TRAINING ON TRIAL: INSIGHTS FROM *BARTZ* AND *KADREY*

By Barbara Bruni[1]

This Article examines the first federal decisions addressing whether the use of copyrighted books to train large language models constitutes fair use: Kadrey v. Meta and Bartz v. Anthropic. Both courts held that training on lawfully acquired works was fair use, while diverging in their treatment of pirated materials and in aspects of their market-harm analysis. Using these cases as doctrinal case studies, the Article argues that generative AI exposes structural tensions within fair use doctrine. In particular, AI training unsettles the traditional relationship between transformativeness under the first factor and market substitution under the fourth. It also places pressure on the idea/expression distinction, complicates the assessment of emerging licensing markets, and brings renewed attention to public-benefit considerations in technology cases. Beyond substitution, the Article identifies a distinct "value-extraction" concern: AI systems may appropriate expressive value from copyrighted works without replenishing the incentive structure copyright is designed to sustain. The Article's primary contribution is to identify and analyze these tensions, while offering some suggestions for how courts might approach them in future cases.

INTRODUCTION

[1] Barbara Bruni is a copyright attorney and independent legal scholar. This Article began as a master's thesis in the WIPO–University of Turin LL.M. program. I am grateful to Professor Alessandro Cogo and Bryan Khan for their guidance in its early stages. I thank Professor Christopher Buccafusco for his thoughtful feedback and Amanda Levendowski Tepski for generously sharing her public-benefits chart, which guided my mapping of the public-benefit analysis. I am also grateful to the Copyright Society fellows for their editorial support.

There are currently nearly one hundred active U.S. copyright lawsuits challenging generative Artificial Intelligence (AI) models.[2] The majority of these allege that AI systems were trained on copyrighted works without authorization.[3] Several of the lawsuits also target the outputs of generative models, claiming they reproduce or closely mimic protected content.[4] These lawsuits concern works such as books, music, news articles, and visual arts, and involve the major generative AI companies (OpenAI, Microsoft, Anthropic, Midjourney, and Stability AI).[5] The lawsuits were initiated by individual authors and artists, such as Paul Tremblay[6] and Sarah Anderson,[7] as well as large media entities like the New York Times,[8] Universal Music Publishing Group,[9] and Disney.[10]

Two cases—*Kadrey v. Meta Platforms, Inc.* ("*Kadrey*") and *Bartz v. Anthropic PBC* ("*Bartz*")—have reached merits rulings in the United States District Court for the Northern District of California.[11] In both cases, the use of lawfully acquired books as training data was considered fair under 17 U.S.C. § 107.[12] The courts diverged, however,

---

[2] This litigation is tracked at ChatGPT Is Eating the World, https://chatgptiseatingtheworld.com (last visited Mar. 31, 2026) .

[3] *See, e.g.*, Complaint 52–54, *Tremblay v. OpenAI, Inc.*, No. 3:23-cv-03223 (N.D. Cal. filed June 28, 2023); Class Action Complaint 87–94, *Authors Guild v. OpenAI, Inc.*, No. 1:23-cv-08292 (S.D.N.Y. filed Sept. 19, 2023) (as filed and later updated on the docket); Complaint 44–47, *Getty Images (US), Inc. v. Stability AI, Inc.*, No. 1:23-cv-00135 (D. Del. filed Feb. 3, 2023).

[4] *See, e.g.*, Complaint 51–52, *Getty Images*, No. 1:23-cv-00135; Complaint 81–112, *Disney Enters., Inc. v. Midjourney, Inc.*, No. 2:25-cv-04820 (C.D. Cal. filed June 11, 2025); Complaint 4, 98–101, *The New York Times Co. v. Microsoft Corp.*, No. 1:23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023).

[5] *See* Lily Argyle et. al, *AI Copyright Litigation v. Licensing: An Exploration*, 72 J. Copyright Soc'y 410 (2025).

[6] *Tremblay*, No. 3:23-cv-03223 (alleging that OpenAI copied plaintiffs' copyrighted books without authorization to train ChatGPT).

[7] *Andersen v. Stability AI Ltd.*, No. 3:23-cv-00201 (N.D. Cal. filed Jan. 13, 2023) (visual artists, including Sarah Andersen and Kelly McKernan, suing Stability AI, Midjourney, and DeviantArt for copyright infringement).

[8] *N. Y. Times Co.*, No. 1:23-cv-11195 (alleging unauthorized use of NY Times articles in training OpenAI's models).

[9] *Universal Music Publ'g Grp., Inc. v. Anthropic PBC*, No. 3:23-cv-01092 (M.D. Tenn. filed Oct. 18, 2023) (claiming infringement of copyrighted song lyrics by Anthropic's Claude model).

[10] *Disney Enters.*, No. 2:25-cv-04820 (alleging Midjourney generated infringing images of Disney and Universal characters).

[11] *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-03400-VC (N.D. Cal. June 25, 2025); *Bartz v. Anthropic PBC*, No. C 24-05417-WHA (N.D. Cal. Sept. 25, 2025). This paper does not discuss *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, because that case involved a specialized legal-research application rather than a generative model. Ross's tool used machine-learning techniques to answer legal queries by drawing on structured Westlaw headnotes. The court held that the headnotes met the minimal originality requirement under *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 345–47 (1991), that Ross had copied them, and that the use was not fair, particularly because Ross's product substituted for Westlaw's functionality and threatened its market position. The court reached this outcome through a conventional substantial-similarity analysis. *Thomson Reuters Enters. Ctr. GmbH v. Ross Intelligence Inc.*, No. 20-613, slip op. at 12–23 (D. Del. Feb. 11, 2025).

[12] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 5, 40; *Bartz*, No. C 24-05417-WHA, slip op. at 1, 31.

on their treatment of pirated works within the training corpus. In *Bartz*, Judge William Alsup found such use infringing, leading to one of the largest settlements in U.S. copyright litigation history.[13] By contrast, in Judge Vince Chhabria's analysis in *Kadrey*, whether the defendant unlawfully obtained the underlying works was not determinative.[14] The two judges also departed in their treatment of asserted market harm under the fourth fair use factor. The *Kadrey* court entertained a theory of "market dilution" according to which AI-generated outputs could potentially flood the market and substitute for authors' original works.[15] Judge Chhabria ultimately dismissed the applicability of market dilution for lack of evidence on the record but suggested that such harm theory might succeed in future cases where plaintiffs can more robustly establish market effects.[16] In *Bartz*, Judge Alsup rejected the plaintiffs' market-dilution theory as too speculative to satisfy factor four, finding no causal connection between the AI-generated outputs and any potential harm to the market for the original works, and emphasizing that the complaint did not allege that the outputs were substantially similar to the plaintiffs' books.[17]

As additional cases make their way through the courts, this Article analyzes *Kadrey* and *Bartz* as doctrinal stress tests for the application of fair use to AI training. It examines where the decisions converge and diverge, how they interpret earlier technology precedents, and identifies tensions with those precedents. The central claim of this Article is that training LLMs on copyrighted works strains fair use doctrine in several respects: it calls into question the alignment between the first and fourth fair use factors because, unlike prior cases, where a transformative use was more likely to serve a distinct market, in the context of AI training market substitution may occur despite transformativeness. Relatedly, the idea/expression dichotomy comes under pressure in the context of market-dilution theories: the concern that large volumes of stylistically similar outputs may supplant demand for the originals and ultimately disincentivize creation clashes with copyright's traditional limitation to protected expression rather than stylistic features. By critiquing the analogy between human learning and machine training, this Article challenges the tendency to treat model training as functionally equivalent to human reading followed by creative production, and instead foregrounds the legally significant act of reproduction inherent in AI training. It further advances a distinct "value-extraction" theory of market failure, arguing that even in the absence of direct

---

[13] *Bartz*, No. C 24-05417-WHA, slip op. at 29−30, 31−32; Memorandum Opinion on Preliminary Approval of Class Action Settlement at 5, *Bartz v. Anthropic PBC*, No. C 24-cv-05417-WHA (N.D. Cal. Oct. 17, 2025) (describing the settlement fund). *See also*, Associated Press, *AI Startup Anthropic Agrees to Pay $1.5bn to Settle Book Piracy Lawsuit*, GUARDIAN (Sept. 5, 2025), https://www.theguardian.com/technology/2025/sep/05/anthropic-settlement-ai-book-lawsuit; ANTHROPIC COPYRIGHT SETTLEMENT WEBSITE, https://www.anthropiccopyrightsettlement.com; Blake Brittain and Mike Scarcella, *Anthropic Agrees to Pay $1.5 Billion to Settle Author Class Action*, REUTERS (Sept. 5, 2025), https://www.reuters.com/sustainability/boards-policy-regulation/anthropic-agrees-pay-15-billion-settle-author-class-action-2025-09-05.

[14] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 20.

[15] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 28.

[16] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 5, 33.

[17] *Bartz*, No. C 24-05417-WHA, slip op. at 28.

substitution, AI training may appropriate expressive value from copyrighted works without replenishing the incentive structure copyright is designed to sustain. The Article also revisits the treatment of licensing markets, cautioning against both recognizing speculative markets in a circular manner and dismissing emerging markets too readily on the ground that a transformative use need not be licensed. Finally, it interrogates the role of public-benefit considerations in AI cases, arguing that courts must articulate more clearly which conception of "progress" or dissemination they are privileging when balancing innovation against incentives for human creation.

Copyright law arose from technological advancement, namely the printing press,[18] and has evolved alongside it—from photography to film and the internet—adapting to new forms of creation and distribution.[19] This history underscores the doctrine's capacity for resilience in the face of new technologies, though the challenges posed by generative AI may differ in scale and character from those that came before it. Unlike earlier technologies, which reproduced existing works in primarily predictable ways, generative AI can produce outputs that range from near-identical reproductions to entirely novel creations, complicating courts' efforts to assess the legality of the technology as a whole while also accounting for its particular applications.[20] Its reliance on ingesting billions of copyrighted works simultaneously also marks a departure in scale from prior copyright

[18] MARK ROSE, AUTHORS AND OWNERS: THE INVENTION OF COPYRIGHT 9 (1993) (noting that the "historical emergence [of copyright] is related to printing technology"); Thomas F. Cotter, *Gutenberg's Legacy: Copyright, Censorship, and Religious Pluralism*, 91 CAL. L. REV. 323, 325 (2003) (noting that "[an] effect of the invention of printing was the development of copyright law"); Brander Matthews, *The Evolution of Copyright*, 5 POL. SCI. Q. 583, 586 (1890) (noting that "it was only after the invention of printing that an author had an awakened sense of the injury done him in depriving him of the profit of vending his own writings"). *See also Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 429–30 & nn. 11–12 (1984) ("Indeed, it was the invention of a new form of copying equipment –the printing press– that gave rise to the original need for copyright protection.").

[19] *See Sony*, 464 U.S. at 430–31 (acknowledging that copyright has repeatedly confronted new technologies and adapted to them, from player pianos to broadcasting to photocopying); Jane C. Ginsburg, *Copyright and Control Over New Technologies of Dissemination*, 101 COLUM. L. REV. 1613, 1614 (2001) ("The setting of the copyright balance is not immutable; rather, each significant technological progress may alter the balance of control between authors and users, in turn eventually prompting a new legal calibration."); Raymond Shih Ray Ku, *The Creative Destruction of Copyright: Napster and the New Economics of Digital Technology*, 69 U. CHI. L. REV. 263, 324 (2002) (observing that "[j]ust as Gutenberg's printing press threatened the dominance of scribes, peer-to-peer networking and MP3s clearly threaten the recording industry.").

[20] *See infra* Section II B 1.

cases, including earlier mass-digitization and peer-to-peer sharing precedents.[21] Further, by enabling outputs "in the style of" existing works, AI challenges the traditional line courts have drawn between unprotectable style and protected expression.[22] This, together with the speed and ease with which AI outputs can be produced, feeds into the fear of "market dilution" expressed in *Kadrey* and noted by the U.S. Copyright Office in its pre-publication report *Copyright and Artificial Intelligence, Part 3: Training*.[23] Finally, the debate over whether training harms a potential licensing market echoes but also stretches the reasoning of previous cases, raising anew the problem of circularity in defining market harm.[24]

This is hardly the first copyright confrontation that has been framed as a struggle for survival between the creative industries and emerging technologies. From the advent of the VCR and digital music-sharing platforms to the rise of search engines and online distribution networks, successive waves of technological innovation have provoked claims that the future of culture—or, conversely, of innovation itself—hung in the balance.[25] With regard to AI training, each side is invoking doomsday scenarios.

---

[21] *See, e.g.*, *Authors Guild v. Google, Inc.*, 804 F.3d 202, 208 (2d Cir. 2015) (noting that Google scanned "more than 20 million books" for its search index); *A&M Records, Inc. v. Napster, Inc.*, 114 F. Supp. 2d 896, 903 (N.D. Cal. 2000) ("Several Napster executives admitted in their depositions that they believed many of the millions of MP3 music files available on Napster were copyrighted"); LAION-5B Dataset, LAION (2022), https://laion.ai/blog/laion-5b (describing a training dataset containing over 5 billion image–text pairs); Complaint 5, *Andersen*, No. 3:23-cv-00201-WHO (alleging use of "billions of images" scraped from the internet to train Stable Diffusion). *See also* U.S. COPYRIGHT OFFICE, *Copyright and Artificial Intelligence, Part 3: Training* (pre-publication draft) (2025) 10 ("Generative AI models are well known for requiring… millions and billions of works for training purposes" (quotation marks and citations omitted)).

[22] *See infra* Section II B 1.

[23] *See supra* note 15 and accompanying text; U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3: Training*, *supra* note 21, at 10. Edward Lee traces the origin of the market dilution theory to the comments submitted to the Copyright Office following its 2023 notice of inquiry on Artificial Intelligence. Edward Lee, *Copyright Dilution Under Constitutional Scrutiny*, 25 CHI.-KENT J. INTELL. PROP. (2026) (at 9–10). *See also* U.S. COPYRIGHT OFFICE, *Artificial Intelligence and Copyright, Notice of Inquiry*, 88 Fed. Reg. 60,876 (Aug. 30, 2023), https://www.copyright.gov/ai/docs/Federal-Register-Document-Artificial-Intelligence-and-Copyright-NOI.pdf.

[24] *See infra* Section II B 1.

[25] *See, e.g.*, Home Recording of Copyrighted Works: Hearing Before the Subcomm. on Courts, Civil Liberties, & the Admin. of Just. of the H. Comm. on the Judiciary, 97th Cong. 8 (1982) (statement of Jack Valenti, President, Motion Picture Ass'n of Am.) (declaring that "the VCR is to the American film producer and the American public as the Boston Strangler is to the woman home alone"); Computer & Commc'ns Indus. Ass'n, Amicus Curiae Br. in Supp. of Respondents at 5−6, *Am. Broad. Cos. v. Aereo, Inc.*, No. 13-461 (U.S. Apr. 2, 2014) (warning that petitioners' theory "would endanger the thriving cloud computing industry" and "put billion-dollar business operations in jeopardy"). Peter K. Yu recounts Justice Stephen Breyer's remark during oral argument in *Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd.*, 545 U.S. 913 (2005): "For all I know, the monks had a fit when Gutenberg made his press." Peter K. Yu, *Monks, Medieval Scribes, & Middlemen*, 2006 MICH. ST. L. REV. 1, 3 (2006) (comparing contemporary fears of new technologies to historical reactions to innovation).

Proponents of a permissive fair-use approach to AI training warn that if licensing were required, the result would be crippling transaction costs in a market that does not yet exist, slowing the pace of innovation and potentially handing a competitive advantage to jurisdictions less encumbered by copyright, such as China.[26] Others caution that such a regime would entrench incumbent firms who have their own data and the resources to negotiate licenses, while effectively excluding smaller entrants, consolidating power in a handful of dominant players and undermining competition.[27] Commentators also predict that licensing mandates could lead to a depletion in the quality and diversity of AI models, as the costs of access restrict the breadth of training corpora.[28] Conversely, allowing unlicensed training raises equally grave concerns. Scholars and rightsholders argue that uncompensated use risks eroding the economic incentives of authors, reducing their income and professional opportunities.[29] Over time, such disincentives may

[26] *See* U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 34. The report contains a wide array of comments on the subject from scholars and interested parties collected in response to its 2023 notice of inquiry. *See, e.g.*, a16z Comment at 4 (asserting that "imposing the cost of actual or potential copyright liability on the creators of AI models will either kill or significantly hamper their development. . . . The result will be far less competition, far less innovation, and very likely the loss of the United States' position as the leader in global AI development."); Stability AI Comment at 6 (calling it "doubtful" that generative AI would be possible without the fair use defense" and arguing that "the U.S. has established global leadership in AI due, in part, to a robust, adaptable, and principles-based fair use doctrine that balances creative rights with open innovation"). *See also* Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 770 (2021) (warning that mandatory licensing would impose excessive transaction costs and impede innovation); Edward Lee, *Fair Use and the Origin of AI Training*, 63 HOUS. L. REV. 104, 106 (2025) (arguing that imposing licensing obligations on AI training could disadvantage U.S. firms relative to Chinese competitors, who receive government support); *id*. at 226 (quoting David Sacks, White House AI Czar, Post on X (June 24, 2025, 10:10 AM), https://x.com/davidsacks/status/1937558998166954092) ("There must be a fair use concept for training data or models would be crippled. China is going to train on all the data regardless, so without fair use, the U.S. would lose the AI race.").

[27] *See* U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 34 & n. 200 (citing EFF Comments at 4 ("The effect of requiring authorization would be to limit competition to companies that have their own trove of images or strike a deal with such a company, resulting in all the usual harms of limited competition (higher costs, worse service, security risks).")). Lee, *Fair Use and the Origin of AI Training*, *supra* note 26, at 224. (commenting on *Bartz v. Anthropic PBC* and noting that "Judge Alsup's approach unwittingly favors well-financed AI companies over small startups and university researchers by requiring the purchase and scanning of all books used to train AI models—which cost Anthropic 'many millions of dollars.'").

[28] *See* Lemley & Casey, *Fair Learning*, *supra* note 26, at 770; U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 86−87 (noting commenters' concerns that licensing obligations would make it difficult or impossible to secure the volume and diversity of works required for high-quality AI training, potentially forcing researchers to rely on narrower, less representative datasets and limiting meaningful participation to only the most well-resourced actors).

[29] *See* U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 33 (summarizing rightsholders' concerns that unlicensed AI training could reduce demand for their works, erode licensing markets, and undermine the economic incentives and professional opportunities necessary to sustain creative labor); Frank Pasquale and Hoachen Sun, *Consent and Compensation: Resolving Generative AI's Copyright Crisis*, 110 VA. L. REV. 207, 212–22 (2024) (arguing that

diminish the supply of high-quality, human-created works, forcing AI systems to rely increasingly on synthetic data (data produced by AI models) thereby undermining the long-term sustainability and reliability of the models themselves.[30]

Part I traces the historical development of fair use and its application to earlier waves of digital innovation. It examines how courts have distinguished functional from expressive uses, justified intermediate copying, and incorporated an increasingly expansive conception of "public benefit" into the doctrinal analysis. Part II A analyzes the first two federal decisions to address AI training on the merits—*Kadrey v. Meta* and *Bartz v. Anthropic*. It compares their treatment of the four statutory factors, highlighting points of convergence and divergence in how the courts assess transformativeness, market harm, licensing markets, unlawful acquisition, and the allocation of evidentiary burdens in the fair-use defense. Part II B situates these cases within broader doctrinal and theoretical debates. It critiques the analogy between human learning and machine training by referencing first-sale and digital-copying precedents, and examines the emerging breakdown between fair use factors one and four in the AI context. It also develops a distinct "value-extraction" theory of market failure and considers how courts should approach public-benefit analysis and institutional-competence constraints in future AI-related disputes.

I. FAIR USE IN THE DIGITAL AGE

A. Historical Development and Purpose

Fair use in the United States originated as a judge-made doctrine in the nineteenth century with the 1841 case of *Folsom v. Marsh*.[31] By the early twentieth century, fair use

---

uncompensated AI training threatens to undermine authors' livelihoods and destabilize the creative ecosystem); Daniel J. Gervais, et al., *The Heart of the Matter: Copyright, AI Training, and LLMs*, 71 J. COPYRIGHT SOC'Y U.S.A. 482, 517 (2025).

[30] *See* U.S. COPYRIGHT OFFICE, *Identifying the Economic Implications of Artificial Intelligence for Copyright Policy* 12 (2024) (observing that continued advances in AI depend on access to ever-growing bodies of training materials and warning that diminished incentives for human creators may weaken long-term AI capabilities by restricting the supply of creative inputs); Ilia Shumailov et al., *AI Models Collapse When Trained on Recursively Generated Data*, NAT. MACH. INTELL. (2024), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11269175/ (warning that over-reliance on synthetic data degrades model quality).

[31] In that case, Justice Joseph Story confronted whether the defendant's abridgment of George Washington's letters infringed the plaintiff's copyright. Judge Story rejected a mechanical rule based on the amount copied and instead articulated a balancing test, weighing "the nature and objects of the selections made, the quantity and value of the materials used, and the degree in which the use may prejudice the sale, or diminish the profits, of the original work." *Folsom v. Marsh*, 9 F. Cas. 342, 348 (C.C.D. Mass. 1841) (No. 4,901) (defendant's two-volume *Life of George Washington* copied roughly 353 pages from plaintiff's twelve-volume edition of Washington's writings). For the prehistory of fair use leading to *Folsom v. Marsh*, see generally Michael Sag, *The Prehistory of Fair Use*, 76 BROOK. L. REV. 1371 (2011) (arguing that fair use doctrine long predates *Folsom v. Marsh* and tracing its origins to eighteenth- and nineteenth-century English and American "fair abridgment" and quotation cases); Lyman Ray Patterson, *Understanding Fair Use*, 55 LAW & CONTEMP. PROBS. 249, 252–58 (1992) (surveying pre-*Folsom* case law and treatises treating fair use as an equitable limitation on copyright).

was firmly embedded in American common law, applied flexibly to accommodate teaching, scholarship, criticism, and technological change.[32] Congress eventually codified this doctrine in the Copyright Act of 1976, adopting § 107's four statutory factors modeled on Judge Story's original test in *Folsom v. Marsh*: (1) the purpose and character of the use, including whether it is commercial; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used; and (4) the effect of the use on the potential market for or value of the original.[33] The statutory list of factors is non-exhaustive, allowing judges to weigh additional considerations and tailor the fair use analysis to the circumstances of each case.[34] The doctrine, which is an affirmative defense,[35] emerged to ensure that copyright's limited monopoly does not unduly inhibit the very progress it is meant to encourage.[36] It serves to maintain the balance between incentivizing authors to create and the public interest in the free flow of information and ideas.[37]

---

[32] *See* Sag, *The Prehistory of Fair Use*, *supra* note 31, at 1283–84 (identifying roughly 40–50 U.S. decisions invoking "fair use" between 1841 and 1976).

[33] Copyright Act of 1976, Pub. L. No. 94-553, § 107, 90 Stat. 2541, 2546 (codifying fair use).

[34] *See* Senate and House Committee Reports, quoted in L.E. SELTZER, EXEMPTIONS AND FAIR USE IN COPYRIGHT: THE EXCLUSIVE RIGHTS TENSIONS IN THE 1976 COPYRIGHT ACT 19–20 (1978) (stating that fair use is an "equitable rule of reason" with no fixed definition, that Congress did not intend to "freeze" the doctrine in the statute, and that courts must continue to apply it flexibly on a case-by-case basis).

[35] *See Campbell v. Acuff–Rose Music, Inc.*, 510 U.S. 569, 590 (1994) ("Since fair use is an affirmative defense, its proponent would have the burden of proof in litigation."); *Dr. Seuss Enters., L.P. v. ComicMix LLC*, 983 F.3d 443, 459 (9th Cir. 2020) ("Not much about the fair use doctrine lends itself to absolute statements, but the Supreme Court and our circuit have unequivocally placed the burden of proof on the proponent of the affirmative defense of fair use."); *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 549 (1985) ("Fair use is a privilege in others than the owner of a copyright to use the copyrighted material in a reasonable manner without his consent."); *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1 at 11–12 (2021) (recognizing fair use as an "affirmative defense" and emphasizing its role in balancing incentives to create with the public's access to information); U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 32. *But see* Lydia Pallas Loren, *Fair Use: An Affirmative Defense?*, 90 WASH. L. REV. 685, (arguing that fair use is not a true affirmative defense because § 107 defines the limits of infringement; thus, a fair use means no infringement occurs, and the burden of persuasion should remain with the plaintiff rather than the defendant).

[36] *See Campbell*, 510 U.S. at 577–78 ("The fair use doctrine thus permits courts to avoid rigid application of the copyright statute when, on occasion, it would stifle the very creativity which that law is designed to foster…. Although such cases are generally presented as affirmative defenses, the doctrine is not simply excused infringement; rather, it is a means of balancing the interests of authors and the public.").

[37] *See Twentieth Century Music Corp. v. Aiken*, 422 U.S. 151, 156 (1975) ("Creative work is to be encouraged and rewarded, but private motivation must ultimately serve the cause of promoting broad public availability of literature, music, and the other arts."); *Sony*, 464 U.S. at 429 ("The monopoly privileges that Congress may authorize are neither unlimited nor primarily designed to provide a special private benefit."); *Fogerty v. Fantasy, Inc.*, 510 U.S. 517, 526 (1994) (copyright's ultimate purpose is "enriching the general public through access to creative works"). *See also* Wendy J. Gordon, *Fair Use as Market Failure: A Structural and Economic Analysis of the Betamax Case and Its Predecessors*, 82 COLUM. L. REV. 1600, 1614 (1982) (copyright must

Courts have long emphasized the primacy of the first and fourth fair use factors in the analysis, often treating the second and third as comparatively less significant in determining outcomes.[38] This emphasis reflects fair use's central concern with market substitution: factor four asks directly about the effect of the challenged use on the market for the original, while factor one—focusing on the purpose and character of the use—determines whether the secondary work serves a different function and thus avoids competing in the same market. Factors one and four are therefore deeply interconnected: a use with a distinct purpose and character is less likely to supplant the market for the original work.[39] In *Campbell v. Acuff-Rose Music*, the Supreme Court introduced the concept of "transformativeness" under the first factor analysis.[40] A second work is transformative when it "adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message."[41] When a secondary work is transformative, it may permissibly copy even the entirety of the original work, so long as the copying is justified by the transformative purpose.[42] A determination that a use is transformative is highly predictive of the overall outcome: empirical studies have shown that courts ruled in favor of fair use in at least 90% of cases where the use was deemed transformative.[43] Commentators have criticized this relationship as circular, noting that courts sometimes rely on a finding of transformativeness under factor one to dismiss potential market harm under factor four, rather than conducting an independent inquiry into substitution or licensing markets.[44]

---

balance "incentives to create" with "the public's interest in the dissemination and use of information").

[38] *See, e.g.*, *Campbell*, 510 U.S. at 579, 590; *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 926 (2d Cir. 1994). *See also* Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions Updated*, *1978–2019*, 10 N.Y.U. J. INTELL. PROP. & ENT. L. 1*,* 2–3 (2020) (reporting on empirical data that factors one and four continue to drive the fair use test, with factor four remaining "the single dominant factor in courts' adjudication of the fair use defense.").

[39] *See Harper & Row*, 471 U.S. at 566 (explaining that a use aimed at serving the same market as the original weighs heavily against fair use); *Sony*, 464 U.S. at 451 (considering both the purpose of the use and its impact on potential markets, finding noncommercial home recording less likely to cause market harm); *Castle Rock Ent., Inc. v. Carol Publ'g Grp., Inc.*, 150 F.3d 132, 145 (2d Cir. 1998) (observing that when the secondary use serves "the same intrinsic purpose as the original," it "seriously impairs" the market for the original).

[40] *See Campbell*, 510 U.S. at 579, 590. The Court drew the concept of "transformativeness" from Judge Pierre N. Leval's seminal article, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990).

[41] *Campbell*, 510 U.S. at 590.

[42] *Id*. at 587–89.

[43] *See* Jiarui Liu, *An Empirical Study of Transformative Use in Copyright Law*, 22 STAN. TECH. L. REV. 163, 174 (2019) (finding that from 1990 to 2017 courts ruled in favor of fair use in 94% of cases where the use was considered transformative); Clark D. Asay et al., *Is Transformative Use Eating the World?*, 61 B.C. L. REV. 905, 934–35 (2020) (analyzing all reported fair use decisions from 1991 to 2017 and finding that when courts deemed a use transformative, they also found it fair in roughly 90% of cases).

[44] *See, e.g.*, D. Asay et al., *Is Transformative Use Eating the World?*, *supra* note 43, at 905 (empirically examining how courts apply the concept of "transformative use" and finding it has

Possibly in response to the perceived overbreadth of the "transformative" doctrine,[45] the Supreme Court recalibrated the first factor in the 2023 decision of *Andy Warhol Foundation v. Goldsmith* ("*Warhol*").[46] In ruling for Goldsmith, the Court emphasized that fair use must be assessed on a use-by-use basis, narrowing the inquiry to the specific act alleged to be infringing rather than the secondary work in its entirety.[47] The Court also clarified that the mere addition of a "new meaning or message" is not dispositive; what matters is whether the defendant's use has a different purpose or character than the original, and whether that purpose avoids market substitution.[48] The amount of "transformativeness" necessary to render a secondary use fair must exceed that required merely to qualify as a derivative work; otherwise, the author's exclusive right to prepare derivative works would be rendered meaningless.[49] Post-*Warhol*, transformativeness remains an important consideration but, as the Court explained, it is "a matter of degree" rather than a binary threshold: the greater the divergence in purpose, character, and market function between the two uses, the more likely the first factor will favor fair use.[50] Conversely, when the secondary use retains a similar purpose or targets the same market—as in *Warhol*—even substantial aesthetic differences may not suffice to tilt the balance in favor of fair use.[51]

The fourth factor, the "effect of the use upon the potential market for or value of the copyrighted work," has long been regarded as the most significant, because it directly addresses the central concern for market substitution.[52] This factor extends beyond proof

become pervasive in fair-use adjudication, raising concerns about the doctrine's scope and coherence); 4 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 13.05[B][6], at 13-224.20 (2021) ("It would seem that the pendulum has swung too far in the direction of recognizing any alteration as transformative, such that [the transformative-use] doctrine now threatens to swallow fair use."). *See also* Lemley & Casey, *Fair Learning*, *supra* note 26, at 782 ("Transformative use has arguably swallowed fair use doctrine in the past twenty-five years").

[45] Jane C. Ginsburg, *Comment on Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 992 F.3d 99 (2d Cir. 2021)* (Colum. Pub. Law Rsch. Paper No. 14-691, May 4, 2021) (forthcoming in J. INTELL. PROP. L. & PRAC.) (arguing that the Supreme Court in *Warhol* appropriately recalibrated the first factor by rejecting a broad, meaning-or-message test and restoring the centrality of statutory considerations, especially purpose of the use and potential market harm).

[46] *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508 (2023). The dispute arose from a photograph that Lynn Goldsmith took of the musician Prince in 1981. *Id*. at 514. In 1984, *Vanity Fair* obtained a license from Goldsmith to use one of her portraits as an artist reference for a single illustration, which Andy Warhol produced for the magazine. *Id*. at 516. Unbeknownst to Goldsmith, Warhol also created a larger "Prince Series" of sixteen works derived from the same photograph. *Id*. After Prince's death in 2016, the Andy Warhol Foundation licensed one of those works_"Orange Prince"_to Condé Nast for use on a commemorative magazine cover without Goldsmith's authorization, prompting the litigation. *Id*. at 517–18.

[47] In *Warhol*, this meant that the Court did not evaluate the creation or display of Warhol's *Prince Series* as a whole, but only the 2016 licensing of "Orange Prince" to Condé Nast, which directly competed with Goldsmith's ability to license her photograph for the same purpose. *Id*. at 521.

[48] *Id*. at 524–26.

[49] *Warhol*, 598 U.S. at 528–29.

[50] *Id*. at 528.

[51] *Id*.

[52] *See Harper & Row*, 471 U.S. at 566 ("The fourth factor is undoubtedly the single most important element of fair use."); *Campbell*, 510 U.S. at 590 (reaffirming the centrality of the fourth factor

of actual economic harm to include potential markets that the copyright owner might reasonably exploit.[53] Courts therefore inquire whether the secondary use usurps the existing market for the original or any traditional, reasonable, or likely-to-be-developed licensing markets.[54] As the Supreme Court explained in *Sony Corp. of America v. Universal City Studios*, the inquiry requires consideration "not only of harm presented by the particular actions of the alleged infringer, but also of the impact that would result if the challenged use should become widespread."[55] If the secondary use competes in the same markets as the original, courts may presume harm without requiring direct evidence of economic loss or diminished value.[56]

B. Adapting Fair Use to Emerging Technologies

As courts have adapted fair use to evolving technological environments, they have articulated new analytical frameworks to address the challenges presented. Two concepts that bear particular relevance to AI training are the distinction between functional and expressive secondary uses of copyrighted works, and the related recognition of intermediate copying as potential fair use when necessary to enable

---

while emphasizing that all factors must be weighed together); *Texaco*, 60 F.3d at 926 (stating that "the fourth factor is the most important, and the first factor is critical as well.").

[53] *Harper & Row*, 471 U.S. at 568 ("[T]he licensing of excerpts is a traditional, reasonable, and likely to be developed market."); *Campbell*, 510 U.S. at 590 (courts must consider harm to "potential licensing markets").

[54] *Texaco*, 60 F.3d at 929–31 (holding that loss to a potential, traditional licensing market weighs against fair use).

[55] *Sony*, 464 U.S. at 451; *Harper & Row*, 471 U.S. at 568 (citing *Sony*, 464 U.S. 417); *Campbell*, 510 U.S. at 590 (reaffirming that courts must consider "whether unrestricted and widespread conduct of the sort engaged in by the defendant … would result in a substantially adverse impact on the potential market for the original" (quoting 3 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 13.05[A][4], at 13-102.61 (1993))).

[56] *Texaco*, 60 F.3d at 930 (explaining that a copyright holder "need not show with certainty that future harm will result" and that market substitution suffices to weigh against fair use); *Castle Rock*, 150 F.3d at 145 ("Where the secondary work serves the same intrinsic purpose as the original, it seriously impairs the market for the original and its derivatives."). *See also* Pamela Samuelson, *Unbundling Fair Uses*, 77 FORDHAM L. REV. 2537, 2550 (2009) (noting that courts often "presume market harm when the defendant's use competes in the same market as the original work, without requiring direct proof of lost sales").

functional uses.[57] The distinction between functional and expressive uses considers whether a secondary use targets the communicative, creative content of the original (expressive), or instead utilizes the work for a technical or informational purpose (functional).[58] Courts have found that uses serving a functional rather than expressive purpose are more likely to qualify as fair, especially when they do not substitute for the original in the market.[59]

In *A.V. v. iParadigms*, for example, the Fourth Circuit upheld the archival use of student essays within a plagiarism detection system.[60] Since the essays functioned as reference points in a technical analysis but were not republished, displayed, or exploited for their expressive content, the court characterized the use as fundamentally functional and non-substitutive.[61] The fact that there was no licensing market for student essays was also relevant.[62] Similarly, in *Authors Guild v. Google*, the Second Circuit determined that Google's full-text scanning of millions of books, which enabled keyword search, bibliographic research, and data analysis, did not appropriate or communicate the expressive content of the works.[63] Instead, the books were used as raw material to extract information about language patterns, citation contexts, and term frequency across a vast corpus aimed at enabling searches and facilitating scholarly and statistical inquiry rather than substituting for reading the books.[64] The court held that this use did not compete with the traditional market for books and did not supplant any reasonable licensing

[57] Scholars generally distinguish between "expressive" and "non-expressive" uses, while courts have more often used the term "functional" to describe uses that do not exploit a work's protected expressive content. In this paper, the terms "non-expressive" and "functional" are treated as interchangeable, with a preference for "functional" to avoid confusion. *See*, *e.g.*, *Authors Guild v. Google*, 804 F.3d at 217–25 (characterizing Google Books' search, indexing, and text-mining uses as "non-expressive," "non-substitutive," and "purely functional"); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818–19 (9th Cir. 2003) (describing a search engine's display of thumbnails as "functional" because it does not exploit the images' aesthetic content); Lemley & Casey, *Fair Learning*, *supra* note 26, at 764–70; Daniel Gervais, *The Heart of the* Matter, *supra* note 29, at 728–33; Pamela Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, 74 FORDHAM L. REV. 1473, 1484–90 (2006); David Opderbeck, *Fair Use and Machine Learning*, 76 WASH. & LEE L. REV. 1567, 1617–21 (2019). The U.S. Copyright Office likewise employs the expressive/non-expressive terminology in its discussion of AI training. *See* U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3 supra* note 21, at 44–45.

[58] *See* Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887, 1900–1906 (2024) (explaining the difference between expressive and non-expressive uses).

[59] *See* Sag, *Fairness and Fair Use in Generative AI*, *supra* note 58, at 1903.

[60] *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 639–40 (4th Cir. 2009).

[61] *A.V. v. iParadigms*, 562 F.3d at 639–40 (emphasizing that the essays were used solely to compare text strings and not for their expressive value).

[62] *Id*. at 643–44.

[63] *Authors Guild*, 804 F.3d at 217 ("Google does not sell the scans or otherwise make them available to the public.").

[64] *Id*. at 217–18 ("Google's use of the copyrighted works is highly transformative; Google uses the books' words not to supplant the books, but to enable a search function."); *id*. at 217 (noting the utility of "n-gram" research and linguistic data analysis made possible by the corpus).

market.[65] In *Kelly v. Arriba Soft*, the Ninth Circuit held that a search engine's reproduction and display of low-resolution thumbnails of copyrighted images constituted fair use because the use was "highly transformative," serving a functional, indexing purpose rather than an expressive or aesthetic one.[66] Although Arriba copied the original images entirely, the court emphasized that full copying was necessary for image recognition in search results.[67] The thumbnails were too small and low-quality to substitute for the originals, so there was no cognizable market harm.[68]

"Functionality" has emerged in case law as a compelling indication of transformativeness under the first fair use factor. Courts have consistently held that when a copyrighted work is used as a tool rather than for its communicative or aesthetic qualities, it is likely transformative.[69] In these scenarios, the original work is being mined for structural or factual properties that support unrelated, socially valuable functions: a novel's text becomes linguistic data; a photograph becomes a search index entry; software code becomes a system architecture to be understood.

Whereas a "functional" use describes the end goal (e.g., indexing, search, plagiarism comparison),"intermediate copying" is the means that often enables non-expressive uses: courts have permitted temporary or incidental reproductions of protected expression when such copying is reasonably necessary to access or analyze unprotected elements (functional interfaces, program structure, factual/linguistic features) and the resulting use does not expose the original expression.[70] In *Sega Enterprises Ltd. v. Accolade*, the court held that Accolade's copying of Sega's object code to discover the functional requirements for compatibility constituted fair use, even though the process required wholesale copying of expressive code.[71] In *Sony Computer Entertainment v. Connectix*, the Ninth Circuit upheld Connectix's copying of Sony's BIOS code to reverse-engineer it and create an emulator that allowed PlayStation games to run on non-Sony hardware, emphasizing that the copying was necessary to access unprotected functional elements.[72] The Supreme Court's decision in *Google v. Oracle America* extended this logic to software interfaces, holding that Google's reuse of Java declaring code was fair because it enabled programmers to call familiar functions on a new platform (Android smartphones), with copying limited to what was strictly needed and without market

---

[65] *Id*. at 223–24 ("There is no evidence that Google Books provides a significant market substitute for the protected aspects of the originals… We know of no Supreme Court case, or for that matter any other court case, that has found fair use to be inappropriate in the absence of significant substitutive competition.").
[66] *Kelly*, 336 F.3d at 818–19.
[67] *Id*. at 820.
[68] *Id*. at 821–22.
[69] *Authors Guild*, 804 F.3d at 217–18 (Google's scanning is "highly transformative" because it uses "the books' words not to supplant the books, but to enable a search function"); *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522–23 (9th Cir. 1992) (reverse engineering to access unprotected functional elements of code served a new purpose unrelated to the code's expressive function).
[70] James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 661–65 (2016) (summarizing case law on intermediate copying).
[71] *Sega Enters.*, 977 F.2d at 1522–23.
[72] *Sony Computer Entm't, Inc. v. Connectix Corp.,* 203 F.3d 596, 602–04 (9th Cir. 2000).

substitution.[73] Across these cases, courts have stressed that while expressive elements were copied, the purpose was to reach functional requirements, and the necessity of copying to achieve interoperability or new technological applications tipped the analysis toward fair use.[74] Furthermore, although courts recognize computer code as "writing" protected under the Copyright Act, they also emphasize its functional character and extend it less protection than that afforded to traditionally expressive works such as novels, poems, or films.[75]

Under the fourth fair use factor analysis, the issue of "circularity" has emerged in technology-related copyright cases when courts assess licensing markets.[76] This issue gained attention following the Second Circuit's decision in *Am. Geophysical Union v. Texaco* ("*Texaco*").[77] In that case, Texaco had a corporate subscription to scientific journals published by the plaintiffs.[78] Texaco's employees photocopied articles from the journals to use in their research.[79] The publishers sued, claiming lost revenues based on royalties it would have received for the additional copies.[80] The Second Circuit agreed, treating the availability of a permissions market run by the Copyright Clearance Center (CCC) as evidence of cognizable market harm under the fourth factor.[81] The dissenting Judge called out this reasoning as circular, warning that it threatened to collapse the fourth factor into a rule that any monetizable use must be licensed.[82] Although the Second Circuit limited the relevant licensing markets to those "traditional, reasonable, or likely to be developed," the decision in *Texaco* has been criticized for inviting to construe

---

[73] *Oracle*, 539 U.S. at 11–12.

[74] *See Lexmark Int'l, Inc. v. Static Control Components, Inc.*, 387 F.3d 522 (6th Cir. 2004) (intermediate copying of printer software for interoperability considered fair); *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832 (Fed. Cir. 1992) (reverse engineering analyzed in light of *Sega*; intermediate copies can be fair). *See also* Samuelson, *Unbundling Fair Uses*, *supra* note 56, at 2553–54 (observing that intermediate copying cases turn on functionality and interoperability). *But see MAI Systems Corp. v. Peak Computer, Inc.*, 991 F.2d 511, 518–19 (9th Cir. 1993), which held that loading software into RAM created an infringing copy and refused to apply a functional or interoperability-oriented fair-use rationale—a position widely regarded as in tension with the later *Sega*, *Connectix*, *Lexmark*, and *Atari* cases.

[75] *See Oracle*, 539 U.S. at 13–4 ("[S]ome works are closer to the core of copyright than others," and the Java declaring code is, "if copyrightable at all, further than are most computer programs … from the core of copyright.") (internal quotation marks and citations omitted). *See also Lotus Dev. Corp. v. Borland Int'l, Inc.*, 49 F.3d 807, 815–19 (1st Cir. 1995), aff'd by an equally divided Court, 516 U.S. 233 (1996) (menu command hierarchy is a "method of operation" under 17 U.S.C. § 102(b)); *Computer Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 706–12 (2d Cir. 1992) (filtering out functional constraints and efficiency-driven elements because they sit outside protectable expression).

[76] For discussion of circularity concerns in fair use analyses involving new technologies, see Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, *supra* note 57, at 1490–91; Edward Lee, *Technological Fair Use*, 83 S. CAL. L. REV. 797, 852 (2010).

[77] *Texaco*, 60 F.3d 913.

[78] *Id*. at 930–31.

[79] *Id*. at 918–20.

[80] *Id*.

[81] *Texaco*, 60 F.3d at 918–20.

[82] *Id*. at 937–38 (Jacobs, J., dissenting).

hypothetical or rights-holder-constructed licensing markets as dispositive of harm even where no entitlement to such licenses exists.[83] Similar claims of market harm have been raised in subsequent technology cases but, unlike in *Texaco*, courts have generally rejected such claims where the use was considered permissible under the first fair use factor, and the asserted licensing market outside the scope of the plaintiff's rights.[84] Two conflicting impulses are at play here: to protect rights-holders' existing or reasonably anticipated markets before a secondary use inflicts harm sufficient to undermine the incentives that lie at the core of copyright, and to refrain from judicially constructing new markets for rights-holders, which risks suppressing legitimate transformative uses and undermining the public interest in follow-on innovation. Courts have repeatedly warned that copyright should not be used to "leverage" rights into markets that were never part of the author's incentive, especially when doing so would chill socially valuable uses.[85] In situations involving emerging technologies where markets are being established for the first time, it can be difficult to determine which activities give rise to a cognizable market, particularly when the technology enables uses that rights-holders had not previously pursued.

Finally, considerations of the public benefit can carry significant weight in a fair use analysis, especially when the record on market harm is ambiguous.[86] But how should as broad a concept as "public benefit" be interpreted? Its constitutional roots lie in the

---

[83] *Id*. at 931–32. *See* Samuelson, *Unbundling Fair Uses*, supra note 56, at 2585 (noting criticism on *Texaco's* market analysis and circularity); Lydia Pallas Loren, *Redefining the Market Failure Approach to Fair Use*, 5 J. INTELL. PROP. L. 1, 36–41 (1997) (warning against letting rights-holder-created licenses dictate factor four); Mark A. Lemley, *Should a Licensing Market Require Licensing?* 70 LAW & CONTEMP. PROBS. 185, 189–191 (2007) (arguing that recognizing a licensing market simply because a rights holder offers licenses creates circularity in factor-four analysis).

[84] *See, e.g.*, *Connectix*, 203 F.3d at 607–08 (rejecting Sony's claimed market for licensing emulators as a basis for market harm because "Sony cannot preempt a transformative market by licensing it away"); *Kelly*, 336 F.3d at 821 (declining to recognize an asserted market for licensing thumbnails and cautioning against treating a rightsholder's willingness to license as dispositive of harm); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1168–69 (9th Cir. 2007) (rejecting asserted licensing market for mobile-image thumbnails as circular where the market was created only after the defendant's use); *Authors Guild*, 804 F.3d at 223–24 (holding plaintiffs could not claim market harm by positing a licensing market for full-text search because a copyright owner "cannot prevent" a fair use by "creating a licensing market for it").

[85] *See Campbell*, 510 U.S. at 592 (rejecting a presumption of market harm and noting that copyright owners cannot claim control over "transformative" or "critical" markets that were never part of their legitimate licensing expectations); *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 614–15 (2d Cir. 2006) (holding that rights-holders are not entitled to "transformative markets," and that copyright does not grant a monopoly over all conceivable downstream uses).

[86] *See* Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, *supra* note 57, at 1490 ("When evidence of market effects is more equivocal, as in *Google LLC v. Oracle America, Inc.*, courts tend to weigh evidence of harms and benefits in relation with other factors, including public benefits"). The discussion of "public benefit" in this Section encompasses related judicial concepts such as the "public interest," "public welfare," and broader "societal benefits," which courts often invoke interchangeably when assessing fair use.

Progress Clause, which empowers Congress "[t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."[87] The idea of public benefit is also implicit in the statutory preamble to § 107, which lists paradigmatic fair use purposes such as criticism, comment, news reporting, teaching, scholarship, and research, as well as in Congress's decision to codify fair use in open-ended terms, emphasizing that the four factors are not exclusive and must be read in light of copyright's constitutional purpose.[88] This reflects the utilitarian theory underlying copyright, which aims to balance private incentives for creation with the broader societal goal of enriching the public domain and fostering the dissemination of knowledge.[89]

Malla Pollack's historical study of the Progress Clause shows that "progress" in the 1780s overwhelmingly referred not to qualitative improvement but to *spread*—the diffusion and public circulation of knowledge—and that Congress may create exclusive rights only insofar as they advance dissemination rather than entrench private control.[90] Wendy Gordon similarly reframes "harm" as a public-regarding concept: lost licensing

---

[87] U.S. Const. art. I, § 8, cl. 8. *See also Campbell*, 510 U.S. at 575–76 ("From the infancy of copyright protection, some opportunity for fair use of copyrighted materials has been thought necessary to fulfill copyright's very purpose, 'to promote the Progress of Science and useful Arts.'"); David Nimmer, *"Fairest of Them All" and Other Fairy Tales of Fair Use*, 66 LAW & CONTEMP. PROBS. 263, 280–81 (2003) (linking fair use to constitutional limits, stressing its role as a structural safeguard to align copyright with the Clause's purpose); Leval, *Toward a Fair Use Standard*, *supra* note 40 (arguing that fair use must be understood as serving the constitutional goal of copyright_encouraging progress_rather than as a grudging exception to authors' rights). Sometimes U.S. Const. art. I, § 8, cl. 8 is referred to as the "Copyright Clause" or, more broadly, the "Copyright and Patent Clause." Malla Pollack has argued that the provision is more accurately described as the "Progress Clause," because the promotion of progress was the Framers' express objective. *See generally* Malla Pollack, *What Is Congress Supposed to Promote?: Defining "Progress" in Article I, Section 8, Clause 8 of the United States Constitution, or Introducing the Progress Clause*, 80 NEB. L. REV. 754, 755 & n.1 (2001).

[88] *See supra* note 34 and accompanying text. *See also* Amanda Levendowski Tepski, *Fairer Public Benefit in Copyright Law,* 47 CARDOZO L. REV. 119, 128 (2025).

[89] *See Sony*, 464 U.S. at 429–30 (emphasizing that "the monopoly privileges that Congress may authorize are neither unlimited nor primarily designed to provide a special private benefit," but are instead "a means by which an important public purpose may be achieved," such that copyright's "limited grant is intended to motivate the creative activity of authors and inventors … and to allow the public access to the products of their genius after the limited period of exclusive control has expired"); *id.* at 429 (quoting *Fox Film Corp. v. Doyal*, 286 U.S. 123, 127 (1932), for the principle that "copyright law, like the patent statutes, makes reward to the owner a secondary consideration" and that "[t]he sole interest of the United States and the primary object in conferring the monopoly lie in the general benefits derived by the public from the labors of authors"); *id.* (citing *United States v. Paramount Pictures, Inc.*, 334 U.S. 131, 158 (1948), for the view that any reward to authors "serves to induce release to the public of the products of [their] creative genius"); *id.* at 430 (explaining that the Constitution assigns Congress the task of defining the scope of copyright's "limited monopoly" in order to ensure "appropriate access" for the public, a balance continuously adjusted in light of "society's competing interest in the free flow of ideas, information, and commerce"). *See also*, Pollack, *What Is Congress Supposed to Promote?*, *supra* note 87, at 759–67; Gordon, *Fair Use as Market Failure*, *supra* note 37, at 1614–18.

[90] *See* Pollack, *What Is Congress Supposed to Promote?*, *supra* note 87, at 755–57.

revenues do not constitute copyright-relevant injury unless they materially impair incentives to create or impede public access to knowledge.[91] Christopher Sprigman advances a complementary, effects-based view, contending that courts should adopt a "rule of reason" that evaluates whether challenged uses actually diminish creative incentives or instead enhance competition, follow-on creativity, and dissemination.[92] Most recently, Amanda Levendowski Tepski demonstrated that public benefit has been a recurring, though undertheorized, element in fair use cases involving novel technologies, where courts have routinely weighed values such as expression, knowledge, competition, and efficiency as part of that inquiry.[93] Levendowski Tepski traces the first explicit mention of "public benefit" in the 1966 case of *Rosemont Enterprises v. Random House*, and, since then until 2023, counted thirty-eight judicial decisions that invoked the public benefit, of which nineteen concern novel technologies.[94] The most significant case until now has been *Google v. Oracle*, which explicitly asked courts to consider the public benefit, where appropriate, under the fourth fair use factor.[95] Both the concept of

[91] Wendy J. Gordon, *The Concept of "Harm" in Copyright*, in RESEARCH HANDBOOK ON THE FUTURE OF EU COPYRIGHT 459, 460–66 (Estelle Derclaye ed., 2009) (arguing that lost licensing revenues constitute harm only if they undermine incentives to create or reduce public access to knowledge).
[92] Christopher Sprigman, *Copyright and the Rule of Reason*, 7 J. TELECOMM. & HIGH TECH. L. 317, 320–25 (2009) (advocating a flexible, effects-based assessment of whether a use actually reduces creative incentives or instead enhances competition, follow-on creativity, and dissemination).
[93] *See* Levendowski Tepski, *Fairer Public Benefit in Copyright Law, supra* note 88 at 124–25. Levendowski Tepski defines "novel technologies" as "either technologies of a court's first impression … or unique uses of existing technologies," explaining that a technology is of "first impression" when a court confronts it for the first time, with no prior precedent addressing its operation or legal implications. *Id.* at 134–35.
[94] Levendowski Tepski, *Fairer Public Benefit in Copyright Law, supra* note 88 at 130, 133.
[95] *Oracle*, 593 U.S. at 13–15. In *Google v. Oracle*, the Court emphasized that the fourth factor "requires a court to consider the public benefits the copying will likely produce," explaining that fair use must account for the ways in which secondary uses can "further the broader public interest in creativity." *Id.* at 395–96. *See also* Levendowski Tepski, *Fairer Public Benefit in Copyright Law, supra* note 88, at 133. Before *Google v. Oracle*, courts had taken into account the public benefit under the first fair-use factor (*see*, *e.g.*, *Perfect 10*, 508 F.3d at 1165 ("We conclude that the significantly transformative nature of Google's search engine, particularly in light of its public benefit, outweighs Google's superseding and commercial uses of the thumbnails in this case."); *Texaco*, 60 F.3d at 923 ("The greater the private economic rewards reaped by the secondary user (to the exclusion of broader public benefits), the more likely the first factor will favor the copyright holder and the less likely the use will be considered fair."); *Kelly*, 336 F.3d at 818–19 (similarly concluding that the first factor weighs in favor of Arriba due to the public benefit of the search engine and the minimal loss of integrity to Kelly's images)); and the fourth fair use factor (*see*, *e.g.*, *Sony*, 464 U.S. at 454–55 (emphasizing that copyright's purpose is to serve the public and relying on the public benefits of time-shifting in evaluating market harm); *Mattel, Inc. v. Walking Mountain Prods.*, 353 F.3d 792, 804 (9th Cir. 2003) ("Finally, the public benefit in allowing artistic creativity and social criticism to flourish is great.")). Other cases recognize public benefits more generally without allocating it under a specific factor. *See, e.g.*, *Hachette Book Grp., Inc. v. Internet Archive*, No. 20-4160, 2023 WL 2623787, at *25–26 (S.D.N.Y. Mar. 24, 2023) (noting that controlled digital lending provides certain public benefits but holding that such benefits cannot override the statutory market in the fair-use analysis); *Sega*, 977 F.2d at 1523–24

"benefit" and the relevant "public" have evolved over time with shifting social norms and technological change. In earlier jurisprudence, the notion of public benefit was construed narrowly, rooted in the traditional goal of fostering expressive creation and the dissemination of works to the public. In *Campbell v. Acuff-Rose Music*, the Supreme Court found that 2 Live Crew's parody of Roy Orbison's "Oh, Pretty Woman" served the public interest by introducing new expression and critique, highlighting that parody "can provide social benefit, by shedding light on an earlier work, and, in the process, creating a new one."[96] In *Sony Corp. v. Universal City Studios*, the Court treated time-shifting as beneficial not because it advanced technology per se, but because it facilitated greater access to protected expression.[97] The emphasis remained on dissemination rather than more indirect benefits to society. In more recent copyright cases involving new technologies, courts have gradually expanded their understanding of public benefit beyond the traditional focus on authorship and dissemination. In *Google v. Oracle America*, the Supreme Court treated Google's reimplementation of Java API declarations as fair use, emphasizing how it enabled innovation and competition in mobile platforms, contributions to the public that extended well beyond the creation of new works in a narrow sense.[98] Likewise, in *Sega Enterprises v. Accolade*, the Ninth Circuit held that Accolade's intermediate copying of Sega's code to reverse-engineer the Genesis system was a fair use, even though it involved exact reproduction of protected software elements.[99] The court emphasized that the purpose of the copying—to gain access to functional elements necessary for interoperability—served the public interest in fostering competition and innovation in the video game market, notwithstanding the absence of a new expressive work.[100] The benefit to the public in *Sega Enterprises v. Accolade* and *Google v. Oracle America* was systemic: enabling an open and competitive technological environment, rather than simply promoting expressive outputs. The relevant "public" has broadened in parallel: in expressive-work cases, courts have sometimes focused on the copyright-relevant audience, as in *Harper & Row*, where premature disclosure to the reading public was seen as undermining the market for Ford's memoir.[101] By contrast, as seen above in *Sega Enterprises v. Accolade* and *Google v. Oracle America*, courts expanded the notion of the relevant public to include beneficiaries far beyond the immediate users of the technology, recognizing innovation-driven gains that extended beyond the contexts of the cases.[102] This shift reflects the flexibility of the fair use

(recognizing the significant public interest in fostering competition and access to functional information necessary for interoperability).

[96] *Campbell*, 510 U.S. at 579.

[97] *Sony*, 464 U.S. at 454–55.

[98] *Oracle*, 593 U.S. at 12–13.

[99] *Sega*, 977 F.2d at 1522–28.

[100] *Sega*, 977 F.2d at 1523–27.

[101] *Harper & Row*, 471 U.S. at 566–69. *See also Salinger v. Colting*, 607 F.3d 68, 74, 83–84 (2d Cir. 2010) (assessing likely market substitution among readers of J.D. Salinger's works and emphasizing the interest of copyright holders in controlling first publication); *Twin Peaks Prods., Inc. v. Publ'ns Int'l, Ltd.*, 996 F.2d 1366, 1377–78 (2d Cir. 1993) (considering harm to the market for television episodes among the show's viewing audience); *Castle Rock*, 150 F.3d at 145–46 (focusing on the fan-based market for derivative works associated with the *Seinfeld* audience).

[102] *See supra* notes 98 and 100 and accompanying text.

doctrine and a broader judicial willingness to see copyright not only as a system for promoting expression, but also as a legal infrastructure that must adapt to accommodate technological development, interoperability, accessibility, and other systemic public goods. As technologies grow more complex and their societal impact more diffuse, courts have recalibrated the fair use inquiry to account for these broader benefits, even in the presence of some market harm and potential disruption.[103] In several cases, courts have had to balance a more traditional public benefit of disseminating and incentivizing expressive works against innovation-oriented benefits such as searchability, accessibility, and technological development. The decisive consideration has generally been market substitution: in *A&M Records v. Napster* and *MGM Studios v. Grokster*, dissemination-based interests prevailed because the technologies enabled direct market substitution, whereas in cases like *Kelly v. Arriba Soft* and *Authors Guild v. Google*, courts favored innovation-based public benefits because they understood the challenged uses not to function as substitutes for the underlying expressive works.[104]

II. AI TRAINING CASES: RECENT DECISIONS

A. Introduction to the Cases

To date, two federal district court decisions have directly addressed fair use in the context of generative AI training: *Kadrey v. Meta*, and *Bartz v. Anthropic*.[105] In *Kadrey*, Judge Chhabria granted summary judgment to Meta, holding that training its Llama model on copyrighted books constituted fair use; in *Bartz*, Judge Alsup likewise found Anthropic's training of its Claude model to be fair use with respect to lawfully acquired works, but that the use of pirated books could not qualify as fair use.[106] This section examines the decisions in parallel, moving through the four fair use factors and noting where the courts invoked considerations of public benefit.

In *Bartz*, Judge Alsup considered claims by authors that Anthropic had trained its "Claude" model on a corpus of books, some legitimately purchased and others obtained from pirated sources.[107] In *Kadrey*, the plaintiffs—over a dozen authors including Richard Kadrey and Sarah Silverman—alleged that Meta had used their books, including materials sourced from unauthorized "shadow libraries," to train its Llama model without

---

[103] *See Oracle*, 593 U.S. at 9–18 (emphasizing that fair use analysis considers public benefit and innovation, not solely market harm); *Sega*, 977 F.2d at 1524–27 (holding reverse engineering of software constitutes fair use in part because it promotes competition and creative innovation).

[104] *See A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015–19 (9th Cir. 2001) (holding that peer-to-peer file sharing displaced licensed consumption and thus caused cognizable substitution); *Grokster*, 545 U.S. at 928–37 (same); *cf. Kelly*, 336 F.3d at 820–22 (holding that search-engine thumbnails did not substitute for high-resolution photographs and emphasizing the public benefits of image search); *Authors Guild*, 804 F.3d at 223–24 (holding that full-text scanning and snippet search did not serve as market substitutes for books and yielded substantial public benefits).

[105] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 26–28; *Bartz*, No. C 24-05417-WHA, slip op. at 20–21.

[106] *Bartz*, No. C 24-05417-WHA, slip op. at 20–21.

[107] *Id*. at 1–9.

consent.[108] Under the first factor, the court in *Bartz* held that training LLMs on copyrighted works was "spectacularly" and "quintessentially" transformative.[109] Anthropic used copies of books to iteratively map statistical relationships among text fragments, enabling its Claude model to generate responses to new prompts in a manner analogous to human learning. Although plaintiffs argued that the model "memorized" expressive elements of their works, the court reasoned that even if LLMs distilled grammar, composition, and style from thousands of texts, this was akin to a student who reads and emulates the best prose—something copyright does not prohibit.[110] Copyright protects expression, not "methods of operation, concepts, [or] principles" embodied in a work.[111] With regard to the pirated books, even though some were later used in the training process, the court followed *Warhol*'s mandate to assess each use separately.[112] The court treated the initial act of downloading millions of books from pirate sites to build a permanent internal library as its own separate use, distinct from training, and found that it was not transformative: it merely substituted for acquiring lawful copies and was motivated by a desire to avoid "legal/practice/business slog."[113]

In *Kadrey*, Judge Chhabria determined, similarly to Judge Alsup, that Meta's use was "highly transformative" under the first factor: LLM training repurposes books into a generative tool distinctly different from their original form.[114] Judge Chhabria reminded that the outputs were not regurgitated and that even though they may be similar, style is not copyrightable.[115] He added that, although Meta's tool is commercial, commercialism is not dispositive and is less important if the use is highly transformative.[116] Judge Chhabria rejected the plaintiffs' argument that unlawful acquisition categorically prevents a finding of fair use, noting that "whether a given use was made in good or bad faith wouldn't seem to affect the likelihood of that use substituting for the original."[117] The court further observed that downloading from shadow libraries could, in theory, affect the analysis if it supported those infringing services or worsened market harm, but the plaintiffs provided no evidence of such effects.[118]

As to the second and third factors, while acknowledging that the authors' works were expressive in nature (a factor typically weighing against fair use), both courts found that copying entire works was reasonable and necessary for achieving the LLM's transformative function.[119] Notably, under the second factor Judge Chhabria rejected

---

[108] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 26–28, 8–15.
[109] *Bartz*, No. C 24-05417-WHA, slip op. at 11, 13.
[110] *Id*. at 12.
[111] *Id*. at 13.
[112] *See supra* note 47 and accompanying text.
[113] *Bartz*, No. C 24-05417-WHA, slip op. at at 20.
[114] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 16. Interestingly, Judge Chhabria entertains and dismisses an argument in the plaintiffs' law professor *amici* arguing that "Meta's use has the same purpose and character as the books because an LLM training on a book is akin to a human reading one." *Id*. Judge Chhabria clarified that this is not how a human reads a book. *Id*. at 17.
[115] *Id*. at 18.
[116] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 19.
[117] *Id*. at 20.
[118] *Id.* at 20–21.
[119] *Bartz*, No. C 24-05417-WHA, slip op. at 25; *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 23.

Meta's argument that it copied plaintiffs' books only to access their "functional elements," and dismissed Meta's analogy to the Ninth Circuit "intermediate copying" cases, where reverse engineering of computer code was permitted to reach unprotected elements.[120] Judge Chhabria emphasized that unlike software, LLM training depends directly on creative expression such as word order, grammar, and syntax, which are protectable expressive choices.[121] He pointed out that books make for especially valuable training data, because they are generally well-written and use proper grammar, are long and consistent.[122] He made an important differentiation with the intermediate copying cases, which extracted non-expressive information: the database in *Authors Guild v. Google* was content-agnostic, meaning that its search functionality was not dependent on the content of the books, whereas Llama's ability to create high quality outputs is dependent on "high-quality expression."[123]

In *Bartz*, under factor four, Judge Alsup deemed that "[t]he copies used to train specific LLMs did not and will not displace demand for copies of Authors' works, or not in the way that counts under the Copyright Act."[124] Concerns that LLMs would generate new works competing with plaintiffs' were analogized to complaints that educating schoolchildren to write would produce competing works—precisely the sort of competition copyright does not bar.[125] No exact copies or infringing knockoffs of plaintiffs' works had been made available to the public, and plaintiffs remained free to sue if such facts later developed.[126] When considering the plaintiffs' alleged licensing market, the court assumed, by viewing the facts in the Authors' favor, that such a market could exist but in mentioning the circularity problem, emphasized that fair use could not shield the Authors from uses they were not entitled to control in the first place.[127] As to the pirated books, Judge Alsup found factor four decisively in plaintiffs' favor: each pirated copy obtained by Anthropic directly substituted for a lawful sale.[128] He further stressed that if such piracy were deemed fair whenever a developer asserted a transformative downstream purpose, the result would "destroy the publishing market."[129]Anthropic also retained the pirated copies indefinitely for broad internal use,

---

[120] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 23 (citing *Harper & Row*, 471 U.S. at 548). *See also supra* notes 70−74 and accompanying text.

[121] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 23−24 ("Word order, word choice, grammar, and syntax are how people express their ideas. *See Harper & Row*, So even though LLMs may only learn about "statistical relationships," those relationships are the product of creative expression. This is true even though, as discussed earlier, Llama consumes that expression in a different way than a person would.").

[122] *Id*. at 9.

[123] *Id*. at 24 (distinguishing *Authors Guild*, 804 F.3d 202).

[124] *Bartz*, No. C 24-05417-WHA, slip op. at 28.

[125] *Id*.

[126] *Id*.

[127] *Id*. ("None of the cases cited by Authors requires a different result. All contemplated losses of something the Copyright Act properly protected — not the kinds of fair uses for which a copyright owner cannot rightly expect to control."). *See also supra* notes 68–75 and accompanying text.

[128] *Bartz*, No. C 24-05417-WHA, slip op. at 30.

[129] *Id.* (quoting Anthropic's hearing statement).

even when they were no longer (or were never) needed for training, which reinforced that these copies functioned as long-term substitutes for lawful acquisitions.[130]

Kadrey departed most markedly from *Bartz* under the fourth factor, at least in reasoning. Judge Chhabria categorically rejected the analogy between human learning and computer learning, remarking on the speed and ease with which LLMs can make copies in contrast to the fallacy of human copying and suggesting that such capabilities can bring about a different type of competition.[131] He introduced the possibility of a "market dilution" theory tailored to generative AI.[132] He observed that, unlike in traditional derivative work disputes, training a large language model could enable the creation of millions of outputs that collectively "flood" the market and indirectly substitute for the authors' works, thereby diluting the originals' economic value.[133] He stated that, given the importance of the fourth factor in the fair use analysis, he could imagine a scenario where a use is highly transformative but nonetheless fails on fair use because of the effect on the market of the original work.[134] He dismissed the argument, contained in the Meta *amici* Brief, that substitution can occur only when the outputs are themselves infringing, by saying that "indirect substitution is still substitution."[135] Judge Chhabria acknowledged that this argument "could decisively win the fourth factor" in a future case but stressed that the plaintiffs had not preserved or substantiated it with evidence.[136] Thus, while market dilution may remain a tool for future litigation, the absence of an evidentiary record compelled the court to resolve factor four in Meta's favor. The plaintiffs' case instead turned on narrower claims of regurgitation and lost licensing opportunities. Both were rejected: regurgitation for lack of proof of meaningful copying, and lost licensing opportunities because a hypothetical market for AI training licenses is inherently circular, existing only if training is first deemed infringing.[137] Recognizing such a market, Judge Chhabria reasoned, would collapse the fair use inquiry into compulsory licensing and "swallow fair use whole."[138] Judge Chhabria further added that the decision was narrowly tailored to the case at hand and did not endorse the legality of wholesale use of works for AI training in all circumstances.[139] Unlike in *Bartz*, where assembling the library was treated as its own non-transformative use, in *Kadrey* Judge

---

[130] *Id.* at 30–31.

[131] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 3 ("…when it comes to market effects, using books to teach children to write is not remotely like using books to create a product that a single individual could employ to generate countless competing works with a miniscule fraction of the time and creativity it would otherwise take). Judge Chhabria adds: "This inapt analogy is not a basis for blowing off the most important factor in the fair use analysis." *Id*.

[132] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 26.

[133] *Id*. at 26–27.

[134] *Id*. at 26 ("… given the fourth factor's importance, it's easy to imagine a situation in which a secondary use is highly transformative but the secondary user nonetheless loses on fair use because allowing people to engage in that kind of use would have too great an effect on the market for the original work.").

[135] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 30-31.

[136] *Id*. at 27.

[137] *Id*. at 28.

[138] *Id*. (citing *Bill Graham Archives*, 448 F.3d at 615).

[139] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 2–3.

Chhabria held that Meta's downloads from shadow libraries had to be evaluated "in light of [their] overall purpose," not in isolation.[140] Even if the downloads were a separate act, he explained that their purpose was the highly transformative use of training LLMs.[141]

It is worth pausing on the issue of the burden of proof in a fair use defense. As fair use is an affirmative defense, the burden of proof rests with the party asserting it.[142] In *Kadrey*, Meta not only raised fair use but also moved for summary judgment, meaning that all disputed facts had to be read in the plaintiffs' favor.[143] Accordingly, Meta bore the burden of presenting evidence that its copying did not threaten to substantially harm the market for the plaintiffs' books.[144] Judge Chhabria acknowledged that Meta had not conclusively shown that its copying would never harm plaintiffs' markets in the future.[145] Yet, he reasoned that when a defendant introduces evidence suggesting a lack of market harm, and the plaintiffs fail to counter with empirical evidence of such harm, the fourth factor should weigh in the defendant's favor.[146] Judge Chhabria rejected the plaintiffs' reliance on *Hachette Book Group v. Internet Archive* in inferring harm, noting that in that case the challenged copies functioned as direct market substitutes for the originals, unlike the training copies at issue in *Kadrey*.[147]

There is no discussion of the public benefit in *Bartz*, and only limited consideration of it in *Kadrey.* In the latter case, under the first factor, Judge Chhabria emphasized Llama's positive function of "the furnishing of valuable information on any subject of public interest," without allowing the public access to the original work.[148] Under the fourth factor, Judge Chhabria recalled the Supreme Court's instruction in *Google v. Oracle*, that "also relevant to this factor are the public benefits the copying will likely produce."[149] He understood those benefits as the promotion of the "broad public availability of literature, music, and the other arts," and found that Llama's abilities to create new expression slightly favored Meta.[150] At the same time, when evaluating potential harms, he rejected Meta's argument that recognizing a licensing market for training data would impede the continued development of generative AI, pointing out Meta's staggering projected revenues over the coming decade.[151]

---

[140] *Id.* at 37 n.14.
[141] *Id*.
[142] *See supra* note 30 and accompanying text.
[143] *See Anderson v. Liberty Lobby, Inc.*, 477 U.S. 242, 255 (1986) (In a summary judgment proceeding, "[T]he evidence of the nonmovant is to be believed, and all justifiable inferences are to be drawn in his favor.").
[144] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 26.
[145] *Id*.
[146] *Id*. at 35 ("All the plaintiffs presented is speculation, and speculation is insufficient to raise a genuine issue of fact and defeat summary judgment.")
[147] *Id*. (citing *Hachette Book Group, Inc. v. Internet Archive*, 115 F.4th 163 (2d Cir. 2024)).
[148] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 17 (citing *Romanova v. American Broadcasting Cos., Inc.*, 138 F.4th 110, 115 (2d Cir. 2025)).
[149] *Id*. at 26 (citing *Oracle*, 593 U.S. at 35).
[150] *Id*. at 5, 37–38.
[151] *Id*. at 6–7, 39 ("the suggestion that the growth of LLM technology would come to a halt (or anything close) doesn't pass the straight face test."). *Id*. at 10 (Meta's revenue

B. Generative AI and the Structural Tensions of Fair Use

Despite a similar outcome on transformativeness under the first factor, the reasoning in *Kadrey* and *Bartz* reveals instability in how fair-use doctrine applies to AI training. The sections that follow identify several concrete difficulties. AI training unsettles the longstanding assumption that greater transformativeness corresponds to reduced market harm under factor four. In addition, market-dilution theories place pressure on the idea/expression dichotomy by treating stylistic similarity as economically relevant harm. Finally, the cases leave unresolved how courts should approach emerging licensing markets, how they should interpret the concept of "public benefit," and what the institutional limits of judicial decision-making should be in this context. Taken together, these difficulties suggest that AI training may require a recalibration of established fair-use principles in light of copyright's constitutional goals, rather than their mechanical application.

1. The Breakdown between Factors One and Four

AI training resists easy categorization within the established functional versus expressive framework.[152] Some stakeholders and commentators argue that the training of AI models on copyrighted works is analogous to established fair uses such as indexing, plagiarism detection, and reverse engineering, and constitutes a functional, non-expressive use that extracts statistical associations rather than expressive content.[153] Others challenge simplistic applications of fair use to AI training and scrutinize the claim that such uses are merely "functional" rather than "expressive." For example, Daniel J. Gervais distinguishes AI training from search and thumbnail cases by underscoring that the initial ingestion takes place exactly due to the work's expressive value, so the functional/expressive line does little work in this context.[154] This view aligns with Judge Chhabria's reasoning in *Kadrey v. Meta*, where he distinguished *Authors Guild v. Google* by observing that search technology would operate identically even if the indexed material consisted of gibberish, whereas the effectiveness and expressive capacities of a generative model are inextricably tied to the quality of its training materials.[155] In support of this position, a recent study by Chakrabarty, Ginsburg, and Dhillon shows that the

---

projections), and *id*. at 38 (rejecting the argument that recognizing a licensing market would impede progress).

[152] For discussion of this framework, *see supra* Section I B.

[153] *See, e.g.*, Lemley & Casey, *Fair Learning*, *supra* note 26, at 752–53; Sag, *Fairness and Fair Use in Generative AI*, *supra* note 58, at 1903; Amicus Brief of Professors Rebecca Tushnet et al., *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-03400-VC (N.D. Cal. Feb. 9, 2024), at 15–18 (arguing that AI training serves functions analogous to established non-expressive uses such as search and indexing). *See also* U.S. COPYRIGHT OFFICE, *Copyright and Artificial Intelligence, Part 3: Training*, *supra* note 21, at 41 (summarizing comments submitted by Samuelson et al., Anthropic, IBM, and BSA in response to the Office's 2023 notice of inquiry).

[154] Gervais, *The Heart of the Matter*, *supra* note 29 at 504–05. *See also* David W. Opderbeck, *Copyright in AI Training Data: A Human-Centered Approach*, 76 OKLA. L. REV. 951, 975–80 (2024).

[155] *See supra* notes 121–23 and accompanying text.

expressive value of copyrighted works is not merely incidental to the training of large language models but essential to their performance.[156] The study asked MFA-trained writers to produce short passages that imitated the styles of well-known authors.[157] The researchers then created matching AI passages under two conditions: one in which the model tried to imitate the author using only the prompt materials, and another in which the model had been fine-tuned on all of the author's books.[158] These human and AI passages were then assessed in blind evaluations by MFA-trained experts and lay readers.[159] When the AI relied only on prompts, expert readers strongly preferred the human-written passages, while lay readers showed mixed preferences.[160] Once the model had been fine-tuned on the author's complete works, both expert and lay readers preferred the AI passages for style and overall quality.[161] The findings confirm the value of books as training materials and suggest that even within large corpora, individual copyrighted works can materially influence a model's capabilities. On the other hand, as discussed below, an AI model may be prompted to produce something entirely different from a stylistically similar paragraph; for instance, it may be used to rephrase a passage in simpler language, provide a translation, or help a user practice conversational fluency in a foreign language.[162]

Another distinctive challenge in applying copyright law to AI-generated outputs is that the relationship between the training input and the resulting output can range from very close to highly remote and/or transformative. On one end of the spectrum, an AI system may produce material that is strikingly similar to a specific input, such as when an image generator reproduces a copyrighted character design with only superficial alterations or when a music model generates a passage nearly indistinguishable from a known song.[163] These situations raise obvious concerns of substitution and direct

---

[156] *See* Tuhin Chakrabarty, Jane C. Ginsburg & Paramveer Dhillon, *Readers Prefer Outputs of AI Trained on Copyrighted Books over Expert Human Writers* (Oct. 15, 2025), https://ssrn.com/abstract=5606570.

[157] *See* Chakrabarty et. al, *Readers Prefer Outputs of AI Trained on Copyrighted Books*, *supra* note 156, at 4.

[158] Chakrabarty et. al, *Readers Prefer Outputs of AI Trained on Copyrighted Books*, *supra* note 156, at 5–6.

[159] Chakrabarty et. al, *Readers Prefer Outputs of AI Trained on Copyrighted Books*, *supra* note 156, at 5, 18–19.

[160] Chakrabarty et. al, *Readers Prefer Outputs of AI Trained on Copyrighted Books*, *supra* note 156, at 1–2, 7–8 (experts preferred human writing; lay readers preferred AI on quality and showed no preference on style).

[161] Chakrabarty et. al, *Readers Prefer Outputs of AI Trained on Copyrighted Books*, *supra* note 156, at 7–8.

[162] *See* U.S. Copyright Office, *Copyright and AI, Part 3*, *supra* note 21, at 45 ("Many AI models, however, are meant to perform a variety of functions, some of which may be distinct from the purpose of the copyrighted works they are trained on. For example, a language model can be used to help learn a foreign language by chatting with users on diverse topics and offering corrective feedback.").

[163] For recent cases claiming these types of infringement*, see, e.g.*, *Getty Images*, No. 1:23-cv-00135 (alleging Stable Diffusion creates and displays images that are highly similar to and derivative of Getty Images' copyrighted works); *Disney Enters.*, No. 2:25-cv-05276 (alleging

infringement. On the other end of the spectrum, the output may diverge substantially from any identifiable source material, for instance when a text generator produces an essay on a novel topic, or when an image model synthesizes a wholly new composition blending countless inputs. This unpredictability, intrinsic to generative technology, is unlike any previous situations the courts have grappled with, where the outputs had predictable results. In *Authors Guild v. Google*, mass digitization produced fixed and functionally limited outputs—search snippets or accessibility copies—closely tied to the source texts.[164] Similarly, in *A.V. v. iParadigms*, plagiarism-detection software generated straightforward similarity reports, again reflecting high fidelity to the underlying material.[165] In both cases, when a user engages these technologies—whether searching for a term in a digitized book or submitting a paper to plagiarism-detection software—the output is determinate and repeatable: there is only one correct result, and it will always be the same. This variability in AI outputs sits uneasily with the Supreme Court's recent "use-by-use" approach in *Andy Warhol Foundation v. Goldsmith*, where the Court emphasized that the first fair use factor must be assessed with reference to "the specific use alleged to be infringing," rather than to the broader transformative potential of the defendant's practice.[166] But how can this be evaluated when a plethora of "uses" is possible?

The tension is amplified in cases where outputs fall in a gray area—bearing stylistic resemblance to a corpus (for instance, "in the style of Van Gogh" paintings or jazz improvisations modeled on Miles Davis) without copying any particular work. This raises the threat of market dilution: the concern that large volumes of stylistically similar outputs will reduce the value and quash demand for the originals.[167] As Judge Chhabria observed, this threat does not affect all authors equally. A reader seeking a book by Sarah Silverman or Barack Obama is unlikely to accept a machine-generated imitation because the author's identity carries substantial market value; the same may not hold for less individualized writing markets, such as magazines or genre fiction, where author-specific demand is weaker.[168] In such markets, a surge of AI-generated substitutes may not only divert immediate readership but also impede emerging authors' ability to develop their skills and distinguish themselves in an already crowded field.[169]

Midjourney generated infringing images of Disney and Universal characters); *Universal Music Publ'g Grp. v. Anthropic*, No. 3:23-cv-01092 (claiming infringement of copyrighted song lyrics by Anthropic's Claude model).

[164] *Authors Guild v. Google*, 804 F.3d 202. *See also Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) ((holding that digitizing books to enable full-text search and provide accessible formats for print-disabled users was a transformative, non-expressive fair use).

[165] *A.V. ex rel. Vanderhye v. iParadigms*, 562 F.3d 630.

[166] *See Warhol*, 598 U.S. at 523–25. *See also supra* notes 47−49 and accompanying text.

[167] *See supra* notes 132−133 and accompanying text.

[168] *See Kadrey*, No. 3:23-cv-03400-VC, slip op. at 31 (giving the example of romance novels); Hearing Tr. at 8–21, *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-03417 (N.D. Cal. May 1, 2025) (giving the example of gun magazines), https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.590.0.pdf.

[169] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 29 ("But it's easy to imagine that AI-generated books could successfully crowd out lesser-known works or works by up-and-coming authors. While

It bears emphasis that factor four of the fair use analysis looks to "the effect of the use upon the potential market for or value of the copyrighted work," and not to the effect of the *copy*. [170] Once the existence of copying is established or conceded, the analysis turns on market consequences of the "use," therefore "indirect substitution" would not contravene a textual interpretation of the Copyright Act. Still, difficult questions arise: is it consistent with fair-use principles to extend protection into the domain of "style"? Such an extension risks unsettling the idea/expression dichotomy by effectively granting copyright owners exclusive rights over non-protectable stylistic elements.[171] Traditionally, the fourth factor has been concerned with market harm caused by substantially similar works that serve as substitutes for the original; collapsing that inquiry into a general prohibition on stylistic similarity would mark a significant doctrinal shift.[172] In *Copyright Dilution Under Constitutional Scrutiny*, Edward Lee argues that the market dilution theory raises two distinct constitutional concerns. It exceeds the limits of the Progress Clause by granting authors de facto exclusive rights over styles or genres rather than over specific expressive works, and it suppresses lawful expression by treating non-infringing works as harmful merely because they share stylistic attributes with training data, thereby imposing content-based speech restrictions.[173] However, AI training differs in a crucial respect from classical cases rejecting protection for style: [174] training a generative model necessarily begins with the creation of verbatim digital copies of the works in the training set.[175] That initial copying sets AI training apart from situations in which an artist merely absorbs and reflects another's style without reproducing underlying works.

Ultimately, unpredictable outputs and speculative harm result in the breakdown of the relationship between fair use factors one and four. Traditionally, courts have assumed that a finding of transformativeness under factor one reduces the risk of market substitution under factor four, on the grounds that the more a secondary use serves a new

---

AI-generated books probably wouldn't have much of an effect on the market for the works of Agatha Christie, they could very well prevent the next Agatha Christie from getting noticed or selling enough books to keep writing."). *See also* ETHAN MOLLICK, CO-INTELLIGENCE: LIVING AND WORKING WITH AI 189 (2024) (presenting research demonstrating that AI systems tend to elevate the output of average performers rather than supplant top-tier creators).

[170] 17 U.S.C. § 107(4).

[171] *See Baker v. Selden*, 101 U.S. 99, 103–04 (1879) (distinguishing unprotectable ideas and systems from protectable expression); *Satava v. Lowry*, 323 F.3d 805, 812–13 (9th Cir. 2003) (denying protection for artistic "style" in glass sculpture, and stressing that copyright does not extend to general aesthetic features).

[172] Lee, *Fair Use and the Origin of AI Training*, *supra* note 26, at 197.

[173] Lee, *Copyright Dilution Under Constitutional Scrutiny*, *supra* note 23, at 21–27.

[174] *See Nichols v. Universal Pictures Corp.*, 45 F.2d 119, 121–22 (2d Cir. 1930) (holding that copyright does not protect general themes, stock elements, or an author's narrative style); *Satava*, 323 F.3d at 812–13 (rejecting protection for an artist's general style and concluding that copyright does not extend to aesthetic style or ideas); *Rentmeester v. Nike, Inc.*, 883 F.3d 1111, 1123–24 (9th Cir. 2018) (distinguishing protectable expression from unprotectable stylistic choices and high-level artistic concepts).

[175] *See* U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 26 (explaining that AI training requires making intermediate copies of copyrighted works during ingestion, tokenization, and embedding, and that such copies implicate the reproduction right under § 106(1)).

purpose, the less it competes with the original.[176] Under a market dilution theory, however, even though the purpose of the secondary use may be transformative, its stylistic similarity to the original, combined with the scale at which outputs can be generated and disseminated, may still cause market displacement. This marks a notable departure from the alignment between the first and fourth factors that has historically underpinned the fair-use framework.[177]

Both Judge Alsup and Judge Chhabria devote surprisingly little attention to the question of licensing markets in *Bartz v. Anthropic*and *Kadrey v. Meta*, remarking that rightsholders are not entitled to claim a market for licenses merely because a defendant could have sought one, and warning that such an approach would "swallow fair use" by circularly presuming infringement whenever no license was obtained.[178] This reasoning risks collapsing factor four into factor one. Under established precedent, the inquiry is whether the relevant market is "traditional, reasonable, or likely to be developed," not merely a hypothetical licensing possibility.[179] Indeed, both opinions mention licensing opportunities in passing, but neither engages substantively with whether markets for licensing training data are in fact emerging or reasonably expected to exist.[180] By reasoning that the use is transformative and therefore not the kind that should be licensed in the first place, both courts arguably replicate a form of the very circularity they say they are trying to avoid. They adopt transformativeness as a proxy for fair use (something the Supreme Court warned against in *Warhol*) and overlook the requirement, grounded in precedent, to account for legitimately cognizable markets that bear on both the purpose-and-character analysis and the assessment of market harm.

In their factor four analysis, the courts in *Kadrey* and *Bartz* also departed on their treatment of pirated books, with Judge Alsup finding it against fair use and Judge Chhabria finding the source of the training data only marginally relevant on the grounds that the use was "highly transformative."[181] Even though there is some ambiguity, courts have generally held that the legality of a defendant's acquisition of copyrighted works bears on the fair-use analysis but is not dispositive; it becomes relevant chiefly where the

---

[176] *See supra* notes 42–43 and accompanying text.
[177] *See supra* notes 39–42 and accompanying text.
[178] *See supra* notes 127 and 137 and accompanying text.
[179] Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, supra note 57, at 1504, citing *Texaco*, 630 F.3d 913.
[180] *See Kadrey*, No. 3:23-cv-03400-VC, slip op. at 10–11 (explaining that Meta initially explored licensing books for training—including budgets reportedly reaching $100 million—but negotiations collapsed because publishers lacked the relevant subsidiary rights, rights were fragmented regionally, and some publishers possessed only pricing proposals; Meta ultimately abandoned licensing after determining that most of the same works were already available on LibGen); *id*. at 13 (noting that plaintiffs stated they would have licensed their books for training but were never approached, and none had ever licensed a book for LLM training); *Bartz*, No. C 24-05417-WHA, slip op. at 3–4 (describing Anthropic's brief effort to pursue large-scale book licensing after hiring former Google Books executive Tom Turvey, followed by the lapse of those discussions even as another tech company reached a licensing deal with a major publisher).
[181] *See supra* notes 112–13, 117–18, 128–30, 140–41 and accompanying text.

unlawful source is causally connected to market substitution.[182] *Bartz* and *Kadrey* differ in their definition of the relevant "use", which in turn impacts the role played by pirated books.[183] In *Bartz*, Judge Alsup applies *Warhol*'s use-by-use framework and treats library building as a discrete use separate from model training, noting that not all books were used for training and some were kept "forever" as part of a "research library" for various "contingent uses."[184] He compares this use to the infringing one in *Texaco*, where copies were made for research purposes outside the scope of the original license.[185] He distinguishes intermediate copying cases like *Sega Enterprises v. Accolade* because in those cases the originals had been properly purchased.[186] From this perspective, the inclusion of pirated books posed an immediate threat of market substitution, because library building was itself evaluated as an expressive, non-transformative purpose that competed with the purpose and the market for purchased editions.[187] An issue with this approach is that having described training as "spectacularly" and "quintessentially" transformative,[188] it seems that, paradoxically, the pirated books might have been more favorably situated under factor one had they all been used for training. In other words, Anthropic's position becomes worse, not better, because certain copies were *not* used. This approach seems to follow an intuition of opportunism by Anthropic without addressing the more difficult question of how unlawful acquisition should weigh in a fair-use analysis when the downstream use is highly transformative.

2. The Limits of the Human-Learning Analogy

In part, Judge Alsup's outcome on pirated books is the logical extension of the analogy between human learning and computer learning that is embraced in *Bartz* and rejected in *Kadrey*.[189] The analogy holds that just as a person reads books, cases, or

---

[182] *See Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 562–63 (1985) (holding that unauthorized acquisition is relevant to fair use but not independently dispositive and emphasizing its weight where the leak caused direct market harm). *Oracle*, 593 U.S. at 34 (expressing skepticism about whether bad faith is relevant in a fair use analysis); *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 478–79 (2d Cir. 2004) (explaining that bad faith may matter only where it "heightens the likelihood of market harm," and rejecting a per se rule against fair use for unlawfully obtained works). *See also Kadrey*, No. 3:23-cv-03400-VC, slip op. at 19–20 ("The law is in flux about whether bad faith is relevant to fair use"); U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 52 (acknowledging ambiguity on the matter but endorsing the view that "the knowing use of a dataset that consist of pirated or illegally accessed works should weigh against fair use without being determinative.").

[183] *Compare Bartz*, No. C 24-05417-WHA, slip op. at 19–20, *with Kadrey*, No. 3:23-cv-03400-VC, slip op. at 20.

[184] *See Bartz*, No. C 24-05417-WHA, slip op. at 21, 19–20. *See also Warhol*, 598 U.S. at 544–545.

[185] *See Bartz*, No. C 24-05417-WHA, slip op. at 20–21(citing *Texaco*, 60 F.3d 913).

[186] *See Bartz*, No. C 24-05417-WHA, slip op. at 22.

[187] *Id.* at 19, 29–30.

[188] *Id*. at 11, 13.

[189] In *Bartz*, Judge Alsup embraces the analogy between human learning and machine learning. Under factor one, he describes Claude as a system that "reads prompts and writes responses as if it were a human," and dismisses the plaintiffs' contention that computers "should not be allowed to

articles and later creates a new work without infringing the copyright of the source materials, LLMs produce new, non-infringing expression based on ingested and processed training data. According to this argument, the original materials were only used as a source to create something new; neither human nor machine outputs are exact copies of the original inputs, so no market substitution—and ultimately copyright infringement—has occurred. This analogy has been adopted by a range of sources. In defending AI training practices in the wake of the AI Act, President Donald J. Trump stated that "when a person reads a book or an article, you've gained great knowledge… that does not mean that you're violating copyright laws."[190] The CEO of OpenAI, Sam Altman, invoked the analogy during an interview ("I do think that in the same way humans can read the internet and learn, Al should be allowed to read the internet and learn.").[191] The analogy matters doctrinally because, if accepted, it shifts the fair-use inquiry away from the legally significant act of reproduction and toward an assessment of the outputs, thereby minimizing the role of input-stage copying that ordinarily triggers the author's exclusive rights. First, by reframing ingestion as the digital equivalent of human reading, it encourages courts to treat reproduction as a preparatory act rather than a copyright-relevant use in its own right. Second, by placing weight on the creative character of the model's outputs rather than on the copying inherent in training, the analogy downplays the possibility of market substitution and recasts the secondary work as the product of the type of follow-on creativity that copyright law typically permits and encourages. The secondary use, then, as in Judge Alsup's view, is a type of competition copyright should not concern itself with.[192] Finally, as a human creator cannot legitimize the use of unlawfully obtained source materials by producing original scholarship or art, neither can an LLM.

Courts have long recognized that analogies between the digital and physical realms often collapse under scrutiny. This doctrinal skepticism reflects the recognition that digital processes frequently involve reproduction even when the analogue physical act does not. The refusal to apply the first sale doctrine to digital goods offers a useful template for understanding why AI training should be treated differently from human learning. The unique properties of digital technologies, such as perfect reproducibility, non-rivalrous use, and the absence of natural scarcity, undermine traditional legal frameworks built around physical goods and spaces. The first sale doctrine, codified in 17

---

do what people do." *Bartz*, No. C 24-05417-WHA, slip op. at 13. Under factor four, he extends the analogy further, suggesting that concerns about market harm resemble the worry that "training schoolchildren to write well would result in an explosion of competing works." *Id.* at 28. Judge Chhabria explicitly rejects this framing in *Kadrey*, calling the analogy "inapt" when applied to LLM training. *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 3. Teaching a child to write, he explains, is "not remotely like" developing an AI system capable of generating "countless competing works" with only a fraction of the time, labor, and creativity required of human authors. *Id.*

[190] President Donald J. Trump, *Transcript: Donald Trump's Address at "Winning the AI Race" Event*, TechPolicy.Press (July 24, 2025), https://www.techpolicy.press/transcript-donald-trumps-address-at-winning-the-ai-race-event.

[191] Sam Altman interviewed by Kevin Roose on the *Hard Fork* (podcast), Nov. 24, 2023 (transcript available at https://podcasts.happyscribe.com/the-daily/hard-fork-an-interview-with-sam-altman.

[192] *Bartz*, No. C 24-05417-WHA, slip op. at 28–30 (treating stylistic similarity in LLM outputs as competition outside copyright's concern and emphasizing the absence of market substitution).

U.S.C. § 109(a), provides that "the owner of a particular copy or phonorecord lawfully made under this title" may sell or otherwise dispose of that copy without the copyright owner's permission.[193] As content increasingly shifted to digital formats, courts confronted whether the first sale doctrine applied to digital files. Before the issue was tackled head on by the Second Circuit in *Capitol Records v. ReDigi*, it came up "unofficially" in *UMG Recordings v. MP3.com*.[194] In that case, a service allowing users to stream digital versions of CDs they owned was held infringing.[195] Although defendant MP3.com had lawfully acquired the physical CDs, it created and stored unauthorized digital copies on its servers so users could access the songs from anywhere.[196] The court rejected MP3.com's argument that it was merely facilitating "space-shifting," emphasizing that liability arose from the digital copying the service undertook to make the CDs accessible.[197] In *Capitol Records v. ReDigi*, ReDigi operated an online marketplace that allowed users to resell "used" digital music files purchased from iTunes.[198] The Second Circuit held that the first sale doctrine does not apply to digital music files, because "it is simply impossible that the same 'material object' can be transferred over the Internet," and found infringement of the reproduction right under § 106(1).[199]

These cases illustrate a consistent judicial refusal to treat digital copying as the functional equivalent of physical-world use, even where the digital practice is designed to mimic well-established consumer behavior. AI training, much like digital resale, depends on the making of intermediate copies through processes of ingestion, tokenization, and embedding that squarely implicate the reproduction right.[200] Human reading, by contrast, does not reproduce expressive content in a fixed form. Courts have stressed that the non-degradable, intangible character of digital works raises market-substitution risks not

---

[193] 17 U.S.C. § 109(a). Rooted in *Bobbs-Merrill Co. v. Straus*, the first sale doctrine reflects a longstanding policy of balancing copyright's exclusivity with the free alienability of lawfully owned physical copies. *Bobbs-Merrill Co. v. Straus*, 210 U.S. 339, 350–51 (1908) (holding that a copyright owner may not restrict the resale price of a lawfully purchased book and establishing the common-law foundation of the first sale doctrine). The Supreme Court reaffirmed this principle in *Quality King Distributors, Inc. v. L'anza Research International, Inc.*, applying § 109(a) to goods that were lawfully made in the United States and then reimported. *Quality King Distribs., Inc. v. L'anza Rsch. Int'l, Inc.*, 523 U.S. 135, 140–41 (1998).

[194] *UMG Recordings, Inc. v. MP3.com, Inc.*, 92 F. Supp. 2d 349 (S.D.N.Y. 2000).

[195] *Id*.

[196] *Id*. at 350–51(explaining that MP3.com purchased thousands of CDs, digitized them, and stored copies on its servers in order to provide its service, and that these reproductions were unauthorized).

[197] *Id***.** at 352.

[198] *Capitol Records, LLC v. ReDigi Inc.*, 910 F.3d 649, 650, 656–57 (2d Cir. 2018).

[199] *Id*. at 655–57. The first sale doctrine only applies to the distribution right, not the reproduction right. 17 U.S.C. § 109(a).

[200] *See* U.S. Copyright Office, *Copyright and AI, Part 3*, *supra* note 21, at 26 (explaining that AI training requires making intermediate copies of copyrighted works during ingestion, tokenization, and embedding, and that such copies implicate the reproduction right under § 106(1)).

present for physical goods.[201] A secondhand physical book is worn and limited in circulation, while a digital file is perfectly replicable at speed and scale, and competes directly with the market for new copies.[202] Human learning is perishable: memories fade, recall is unreliable, and expressive reproduction is imperfect. AI models, however, can retain, reproduce, and even regurgitate copyrighted content with high fidelity, raising substitution risks that approximate the resale of pristine digital files rather than the fallible cognition of human beings. Finally, even though beyond the scope of this paper, most digital files are distributed through end-user license agreements that expressly deny ownership.[203] Likewise, the training of AI models often occurs on the basis of licensed or scraped materials, which are subject to explicit contractual restrictions.[204] Together, the digital-copying cases demonstrate that analogies to physical-world behavior tend to break down once perfect, durable reproduction is involved. Treating AI training as the equivalent of human reading therefore risks overlooking the legally operative act of copying and sits uneasily with courts' refusal to import physical analogies into digital environments where reproduction magnifies the possibility of market substitution.

3. Public Benefit, Market Failure, and Institutional Limits

As discussed above, the "public benefit" operates as an unofficial but recurring consideration that is deeply rooted in the Constitution's Progress Clause and copyright's utilitarian orientation.[205] Although *Bartz* largely bypasses this inquiry and *Kadrey* touches on it only briefly, there is reason to believe that considerations of public benefit will assume greater prominence in future AI cases.[206] AI training disputes present a more complex conflict between two competing public-benefit narratives than we have seen before. Rightsholders frame the public interest in terms of sustaining incentives for

---

[201] *See ReDigi*, 910 F.3d at 657–59 (noting that perfect digital copies "compete directly" with authorized copies).

[202] *See Vernor v. Autodesk, Inc.* 621 F.3d 1102, 1111–12 (9th Cir. 2010) (explaining that transfer of digital files differs from resale of physical goods because each transfer produces a new copy while the original remains intact).

[203] Aaron Perzanowski & Jason Schultz, THE END OF OWNERSHIP 57–62 (MIT Press 2016) (Case Legal Studies Research Paper No. 2020-24) (documenting the shift from ownership of copies to licensed access in digital markets).

[204] *See, e.g.*, *Reddit, Inc. v. Anthropic PBC*, No. CGC-25-6258921 (Cal. Super. Ct. S.F. Cnty. June 4, 2025), removed to *Reddit, Inc. v. Anthropic PBC*, No. 3:25-cv-05643 (N.D. Cal. filed July 3, 2025) (Reddit suing Anthropic, alleging that it illegally "scraped" millions of Reddit posts and comments–many of which were deleted by users–to train its AI chatbot Claude without obtaining permission or licensing the content). *See also Reddit Sues Anthropic, Alleges Unauthorized Use of Site's Data*, WALL STREET JOURNAL (June 4, 2025) (noting that Reddit "has reached formal agreements with OpenAI and Google to license Reddit's valuable human-user data.").

[205] *See supra* notes 87–95 and accompanying text.

[206] The public benefit is identified as a relevant consideration in the U.S. Copyright Office's report on AI training. U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 71. It also appears throughout the literature on fair use and new technologies. *See, e.g.*, Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, *supra* note 57, at 1492–93; Lee, *Fair Use and the Origin of AI Training*, *supra* note 26, at 150; Lee, *Technological Fair Use*, *supra* note 76, at 837–38. *See generally* Levendowski Tepski, *Fairer Public Benefit in Copyright Law, supra* note 88.

human-created expression, warning that generative outputs risk diluting the market for original works, potentially displacing labor markets.[207] AI developers and their supporters, by contrast, emphasize a public benefit grounded in technological innovation, competition, and the expanded expressive capacity that LLMs enable, which they argue would be undermined by licensing obligations for training data.[208] In earlier cases dealing with a conflict between dissemination and technological innovation—*Napster*, *Grokster*, *Kelly*, and *Google Books*—the market-substitution inquiry provided a workable dividing line, in part because the uses and outputs at issue were easier to characterize.[209] When substitution was evident, courts favored the public benefit of preserving incentives for creation; when it was remote, they prioritized innovation-centric benefits such as searchability and new informational tools. The outputs in those cases were relatively stable, making substitution predictable. In AI training cases, however, the substitution inquiry becomes substantially more difficult to apply. Because LLM outputs are variable, probabilistic, and capable of both highly transformative and highly imitative expression, it is far harder to determine ex ante whether any given output will operate as a substitute for the underlying works.[210] The causal pathways to market harm are diffuse and, as seen in *Kadrey*, evidence of substitution is difficult to assemble in a fast-moving technological environment.[211]

Beyond substitution and the threat of market dilution, AI training also raises a distinct form of market failure that current fair-use doctrine has not fully addressed. In the context of AI training, the concern most frequently invoked is that of prohibitively high transaction costs.[212] There is, however, another market failure relevant to AI training that has received little attention.[213] If AI training is treated as fair use, once LLMs have been

---

[207] *See supra* note 29 and accompanying text.
[208] *See supra* notes 27–28 and accompanying text.
[209] *See supra* note 104 and accompanying text.
[210] *See supra* Section II B 1.
[211] *See supra* note 136 and accompanying text. In *Kadrey*, plaintiffs submitted an expert report by economist Daniel F. Spulber (Exhibit 76) addressing potential substitution effects and market harm arising from LLM training on copyrighted books; the court ultimately found the evidence insufficient. Daniel F. Spulber, Expert Report, Ex. 76, 190–230, *Kadrey*, No. 3:23-cv-03417-VC.
[212] *See* U.S. COPYRIGHT OFFICE, *Copyright and AI, Part 3*, *supra* note 21, at 86–87 (reporting on comments received from representatives from technology interests arguing that the logistical and financial challenges to license training data would be insurmountable.); Lemley & Casey, *Fair Learning*, *supra* note 26, at 748–49 (observing that AI training sets "contain millions of different works with thousands of different owners," making it infeasible to "license all of the underlying photographs, videos, audio files, or texts for the new use"). *See also* Gordon, *Fair Use as Market Failure*, *supra* note 37, at 1628 (describing transaction costs as a type of market barrier that, when exceeding an anticipated benefits, prevent a transaction from occurring).
[213] In economic terms, a market failure exists when markets, left to private bargaining, cannot reach outcomes that maximize overall welfare, usually because of some structural feature of the markets that prevents willing buyers and sellers from transacting efficiently. This results in a loss of value to society because potential gains from trade cannot be realized. *See* ROBERT COOTER & THOMAS ULEN, LAW AND ECONOMICS 38–43 (6th ed. 2012) (identifying market failures as situations where bargaining cannot achieve efficient outcomes due to transaction costs, externalities, or information problems, resulting in foregone gains from trade).

trained on copyrighted works and have extracted value from them, many of the outputs they produce may lack the human authorship required for copyright protection and therefore fall into the public domain.[214] There may also be instances where an individual or a company takes such AI-generated material, claims authorship, and registers or markets it as their own, despite having contributed no original creative input to it. As a result, value derived from copyrighted inputs may not be recaptured through copyright at the output stage. This scenario raises the classic problems of free riding and, over time, public-goods underproduction, which copyright was set up to solve in the first place.[215] In a typical creative economy, the value generated from existing works circulates back into the system: successful works finance new ones; licensing revenues support authors, editors, and publishers; and each generation of copyrighted works provides incentives for the next.[216] AI training disrupts this feedback loop. When LLMs ingest copyrighted expression, they appropriate some of its creative value (structure, style, patterns, and richness) but the outputs produced by the model do not necessarily replenish the pool of copyrighted works or return compensation to the original creators. The value may instead be captured at the level of the model, monetized by the developer, and re-emerge as content that does not itself generate equivalent copyright-based incentives. Over time, this one-directional transfer may erode the very incentive structure copyright is meant to sustain. This dynamic also weakens a central argument made by AI developers: that training promotes follow-on creativity.[217] That claim rests on the assumption that new works produced with AI tools will themselves contribute to the creative ecosystem. This value-extraction problem is distinct from market-dilution theory. Market dilution concerns the risk that AI outputs may serve as substitutes for copyrighted works, diverting audience demand. The argument advanced here does not depend on substitution. Even absent displacement, AI training can extract creative value without

---

[214] *See* U.S. COPYRIGHT OFFICE, *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*, 88 Fed. Reg. 16190, 16191–92 (Mar. 16, 2023) (clarifying that copyright protects only human authorship and that purely machine-generated material is not eligible for copyright protection); *Thaler v. Perlmutter*, No. 23-5233, slip op. at __ (D.C. Cir. Mar. 18, 2025) (affirming that a work generated autonomously by a computer system is not copyrightable absent human creative control).

[215] Free riding occurs when a party reaps the benefits of a creative work without bearing the costs of producing it, thereby weakening the incentives necessary for continued production. Public-goods underproduction arises because creative works—non-rivalrous and non-excludable absent legal entitlements—tend to be undersupplied when creators cannot capture enough of their social value. *See* WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 18–19 (2003); Gordon, *Fair Use as Market Failure*, *supra* note 37, at 1610–12.

[216] *See* U.S. COPYRIGHT OFFICE, *Identifying the Economic Implications of AI for Copyright Policy*, *supra* note 30, at 1 (explaining that copyright's economic purpose is to enhance long-run social welfare by enabling consumption of creative works and facilitating scientific and cultural innovation, and noting that existing works generate both immediate welfare gains and dynamic gains by serving as inputs into future creative output, such that successful works help finance and incentivize the creation of new ones).

[217] *See* Anthropic PBC, *Comments to the U.S. Copyright Office* 5–6 (2023); OpenAI, *Comments to the U.S. Copyright Office* 1 (2023); Microsoft, *Comments to the U.S. Copyright Office* 1 (2023). All comments available at: https://www.regulations.gov/docket/COLC-2023-0006/comments.

replenishing the ecosystem, potentially producing a structural underproduction of new protected works rather than merely a redistribution of demand.

This account is not without limits. First, it presupposes the current state of copyright law under which purely AI-generated outputs are generally uncopyrightable—a rule that could evolve. Congress or the courts might eventually recognize a new category of protection for machine-assisted works or establish sui generis rules for AI-generated content or clarify the threshold of human creative control sufficient to support authorship, which could alter the incentive dynamics described above.[218] To the extent that AI-generated outputs incorporate meaningful human selection, arrangement, or modification, they may themselves qualify for copyright protection, thereby partially restoring the feedback loop described above. Second, even in traditional markets, the benefits of copyright protection do not always flow directly to creators; they are often captured by intermediaries such as large publishers, labels, or platforms.[219] Thus, it is not obvious that additional protection for training data would translate into meaningful gains for most authors. Third, because the market for generative AI remains both nascent and rapidly developing, it is difficult to assess with confidence any resulting contraction in the market for expressive works. Finally, some may argue that the creative ecosystem has always tolerated (and even benefited from) forms of unremunerated borrowing—quotation, influence, re-use—that do not feed value directly back to original creators.[220] The challenge, then, is determining whether AI training represents a difference in degree or a difference in kind.

When the public-benefits inquiry is extended into domains such as foreign-policy competitiveness, national security concerns (including the "AI arms race"), impacts on labor markets, and the long-term trajectory of foundational technologies, it becomes increasingly unclear how well-equipped courts are to conduct evaluations that reach so far beyond traditional copyright analysis.[221] Amanda Levendowski Tepski has shown that courts struggle to apply fair-use doctrine consistently in technologically complex cases, in part because judges often lack technological fluency and rely on analogies or heuristics that fail to capture the underlying technical realities.[222] Edward Lee makes an analogy to the *Bleistein* nondiscrimination principle, which cautions judges against evaluating the

---

[218] *See* U.S. COPYRIGHT OFFICE, *Copyright and Artificial Intelligence: A Report to Congress, Part 1* 19–27 (2024) (noting congressional authority to revise the authorship standard and discussing potential alternative regimes).

[219] *See generally* Carys J. Craig, *The AI–Copyright Trap*, 55 COLUM. J.L. & ARTS 481 (2022).

[220] For the argument that not all free riding on creative works is harmful and that some forms of unremunerated borrowing promote creativity, see Mark A. Lemley, *Property, Intellectual Property, and Free Riding*, 83 TEX. L. REV. 1031 (2005) (arguing that the law tolerates and even values many forms of free riding because they support cumulative innovation and expressive progress).

[221] For a summary of these threats, *see supra* notes 28–29 and accompanying text. *See also* Lee, *Fair Use and the Origin of AI Training*, *supra* note 26, at 221 (arguing that copyright law should not hinder the technological progress of the U.S. and its dominance in a global "AI arms race"); Qualcomm, *Comments to the U.S. Copyright Office* at 3 (warning that uncertainty in U.S. copyright law governing AI training may drive AI development to jurisdictions with clearer rules, threatening U.S. competitiveness in high-technology and software industries).

[222] *See* Levendowski Tepski, *Fairer Public Benefit in Copyright Law, supra* note 88 at 51–2.

merit or worth of expressive works, to argue that courts are structurally ill-positioned to make predictive judgments about the social value of emerging technologies because technological benefits are inherently unpredictable."[223] Lee further emphasizes that economic analysis of technological impacts is "inconclusive, if not indeterminate," and that courts lack competence to perform cost-benefit balancing when the values at stake are incommensurable, especially given the First Amendment dimensions of technologies facilitating expression.[224] Taken together, these critiques suggest that expecting courts to determine how AI training aligns with national innovation strategy, international competition, or long-term industrial capacity risks pushing fair-use doctrine beyond the institutional competencies of the judiciary.

CONCLUSION

The first judicial opinions addressing the use of copyrighted works for AI training—*Bartz* and *Kadrey*—illustrate how emerging technologies test the contours of established fair-use doctrine. Such challenges are not unprecedented, but generative AI, which builds exponentially rather than incrementally on prior technologies, distinguishes itself in scale, speed, and unpredictability. As the cases reveal, the traditional distinction between expressive and functional secondary uses becomes difficult to apply when a system is not content-agnostic and mines the underlying work for both expressive and functional features. Likewise, the doctrine of intermediate copying offers limited guidance when copying serves not merely as a bridge to unprotected functional elements but to outputs that may plausibly substitute for the original works in the market. Perhaps most striking is the breakdown between the first and fourth fair-use factors. Historically, factors one and four have operated on a sliding scale: the more transformative the secondary purpose, the less likely the use is to substitute for the original work. AI destabilizes this relationship: a use may be highly transformative, like repurposing books into statistical training material, but also lead, in the aggregate and at scale, to outputs that plausibly compete with the market for the underlying works. A "different purpose" no longer reliably signals a "different market." Generative AI also strains the Supreme Court's directive in *Warhol* to evaluate "the specific use alleged to be infringing." A single act of ingestion may produce an unlimited range of heterogeneous outputs, some harmless, some highly derivative, and some potentially substitutive, making it difficult to identify which "use" is the relevant object of analysis. At the same time, AI presses on the idea/expression dichotomy: although style is unprotectable, AI systems can reproduce stylistic signatures with unprecedented fidelity and scale. Because the fourth factor is the most important and "the problem of substitution is the *bête noire* of copyright law" courts

[223] Lee, *Technological Fair Use*, *supra* note 76, at 839–41 (explaining that courts cannot reliably predict or quantify the public benefits of new technologies and lack tools for balancing those benefits against copyright harms).
[224] Lee, *Technological Fair Use*, *supra* note 76, at 840–41 (arguing that Congress, not courts, is institutionally equipped to make predictive judgments about technological development).

may eventually need to revisit certain assumptions to preserve the incentive structures at the heart of copyright's constitutional goals.[225]

Licensing markets, nascent when *Bartz* and *Kadrey* were filed but more developed today, should play a more meaningful role in future fair-use analysis.[226] Courts should not dismiss them reflexively out of concern for circularity. While recognizing a licensing market merely because rightsholders demand one would improperly collapse factor four, treating emerging, functioning markets as irrelevant is equally inconsistent with fair-use doctrine. These markets are not dispositive, but they are undeniably relevant. The case law also shows the need for greater doctrinal clarity about what counts as the relevant "use" when evaluating the impact of unlawful acquisition.

The treatment of pirated books in *Bartz* illustrates the challenge of defining the relevant "use" in the AI-training context. When acquisition, storage, and training operate within the same technical workflow, applying *Warhol*'s use-by-use analysis risks producing results that turn on formal distinctions rather than on a meaningful assessment of copyright-relevant effects.

The analogy between machine learning and human learning should be abandoned. It is descriptively inaccurate and doctrinally misleading. Machines do not "read," forget, or synthesize ideas in any manner analogous to human cognition; they reproduce, transform, and retain expression through copying processes that directly implicate the exclusive rights granted by copyright law. Collapsing these distinctions obscures the reproduction at the heart of training and risks importing assumptions about human creativity into contexts where they do not belong, including with respect to claims that AI promotes follow-on creativity. Courts should scrutinize such assertions, particularly when the human author's contribution to the LLM's output is minimal or purely mechanical. This is not the type of creativity copyright seeks to incentivize. The value-extraction dynamic described above reinforces this point: when AI systems absorb expressive value from copyrighted works but generate outputs that fall into the public domain or are misappropriated by users who claim authorship despite minimal input, the creative ecosystem does not recover the value expended in producing the original works. Rather than fueling new protected works, the value dissipates. Courts should recognize that this undermines, rather than advances, copyright's incentive structure.

Considerations of the public benefit have long played a role in fair-use jurisprudence, particularly in technology cases, and since *Google v. Oracle* have been explicitly located within the fourth factor analysis. Courts should remain attentive to the original meaning of dissemination in the Progress Clause—namely, the spread of knowledge—while also recognizing that AI may undermine it by diminishing incentives for the production of new human expression. Public-benefit analysis must therefore account for both innovation-oriented gains and the potential erosion of the creative ecosystem that

---

[225] *See Warhol*, 598 U.S. at 536 (quoting Pierre N. Leval, *Toward a Fair Use Standard*, 103 Harv. L. Rev. 1105, 1111 (1990)).

[226] *See* Copyright Alliance, *AI Licensing Database* https://copyrightalliance.org/artificial-intelligence-copyright/licensing/copyright-owners/ (last visited Apr. 2, 2026) (listing licensing agreements between major rightsholders—such as Springer, Condé Nast, and the *Financial Times*—and AI developers).

copyright is designed to sustain. As more empirical evidence becomes available, courts should incorporate economic analysis into fair-use adjudication, as they have done in earlier generations of technology cases, while remaining mindful of the limits of judicial competence. Evaluating macro-level effects like national competitiveness, geopolitical dynamics of an "AI arms race," environmental costs, or long-term labor-market disruptions extends beyond the judiciary's institutional mandate. These questions belong to Congress. What courts can and should do is ground their decisions in the evidentiary record, established economic reasoning, and copyright's constitutional purpose, without venturing into speculative or policy-laden predictions that exceed their role.

Ultimately, the challenges posed by generative AI call for applying fair use with the sensitivity and flexibility that have always been central to the doctrine. As Judge Chhabria cautioned in *Kadrey*, courts must resist the temptation to "robotically" import analogies and reasoning from earlier technological cases without attending to the distinctive context of AI.[227] Fair use has endured precisely because it adapts to significant changes in technology, but its adaptability is not unbounded: it remains anchored in the Constitution's directive to promote the progress of knowledge and the creation of new expressive works.

[227] *Kadrey*, No. 3:23-cv-03400-VC, slip op. at 32.