI Tools in 2025*, Exploding Topics (10 October 2025), <https://explodingtopics.com/blog/most-popular-ai-tools>.

3 *Number of AI Tool Users Worldwide From 2020 to 2031*, Statista (December 2024), <https://www.statista.com/forecasts/1449844/ai-tool-users-worldwide>.

overcautious development and liability exposure, while creators lack clear enforcement mechanisms. Moreover, if courts ultimately reject fair use defences, an overwhelming flood of copyright claims could paralyse the judiciary while stifling both technological innovation and artistic protection.

The European Union's (EU's) regulatory response, while comprehensive, creates nearly insurmountable compliance burdens for AI developers, particularly in documenting training data for large language models. This threatens to drive innovation to less regulated jurisdictions, undermining the very protections these laws seek to establish.

The current legal landscape regarding AI training data and copyright presents a stark contrast between European regulation and American uncertainty. In the U.S., there is a notable regulatory void: no comprehensive federal legislation addresses the use of copyrighted materials as AI training data. The U.S. Copyright Office has yet to complete its promised guidance on this issue, having only addressed digital replicas while explicitly deferring the question of training data to a future report.⁴ This regulatory gap has prompted executive action, with President Biden's Executive Order directing the United States Patent and Trademark Office (USPTO) to make recommendations on both AI-generated works and training data usage.⁵ While U.S. common law precedents exist that could potentially establish fair use standards for AI training, the absence of direct judicial resolution of these issues leaves both AI developers and content creators in an uncertain position.

Meanwhile, the EU has established a regulatory framework through the EU Copyright Directive and AI Act, though this framework presents its own challenges. The Directive creates two approaches to text and data mining: a robust exception for research organizations and cultural heritage institutions and a separate, broader exception which allows rightsholders to opt out.⁶ The AI Act's documentation requirements for training data, however, may prove impractical for many AI systems, particularly large language models trained on internet-scale data sets.⁷

This regulatory asymmetry – between the EU's potentially overburdensome approach and the U.S.'s current vacuum – creates a challenging environment for global AI development and deployment.

This Note advocates for a targeted amendment to U.S. copyright law establishing a clear statutory exception for the use of copyrighted materials in AI training, drawing from but simplifying the EU's approach to text and data mining and documentation. This solution would resolve the current regulatory paralysis that has left both content creators and AI developers in legal limbo, while avoiding the documentation burdens that make the EU framework impractical for large language models. By creating a categorical safe harbour for training data – rather than waiting years for case-by-case judicial development of fair use doctrine – this

4 U.S. Copyright Office, Copyright and Artificial Intelligence: Part 1: Digital Replicas Report (2024), <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf>.

5 Exec. Order No. 14,110, 88 Fed. Reg. 75,187 (1 November, 2023).

6 Council Directive 2019/790, Art. 3, 2019 O.J. (L 130) 92 (EU).

7 Council Regulation 2024/1689, 2024 O.J. (L 32) 1 (EU).

approach would democratize AI innovation beyond major tech companies, provide certainty to content creators, and prevent an overwhelming surge of copyright litigation that could strain judicial resources. This balanced framework would position the U.S. as a leader in AI development while establishing clear guidelines that other jurisdictions could adopt, ultimately fostering a more predictable global environment for AI innovation.

This Note proceeds in three parts. Part I provides foundational context, examining copyright law basics, fair use doctrine, relevant U.S. case law supporting fair use exceptions for AI training data, the EU's regulatory approach and a high-level technical overview of AI training processes. Part II analyses the inadequacies of current approaches, contrasting the EU's burdensome regulations with the U.S.'s regulatory void, and examines how existing frameworks may impede innovation and subsequently proposes a solution through a statutory exception to copyright law for AI training data, addressing both practical implementation and policy considerations. Part III concludes by considering the implications of this proposed framework for future AI development and creative rights protection in the U.S..

B Background

I United States Constitutional Basis for Copyright, Copyright Protection and Common Law Development of Fair Use

Copyright protection, within the U.S. framework, extends to:

[O]riginal works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device. Works of authorship include the following categories: (1) literary works; (2) musical works, including any accompanying words; (3) dramatic works, including any accompanying music; (4) pantomimes and choreographic works; (5) pictorial, graphic, and sculptural works; (6) motion pictures and other audiovisual works; (7) sound recordings; and (8) architectural works.⁸

The idea-expression dichotomy is crucial in understanding what is afforded a copyright protection:

[i]n no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.⁹

Section 102(b) establishes an important distinction between ideas and their expression. The classic example of lovers from rival families illustrates this

⁸ 17 USC § 102(a) (2018).

⁹ 17 USC § 102(b) (2018).

perfectly. While *Romeo and Juliet* and *West Side Story* share this underlying idea, each work's specific expression is protected. This foundation of copyright law provides the necessary context for examining the exclusion rights of copyright holders.

In the U.S. copyright exclusivity is granted from federal law with a basis in the Constitution:

[T]he owner of copyright ... has the exclusive rights to do and to authorize any of the following: (1) to reproduce the copyrighted work in copies or phonorecords; (2) to prepare derivative works based upon the copyrighted work; (3) to distribute copies or phonorecords of the copyrighted work to the public by sale or other transfer of ownership, or by rental, lease, or lending; (4) in the case of literary, musical, dramatic, and choreographic works, pantomimes, and motion pictures and other audiovisual works, to perform the copyrighted work publicly; (5) in the case of literary, musical, dramatic, and choreographic works, pantomimes, and pictorial, graphic, or sculptural works, including the individual images of a motion picture or other audiovisual work, to display the copyrighted work publicly; and (6) in the case of sound recordings, to perform the copyrighted work publicly by means of a digital audio transmission.¹⁰

These rights, however, are not absolute. The limitations on the exclusive rights described in section 106 are limited by a fair use exception within the following framework:

[F]or purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use, the factors to be considered shall include – (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 in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.¹¹

Fair use serves the purpose of “promoting science and the arts”.¹² The factors should not be treated in isolation when applied.¹³ Congress intended that courts continue common law adjudication of fair use in an effort to avoid rigid application of the copyright statute when it would “stifle the very creativity which it's designed to foster”.¹⁴

Importantly, courts have developed a critical inquiry under the first fair use factor: “whether and to what extent the new work was transformative”.¹⁵

10 17 USC § 106 (2018).

11 17 USC § 107 (2018).

12 *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994).

13 *Ibid.*, at 578.

14 *Ibid.*, at 577.

15 *Ibid.*, at 579.

Transformative is “adding something new, with a further purpose or different character, altering the first with new expression, meaning, or message”.¹⁶ A finding of transformative use furthers the goals set out in the Constitution because “such works lie at the heart of the fair use doctrine’s guarantee of breathing space within the confines of copyright”.¹⁷ If the work “supersedes the objects” of the original, it is not transformative.¹⁸

Whether the new use, however, is transformative is not solely sufficient in a determination of fair use.¹⁹ Rather, “a new meaning or message is only relevant to whether the new use served a purpose distinct from the original, or instead superseded its objects”.²⁰ While transformative use has been thought of to be sufficient in a determination of fair use, it is not; all factors must be considered and given weight.

II Overview of Generative AI as It Pertains to Copyright and Fair Use

A comprehensive analysis requires, at a 10,000-foot view, understanding generative AI’s fundamental phase of training. The other phases commonly associated with generative AI creation, learning and generation (output), are largely omitted for the sake of clarity and brevity. Our copyright and fair use analysis demands particular attention to the training phase, where the intersection between AI development and protected works is most pronounced.

At its core, AI training involves the systematic ingestion of vast corpora of existing works, including literary texts, computer code, artistic works and other copyrighted materials. This training data collection phase involves several distinct processes that implicate copyright law.

First, AI developers engage in mass collection and reproduction of works, often through automated processes that crawl the internet and other digital repositories.²¹

Second, these collected works undergo substantial change. The original works are converted into mathematical representations called ‘embeddings’ – numerical abstractions that capture patterns and relationships within the source material but that do not retain the works in their original expressive form.²²

Third, the processed data is incorporated into training data sets that serve as the foundation for the AI system’s learning phase.²³ During this incorporation, the

16 *Ibid.*

17 *Ibid.*

18 *Ibid.*

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

20 *Ibid.*, at 1275.

21 *How AI Is Trained: The Critical Role of AI Training Data*, RWS, <https://www.rws.com/artificial-intelligence/train-ai-data-services/blog/how-ai-is-trained-the-critical-role-of-ai-training-data> (accessed 17 December 2024).

22 *What are Embeddings?*, Cloudflare, <https://www.cloudflare.com/learning/ai/what-are-embeddings> (accessed 17 December 2024).

23 *AI Development Lifecycle*, Palo Alto Networks, <https://www.paloaltonetworks.com/cyberpedia/ai-development-lifecycle> (accessed 18 December 2024).

system does not store verbatim copies of the works but rather retains the mathematical patterns derived from them.²⁴

Having established the foundational concepts of AI training, copyright protection and the scope of exclusive rights – including their limitations under fair use – we now examine the judicial precedents that frame our analysis of AI training data within copyright law.

III Historical Context Surrounding the Genesis of AI Copyright Litigation

The start of litigation surrounded a case titled *Thomson Reuters v. Ross Intelligence*. The facts of the case are largely disputed, but, essentially, Ross, to train a generative AI, scraped memos (i.e. headnotes, which are copyrighted) from Reuters' well-known website, Westlaw.²⁵ The training process, Ross claims, is as follows:

First, it receives the Bulk Memos in its database. Then, it converts the plain-language entries into numerical data. Next, it feeds that data into its machine-learning algorithm to teach the artificial intelligence about legal language. The idea is that the artificial intelligence will be able to recognize patterns in the question-answer pairs. It can then use those patterns to find answers not just to the exact questions fed into it, but to all sorts of legal questions users might ask.²⁶

Because the facts are highly contested, the court decided that the issue of fair use as it relates to generative AI training, *inter alia*, must go to a jury stating that:

[W]e run into a hotly debated question: Is it in the public benefit to allow AI to be trained with copyrighted material? The value of any given AI is likely to be reflected in the traditional factors: How transformative is it? Can the public use it for free? Does it discourage other creators by swallowing up their markets? So an independent evaluation of the benefits of AI is unlikely to be useful yet, even though both the potential benefits and risks are huge. Suffice it to say, each side presents a plausible and powerful account of the public benefit that would result from ruling for it. So a jury must decide the fourth factor – and the ultimate conclusion on fair use.²⁷

In 2025 the district court amended its 2023 opinion on summary judgment by explicitly rejecting fair use. The rejection came with the caveat that it should not be thought of broadly but only in the context of headnotes and training as it relates to non-generative AI.²⁸

This case has led to the increased discussion and contested issue of how a court would treat such a fact-sensitive inquiry under the traditional fair use factors.

24 *Prepare Your Data for Machine Learning*, Integrate.io, <https://www.integrate.io/blog/prepare-your-data-for-machine-learning> (accessed 17 December 2024).

25 *Thomson Reuters Enter. Ctr. GmbH v. ROSS Intel. Inc.*, No. 20-CV-613, 2020 WL 8611175 (D. Del. 3 December 2020).

26 *Ibid.*, at 17-18.

27 *Ibid.*, at 25-26.

28 *Thomson Reuters Enter. Centre GmbH v. Ross Intelligence Inc.*, No. 1:20-cv-00613-SB, 2025 WL 458520, at 1-15 (D. Del. 11 February 2025).

Given the complexity and implication of damages, many of these questions remain unanswered. As a result, how the ageing framework of copyright law applies to new, emerging technologies is up for debate and ripe for answering.

IV Key Case Law Development in the United States

In order to adopt a Fair Use standard that is in accordance with existing law, the application of copyright law to AI training data must be understood through the lens of fair use precedents key to the United States. While no court has directly addressed the copyright implications of AI training, several key decisions provide the analytical framework for evaluating such uses. These cases trace the evolution of fair use doctrine across various technological contexts.

The seminal case *Campbell v. Acuff-Rose* established that transformative use is a crucial factor in fair use analysis. Here, a parody group, 2 Live Crew, made a song mirroring Ray Orbison's *Oh, Pretty Woman*. The court held that just because a parody may be commercial in nature, it did not render it presumptively unfair.²⁹ The court thereby established that the more transformative a new work, the less significant other factors, like commercialism, become.

Several cases have aided in the understanding of “whether and to what extent the new work [is] transformative”.³⁰ To begin with, it is necessary to note that while transformative use is important, finding it is not the be-all and end-all for a finding of fair use. There must be a balancing of all statutory factors.³¹

The first pair of cases under review, the *Authors Guild* cases, deal with mass digitization of copyrighted books. For purposes of this note, I will focus on the cases as they influence our understanding of the first, third and fourth factors of fair use analysis; the second factor is rarely dispositive.

The first case, temporally, involved *HathiTrust*, which is “[a] partnership of major academic and research institutions working to ensure that the cultural record is preserved and accessible long into the future”.³² Under that guise, *HathiTrust* worked with major universities to integrate, through digitization, more than ten million works into a database.³³ Those digital copies, *HathiTrust* claimed (*inter alia*), served to create a database for full-text searching.³⁴ The court concluded that “creation of a full-text searchable database is quintessentially transformative use”.³⁵ Further, the court concluded that full-text copying of entire works was “reasonably necessary to facilitate the services [*HathiTrust*] provides to the public and to mitigate the risk of disaster or data loss”.³⁶ Lastly, the court agreed with the district court that there was no “ ‘specific, quantifiable past [market] harm, or any documents relating to any such past harm,’ resulting from

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

30 *Ibid.*, at 579.

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

32 Overview, *HathiTrust*, <https://www.hathitrust.org/documents/HathiTrust-Overview-Handout.pdf> (accessed 17 December 2024).

33 *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 91(2d Cir. 2014).

34 *Ibid.*, at 97.

35 *Ibid.*

36 *Ibid.*, at 99.

any of ... [HaithiTrusts'] uses of their works (including full-text search)".³⁷ The court thereby concluded that "the doctrine of fair use allows [HaithiTrust] to digitize copyrighted works for the purpose of permitting full-text searches".³⁸

The second case involved Google and their efforts to digitize books for their Google Books project. Working with libraries, Google would take submissions, create a "digital scan of each book, extract a machine-readable text, and create an index of the machine-readable text of each book".³⁹ The result was a digital corpus of millions of books to enable the Google Books search engine.⁴⁰ The idea was that members of the public

c[ould] enter search words or terms ... receiving in response a list of all books in the database in which those terms appear, as well as the number of times the term appears in each book.⁴¹

The court concluded that "Google's unauthorized digitizing of copyright-protected works [and] creation of a search functionality ... are non-infringing fair uses".⁴² This was because

[t]he purpose of the copying [was] highly transformative, the public display of text is limited, and the revelations do not provide a significant market substitute for the protected aspects of the originals.⁴³

Perfect 10 v. Amazon.com's pertinent facts include that Amazon.com and Google (co-defendant) were using Perfect 10's copyrighted images for purposes of creating a thumbnail.⁴⁴ The court concluded that the creation and use of a thumbnail for purposes of aiding an electronic reference tool is transformative because "it can provide social benefit..."⁴⁵ To clarify, the court stated that

a search engine puts images 'in a different context' so that they are 'transformed into a new creation' ... [and] [t]he fact that Google incorporates the entire Perfect 10 image ... does not diminish the transformative nature.⁴⁶

Further, "it was necessary for [Google] to copy the entire image ... [because] it would be more difficult to identify it, thereby reducing the usefulness of the visual search engine".⁴⁷ Lastly, "when a work is transformative ... 'market substitution is

37 *Ibid.*

38 *Ibid.*, at 101.

39 *Authors Guild v. Google, Inc.*, 804 F.3d 202, 208 (2d Cir. 2015).

40 *Ibid.*

41 *Ibid.*, at 209.

42 *Ibid.*, at 229.

43 *Ibid.*

44 *Perfect 10, Inc. v. Amazon.com, Inc.*, 487 F.3d 701 (9th Cir. 2007).

45 *Ibid.*, at 722.

46 *Ibid.*, at 721.

47 *Ibid.*, at 724.

at least less certain, and market harm may not be so readily inferred.”⁴⁸ Under these notions, the court concluded that Amazon’s and Google’s implementation of copyrighted images for an electronic reference tool was fair use.

The following case, *Google v. Oracle*, will be discussed with brevity (to be explained later) as it concerns copying of computer code for the creation of a new application programming interface (API).⁴⁹ The court concluded that “Google’s copying of the ... API was a fair use”.⁵⁰ The court did note that since “computer programs are primarily functional it [is] difficult to apply traditional copyright concepts in that technological world”.⁵¹

MGM v. Grokster contained facts that included “shar[ing] files through peer-to-peer networks...”.⁵² The pertinent conclusion the court illuminated was that the usefulness of a technology or product, that is potentially infringing, may hinge on the amount of copyright that is taken.⁵³ Taking whole works does not preclude a finding of fair use.⁵⁴

Sony v. Universal key facts include that Sony had a technology, called a Betamax,⁵⁵ which it was providing to consumers.⁵⁶ The issue of note and purpose of the Betamax was to record television shows so consumers could watch them on their own time, apart from watching it live.⁵⁷ The court ultimately concluded that if a technology has a minimal harm “to the potential market for, or the value of ... copyrighted works [then the technology] is, therefore, capable of substantial non-infringing uses”.⁵⁸

Warner Bros. Entertainment Inc. v. RDR Books entails an encyclopaedia of the *Harry Potter* origin. The author titled the work “A-to-Z guide to the creatures, characters, objects, events, and places that exist in the world of *Harry Potter*”.⁵⁹ The author created the encyclopaedia relying only on existing *Harry Potter* works by J.K. Rowling.⁶⁰ The court notes that transformative character is diminished where a work is not consistently transformative.⁶¹ While the court recognized that it was necessary to take copious amounts of information from the *Harry Potter* works, the “verbatim copying and close paraphrasing”, weighed against a finding of fair use given the minimal transformative nature of the encyclopaedia.⁶² Lastly, given that the works were minimally transformative and overtook content, the court

48 *Ibid.*

49 *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183 (2021).

50 *Ibid.*, at 1190.

51 *Ibid.*, at 1208.

52 *Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd.*, 545 U.S. 913, 919 (2005).

53 *Ibid.*, at 953.

54 *Ibid.*

55 *Betamax*, Wikipedia, <https://en.wikipedia.org/wiki/Betamax> (accessed 18 December 2024).

56 *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 422 (1984).

57 *Ibid.*

58 *Ibid.*, at 456.

59 *Warner Bros. Ent. Inc. v. RDR Books*, 575 F. Supp. 2d 513, 524 (S.D.N.Y. 2008).

60 *Ibid.*, at 525.

61 *Ibid.*, at 544.

62 *Ibid.*, at 547.

concluded that there was a greater likelihood of market impact on Rowling's books.⁶³ In sum, and having weighed all the factors, the court did not find fair use.

A&M Records v. Napster examines a technology that takes advantage of ripping audio files and dissemination thereafter. Ripping can be described as "a computer owner ... copy[ing] an audio compact disk ('audio CD') directly onto a computer's hard drive by compressing the audio information on the CD into [an] MP3 format".⁶⁴

The court concluded that they are reluctant to find fair use if an "original work is merely retransmitted in a different medium".⁶⁵ Further, the court accepted that in some circumstances copying full works is necessary for the transformative use but declines to give this factor much weight otherwise.⁶⁶ Lastly, given evidence of market impact, the court concluded that Napster had a "deleterious effect on the present and future digital download market".⁶⁷ In weighing all the factors together, the court did not find fair use for ripping of audio files.

The *Cambridge Univ. Press v. Albert* case, most importantly, concerns a university providing students with copyrighted digital excerpts from various publishing houses, including Cambridge University Press.⁶⁸ After the court remanded on the issue of fair use, due to a botched application of the factors by the trial court, they noted that under the third factor, amount and substantiality of the portion used in relation to the copyrighted work as a whole, courts may not

consider the price of the unpaid use ... [because] a publisher's 'marginal cost for authorizing digital copies would be virtually nil., and would not vary no matter how many digital copies were authorized.'⁶⁹

Finally, in *Andy Warhol Foundation for The Visual Arts, Inc v. Goldsmith*, photographer Goldsmith took black and white photos of a then up-and-coming artist, Prince Rogers Nelson (Prince).⁷⁰ Years later, Goldsmith licensed one of those photographs "to Vanity Fair to serve as an 'artist reference for illustration' in the magazine".⁷¹ After, and "[t]o make the illustration, Vanity Fair hired pop artist Andy Warhol ... [who] created a silkscreen portrait of Prince, which appeared alongside an article about Prince [later that same year]".⁷² The conclusion, and resulting holding of this case, shocked many and will thus be quoted at length to show how the court balanced the factors of fair use analysis by attempting to rein in courts in their (often heavy) consideration of transformativeness:

The use of a copyrighted work may nevertheless be fair if, among other things, the use has a purpose and character that is sufficiently distinct from the original. In

63 *Ibid.*, at 549.

64 *A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004, 1009 (9th Cir. 2001).

65 *Ibid.*, at 1015.

66 *Ibid.*, at 1016.

67 *Ibid.*, at 1017.

68 *Cambridge Univ. Press v. Albert*, 906 F.3d 1290 (11th Cir. 2018).

69 *Ibid.*, at 1301.

70 *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258, 1264 (2023).

71 *Ibid.*, at 1267.

72 *Ibid.*

this case, however, Goldsmith's original photograph of Prince, and AWF's copying use of that photograph in an image licensed to a special edition magazine devoted to Prince, share substantially the same purpose, and the use is of a commercial nature. AWF has offered no other persuasive justification for its unauthorized use of the photograph. Therefore, the 'purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes,' §107(1), weighs in Goldsmith's favor.⁷³

Because the lower court held that the other factors (II-IV) weighed in favour of Goldsmith, and were not challenged at the Supreme Court, Warhol was not awarded a finding of fair use.⁷⁴

In sum, the cases described in the preceding paragraphs are critical in both understanding application of the fair use analysis generally and in formulating an argument in a broad fair use exception under existing precedent in the U.S..

V Existing EU Regulation as It Pertains to Copyright and Generative AI Training

The EU Copyright Directive (Directive 2019/790) contains two major exceptions (to directives 96/9/EC Arts. 5(a) and 7(1) and 2001/29/EC Art. 2, which deal with traditional copyright protections⁷⁵) that pertain to text and data mining. Text and data mining is defined as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations..."⁷⁶ Article three outlines specific exceptions for research organizations and cultural heritage institutions, each respectively defined as:

a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research: (a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or (b) pursuant to a public interest mission recognized by a Member State; in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercise a decisive influence upon such an organization ...⁷⁷ [and] a publicly accessible library or museum, an archive or a film or audio heritage institution.⁷⁸

Under this exception, these organizations can perform text and data mining for scientific research purposes, use works they have lawful access to, make copies of databases and copyrighted materials without penalty, store copies with appropriate

⁷³ *Ibid.*, at 1265.

⁷⁴ *Ibid.*

⁷⁵ Council Directive 96/9, 2001 O.J. (L 167) 10 (EC); Council Directive 2001/29, 2001 O.J. (L 167) 10 (EC).

⁷⁶ Council Directive 2019/790, Art. 2, 2019 O.J. (L 130) 92 (EU).

⁷⁷ *Ibid.*

⁷⁸ *Ibid.*

security measures, and retain materials for research verification.⁷⁹ Copyright owners have no option to opt out, but these training exceptions are limited to the organizations listed previously.

Article 4, however, applies more broadly by allowing copying and extraction of lawfully accessible works for text and data mining purposes.⁸⁰ Copyright owners, in this instance, can opt out through machine-readable means.⁸¹

The EU AI Act contains several obligations for general purpose AI model providers. The paramount, for purposes of this discussion, includes documentation requirements. Providers must

draw and keep up to date the technical documentation of the model, including *its training and testing process* ... which shall contain, *at a minimum*, the information set out in Annex XI [emphasis added].⁸²

Annex XI includes, *inter alia*, the following:

[T]he modality (e.g. text, image) and format of inputs and outputs ... a detailed description of the elements of the model referred to in point 1, and relevant information of the process for the development, including the following elements ... information on the data used for training ... the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect unsuitability of data sources.⁸³

Further, developers must assure that they are “keep[ing] up-to-date and mak[ing] available information and documentation to providers of AI systems who intend to integrate the general-purpose AI model into their AI systems”.⁸⁴ Developers must do so without generally violating intellectual property rights and shall include documentation, at a minimum, set out in Annex XII.⁸⁵ Additionally, there must be full compliance with copyright laws as established in Directive 2019/790.⁸⁶

Lastly, in pertinence to our subject, the Act lays out that developers are required to

draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.⁸⁷

In sum, the EU has established a comprehensive framework for AI development and text and data mining through two key pieces of legislation. The Copyright

79 *Ibid.*, at Art. 3.

80 *Ibid.*, at Art. 4.

81 *Ibid.*

82 Council Regulation 2024/1689, Art. 53(a), 2024 O.J. (L 32) 1 (EU).

83 *Ibid.*, annex XI.

84 *Ibid.*, Art. 53(b).

85 *Ibid.*

86 *Ibid.* Art. 53(c).

87 *Ibid.*, Art. 53(d).

Directive (2019/790) creates two distinct exceptions to traditional copyright protections: a specific exception for research and cultural heritage institutions to conduct text and data mining for scientific purposes without the possibility of rightsholder opt-out, and a broader exception allowing any entity to conduct text/data mining subject to rightsholder opt-out through machine-readable means. Complementing these provisions, the EU AI Act mandates that AI model providers maintain detailed documentation of their training processes and data sources, share necessary information with system integrators, comply with copyright laws and publicly disclose summaries of training content.

VI *Recent Developments in the U.S. Related to Generative AI and Copyright Law*

The U.S., compared to the EU, is devoid of legitimate and concrete federal regulation. Worth exploring is the Executive Order issued by the Biden Administration, the (incomplete) guidance from the Copyright Office, and, finally, the vetoed bill from California.

First, Executive Order, in pertinence, states that

[t]o promote innovation and clarify issues related to AI ... within 270 days of the date of this order or 180 days after the United States Copyright Office of the Library of Congress publishes its forthcoming AI study that will address copyright issues raised by AI, whichever comes later, consult with the Director of the United States Copyright Office and issue recommendations to the President on potential executive actions relating to copyright and AI. The recommendations shall address any copyright and related issues discussed in the United States Copyright Office's study, including ... the treatment of copyrighted works in AI training.⁸⁸

Because the Copyright Office has not issued the third part of their guidance, there is no solidified discussion on the issue of copyrighted works in AI training. The Copyright Office has, however, published the guidance as it relates to digital replicas and copyrightability of AI authored works. The discussion of copyrighted materials and AI training is relegated to a footnote which provides that “[t]he Office will address the legal issues involved in use of copyrighted works in AI systems in a subsequent Part of this Report”.⁸⁹ Aside from these two instances, there has been no additional guidance from a federal perspective. States have attempted to produce regulation, but nothing major has resulted.⁹⁰

VII *Why It Is Imperative to Develop a Statutory Fair Use Exception for Copyrighted Materials in AI Training*

The recent wave of AI copyright litigation has exposed companies to potentially billions of dollars in liability. Beyond these direct monetary risks, the astronomical

88 Exec. Order No. 14,110, 88 Fed. Reg. 75,187 (1 November 2023).

89 U.S. Copyright Office, *supra* note 4, at 17, footnote 107.

90 Governor Gavin Newsom, Veto Message on S.B. 1047 (24 September, 2024), <https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf>.

legal fees alone could have far-reaching consequences. Rather than investing in the development of AI technology – which may be one of the most transformative innovations in human history – companies could find themselves diverting vast resources to legal defence.

This legal landscape creates a dangerous incentive: AI companies may choose to relocate to jurisdictions with alternative, lesser copyright regulation. Such an exodus would threaten America's position as a global leader in technological innovation and creativity. To prevent this outcome, we must develop a balanced legal framework that protects both AI developers and users while fostering continued innovation. Avoiding over regulation and documentation requirements, as established in the EU, is also essential in maintaining interest of compliance with any further development of rules.

C Analysis

I Making the Case for Fair Use Under Existing U.S. Precedent

Fair use doctrine serves the fundamental purpose of “promoting science and the arts”.⁹¹ As AI becomes increasingly central to technological innovation, the application of fair use doctrine to AI training data presents novel legal and policy challenges. The four statutory factors of fair use – (1) purpose and character of use, (2) nature of the copyrighted work, (3) amount and substantiality used, and (4) market effect – must be analysed in the context of machine learning and computational creativity. Congress intended courts to continue common law adjudication of fair use specifically to avoid rigid application of copyright law when it would “stifle the very creativity which it's designed to foster”.⁹²

1 (Factor I) Purpose and Character of the Use: The Transformative Nature of AI Training

Courts have developed a critical inquiry under the first fair use factor, namely “whether and to what extent the new work was transformative”.⁹³ Transformative is “adding something new, with a further purpose or different character, altering the first with new expression, meaning, or message”.⁹⁴ A finding of transformative use furthers the goals set out in the Constitution because “such works lie at the heart of the fair use doctrine’s guarantee of breathing space within the confines of copyright”.⁹⁵ If the work “supersedes the objects” of the original, it is not transformative.⁹⁶

Whether the new use, however, is transformative is not sufficient in a determination of fair use.⁹⁷ Rather, a new “meaning or message [is] simply relevant

91 *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994).

92 *Ibid.*, at 577.

93 *Ibid.*, at 579.

94 *Ibid.*

95 *Ibid.*

96 *Ibid.*

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

to whether the new use served a purpose distinct from the original, or instead superseded its objects".⁹⁸

The integration of copyrighted works into AI training data sets is a fundamentally transformative use. Creation of a full-text searchable database, via full-text integration of copyrighted works, is transformative.⁹⁹ Because the original authors, copyright holders, did not create their work for the character and purpose to have a full-text searchable database, the new use is transformative.¹⁰⁰ Thus, the works, when integrated into a training corpus, serve a new purpose distinct from the original. The crux of this lies at the comparative analysis between an AI training corpus or data set and a full-text searchable database under *HaithiTrust*.

A full-text search is

a technique that finds specific information within a large corpus of text. It goes beyond keyword matching and analyzes the content of documents to identify relevant results based on the user's search query.¹⁰¹

A full-text search of a database would then look very similar to, absent the predictive analysis in generative AI output, a training corpus for generative AI.

Digitization of copyrighted books for the creation of a searchable library is transformative.¹⁰² Indeed, the purpose of Google's copying was to make "available information *about* [the copyright holders'] books...".¹⁰³ Generative AI developers may use this precedent to support the integration of copyrighted works into a training corpus to allow for dissemination of information (as a referential tool), *related to those works*, available. An electronic reference tool provides great social benefit and is thus transformative.¹⁰⁴ A court could view a training corpus as an electronic reference tool benefiting OpenAI, but it might also reject this interpretation as overly broad.

Transformative use was found in copying computer code related to APIs from one computing system to a smartphone to create a new platform (Android).¹⁰⁵ This case should be narrowly construed to computer-programming language and APIs because "computer programs are primarily functional [which] makes it difficult to apply copyright concepts in that technological world".¹⁰⁶ Strictly speaking, if the coding LLMs are implicated, *Oracle* may play a role. The training and creation of a corpus, however, draws parallels to *Oracle*: AI training involves technical transformation of creative works into mathematical representations and computational models. While *Oracle* focused on APIs, its reasoning about functional

98 *Ibid.*, at 1282.

99 *Authors Guild v. HaithiTrust*, 755 F.3d 87 (2d Cir. 2014).

100 *Ibid.*

101 *What Is Full-Text Search?*, Google Cloud, <https://cloud.google.com/discover/what-is-full-text-search> (accessed 17 December 2024).

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

103 *Ibid.*, at 206.

104 *Perfect 10 v. Amazon.com*, 508 F.3d 1146 (9th Cir. 2007).

105 *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183 (2021).

106 *Ibid.*, at 1208.

aspects of copyrighted work has broader implications for computational uses of creative content.

Where the technology has substantially non-infringing uses, it is more likely to be fair use.¹⁰⁷ A trifurcation of processes, training (developer), input (user), and output (user/AI interaction), is necessary to apply this precedent. A foundational principle emerges from this separation: AI does not act without human interaction or initiation. The training process itself is best understood through an analogy to a traditional tool: a hammer.

A hammer is made by combining raw materials – smelted iron, mined and refined from the earth, for the head, and wood, cut, shaped, sanded and stained, for the handle – to form a functioning tool. Until that hammer is picked up and has force exerted onto it, it remains lifeless; a tool is only as useful as the human makes it. The comparison to AI is just that: until a human interacts with it, it will sit there serving no purpose other than to collect dust in a user's browser.

With that in mind, the training of an AI and the output of an AI should be kept separate for purposes of applying this precedent. Training, therefore, has substantially non-infringing uses. Training AI involves highly technical aspects of converting massive amounts of text, images, video, code, etc. to numbers related to predictive analysis. Under that presumption, the copyrighted works taken for such training probably did not have the purpose and character to specifically train a generative AI. Thus, this leads to the conclusion that the new use is transformative.

Transformative character, however, is diminished when it “is not consistently transformative”.¹⁰⁸ Restricting our purview to training alone, this is helpful to generative AI developers. Under *RDR*, large verbatim portions of copyrighted material in a context similar to that of the original work will not qualify as transformative.¹⁰⁹ The context of an encyclopaedia written about a book is not the same context as a book being consumed and used by a neural network. Similarly, a training corpus compiled in the context of a neural network is far from any previous context the copyrighted materials were a part of.

Reproduction of full books for print-disabled individuals was not transformative because it serves the same purpose as the original.¹¹⁰ Nor is a work transformative if it is merely retransmitted in a different medium.¹¹¹ The context of generative AI training is a different medium from the traditionally defined online landscape, but it also serves a different purpose from the original works entirely. One may argue that a generative AI exists solely online (as do most of the

107 *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 US 417 (1984).

108 *Warner Bros. Entm't Inc. v. RDR Books*, 575 F. Supp. 2d 513, 544 (S.D.N.Y. 2008).

109 *Ibid.*, at 540.

110 *Authors Guild v. HaithiTrust*, 755 F.3d 87 (2d Cir. 2014).

111 *A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004 (9th Cir. 2001).

copyrighted training materials) but that the neural networks and training infrastructure in which the materials are encompassed, do not.

In sum, the transformative nature of AI training is supported through numerous slight extensions of existing precedent.

2 *(Factor III) Amount and Substantiality of the Portion Used in Relation to the Copyrighted Work*

The third factor asks courts whether the amount used is reasonable *in relation to the purpose of the copy* and is related to whether the work is transformative.¹¹² Generative AI developers often take whole and substantial amounts of copyrighted materials for training purposes.

Further,

a finding of fair use is more likely when small amounts, or less important passages, are copied than when copying is excessive or encompasses the most important parts of the original.¹¹³

A finding of fair use, under this precedent, is buttressed by the fact that in order to build an effective training corpus, compiling extensive and complete articles is essential for not only a useful training corpus but also for functional generative AI. Integrating extensive copyrighted work to enable search functionality is fair use.¹¹⁴

In sum, while generative AI training requires copying substantial portions of works, this comprehensive use is necessary to achieve the transformative purpose of creating functional AI systems. Thus, although the copying is substantial, it is proportional and necessary to achieve the transformative purpose, supporting a finding of fair use under the third factor.

3 *(Factor IV) Effect of the Use Upon the Potential Market for or Value of the Copyrighted Work*

When the “the second use is transformative, market substitution is at least less certain, and market harm may not be so readily inferred”.¹¹⁵ The incorporation of the copyrighted works into a training corpus is likely a transformative use and thereby unlikely to serve as a market substitute. Opponents may argue that the training corpus and output are inextricably bound inasmuch as the corpus contributing to a market substitute. As established previously, however, the creation and use of a generative AI are two separate issues.¹¹⁶

Courts are less likely to call verbatim reproductions transformative. If the new technology directs users to locations benefiting its developer, it becomes more

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

113 *Authors Guild v. Google, Inc.*, 804 F.3d 202, 221 (2nd Cir. 2015).

114 *Ibid.*

115 *Campbell v. Acuff-Rose*, 510 U.S. 569, 591 (1994).

116 See *supra* text on pages 22-23 regarding the making and utilization of a hammer.

commercial.¹¹⁷ If a developer's technology is primarily being used to replace copyrighted content, as opposed to its other, more prominent and advertised uses, it is more likely to weigh against developers. This, generally, is not the case for AI developers. Generative AIs are developed with the intent of having a new market separate from that of the copyright owner's market.

In observing factor three, "courts may not consider the price of the unpaid use".¹¹⁸ The cost of producing digital copies is almost zero.¹¹⁹ Copyright owners are technically not losing out on income resulting from reproduction in a training corpus.

In sum, factor three weighs in favour of a finding of fair use of copyrighted materials in AI training.

4 Conclusion of Fair Use Analysis of Copyrighted Materials in AI Training

The fair use analysis, while inherently fact-specific, strongly supports protecting AI training corpus integration under the doctrine. The highly transformative nature of converting copyrighted works into computational training data significantly reduces concerns about market harm. While this process typically requires copying entire works, such comprehensive use is technically necessary and proportional to achieve the transformative purpose of creating functional AI systems. These factors, taken together, provide compelling support for establishing a clear regulatory framework and national policy that explicitly recognizes AI training data integration as a protected fair use. Such recognition would provide essential clarity for technological innovation while remaining faithful to copyright law's fundamental purpose of promoting progress in science and the useful arts.

II Issues with EU Regulation and Why It Is Overburdensome

EU Regulation poses significant compliance burdens which are met with serious enforcement measures. While effective in its own right, there are several major issues with the EU framework as it relates to documenting AI training material: (i) the regulation creates anti-competitive effects and market concentration, (ii) there is a risk of innovation being forced to foreign jurisdictions due to economic and strategic implications in strict documentation requirements, and (iii) a parallel problem of effectuating the loss of U.S. policy leadership, global or otherwise.

1 Anti-Competitive Effects and Market Concentration

Under the EU framework there is a requirement to maintain a detailed technical documentation of training and testing processes according to an extensive Annex.¹²⁰ Under those requirements exists the necessity to track and document "number of

117 *Perfect 10 v. Amazon.com*, 487 F.3d 701 (9th Cir. 2007).

118 *Cambridge University Press v. Albert*, 906 F.3d 1290, 1301 (11th Cir. 2018).

119 *Ibid.*

120 Council Regulation 2024/1689, *supra* note 82.

data points, scope and main characteristics”¹²¹ of your training data. The number of data points (training material [which influences the number of parameters¹²²]) that exists in an AI training may be billions of texts, images or videos.¹²³ While keeping track of the testing process may not be burdensome, identifying each and every training material’s scope and main characteristics for training is.

Furthermore, the requirement of obligating developers to create and maintain a “sufficiently detailed summary”¹²⁴ of training content is overly broad and unclear. With an eye towards pragmatism, how would one, in light of severe penalty, offer a “sufficiently detailed summary” of *the internet*? After all, most generative AI training sets, and the potentially billions of materials contained therein, are trained on significant portions of the internet.¹²⁵

In order to comply with these vague, untested standards requires significant legal and technical expertise to ensure no stone is left unturned. As a result, companies will spend egregious amounts of money at the front end to maintain these summaries of training content to avoid fines. This perpetuates the existing environment of large, established companies ruling the developer market solely because they have the financial means. What of the talented, smaller developers with a passion to create a useful tool? Should they fall by the wayside because they do not have the upfront capital?

2 Innovation Flight Risk and Economic and Strategic Implications

Under the exceptions allotted in the EU copyright directive for research organizations and cultural heritage institutions,¹²⁶ there is an uneven playing field for developers looking to take advantage of the text and data mining exceptions. This creates not only an uneven playing field, but also when the opt-out mechanism allotted under Article 4 is added in, a deeper divide between those deemed as research organizations and those that are not.

While respecting individual copyright protection is important, allowing an exception to one predetermined subset of developers over another is laden with the propensity to send innovation away from the confines of the borders where that regulation exists. Availability of the best and most up-to-date data is essential in the creation of any highly technical or advanced tool.

Adding to this concern is the innate difficulty of keeping constant up-to-date documentation under these burdens. These complex compliance requirements create barriers to market entry and ensure that even if innovators manage to set up shop, they may be forced out. Lastly, the requirements for public disclosure all but

121 *Ibid.*

122 *Artificial Intelligence Parameter Count*, Our World in Data, <https://ourworldindata.org/grapher/artificial-intelligence-parameter-count> (accessed 17 December 2024).

123 U.S. Gov’t Accountability Office, GAO-25-107651, *Artificial Intelligence: Current and Potential Uses in the Federal Government* (2024).

124 Council Regulation 2024/1689, *supra* note 82.

125 *AI Cannibalization and Model Collapse*, The Week, <https://theweek.com/tech/ai-cannibalization-model-collapse> (accessed 17 December 2024).

126 Council Directive 2019/790, *supra* note 76.

assure that proprietary information will be disclosed to competitors, serving to hinder growth in an already uneven ecosystem.

To avoid the EU's mistake, creating a broad fair use exception without arbitrary divides in enforceability would welcome innovation and developers to make what has the potential to be the most useful technology in U.S. history.

3 *US Policy Leadership Imperative*

Put simply, the U.S. needs to be at the forefront of regulating AI. Allowing EU regulations to become de facto global standards (see GDPR¹²⁷) cedes policy leadership. The U.S. has different market conditions and innovation priorities; it has served as a home for innovators and creators for nearly three centuries. To continue to do so requires insightful and meaningful regulation according to our precedent and history. American fair use doctrine provides a foundation for a more balanced approach.

D Conclusion

Through a targeted amendment to U.S. copyright law establishing a clear statutory exception for the use of copyrighted materials in AI training that does away with the overburdensome EU documentation requirements, the regulatory paralysis that has left both content creators and AI developers in legal limbo would be resolved. A categorical safe harbour for training data, supported through fair use precedent, would democratize AI innovation beyond major tech companies, provide certainty to content creators, and prevent an overwhelming surge of copyright litigation that could strain judicial resources. This balanced framework would solidify the U.S. as a leader in global AI regulation and maintain our stature as a home of innovation and competition.

127 Regulation 2016/679, 2016 O.J. (L 119) 1 (EU).