ARTICLE

FAIR USE AND THE ORIGIN OF AI TRAINING

Edward Lee[*]

ABSTRACT

Whether the training of artificial intelligence (AI) models with copyrighted works is a fair use—or not—is the most important copyright issue U.S. courts will decide in numerous lawsuits filed against AI companies. This Article offers four contributions to this debate. First, the Article traces the origin and history of the practice of using large and diverse datasets, including with many copyrighted works, to train AI models. This practice originated with AI researchers at universities who discovered a key insight: *scaling*, or using larger and more diverse datasets, actually worked in developing *and improving* AI models. This seminal breakthrough, which took decades to figure out, propelled the advances in AI witnessed today. This Article recommends that courts in the copyright litigation against AI companies evaluate not only this history of AI training, but also whether this university-based AI training has a fair use purpose or not.

Second, this Article explains how the Supreme Court's teachings in the fair use cases *Warhol*, *Google*, and *Sony* apply to the copyright claims filed against AI. The history of AI training through scaling illuminates its purpose at both universities and companies. Using copyrighted works to train AI models serves a fair use or "transformative" purpose, namely, the *further* purpose to research, develop, create, and improve new AI—a technological

* Professor of Law. Santa Clara University School of Law. Many thanks to Taylor Dalton, Eric Goldman, Hal Krent, Mark Lemley, Brian Love, Kerry Macintosh, Tim McFarlin, Andrew Moshirnia, Tyler Ochoa, Lisa Ramsey, Zahr Said, David Schwartz, Nick Serafin, David Sloss, David Yosifon, and Peter Yu for comments on earlier drafts.

innovation with profound significance for the United States, including its economic standing in the world. And the degree of this transformativeness in developing AI is *increased*, not diminished, by deploying AI models to enable people to generate new, non-infringing works. But, under *Warhol*, the outputs of AI generators, once deployed to the public, must be analyzed on a use-by-use basis. Substantially similar outputs, including regurgitations, may be infringing.

Third, this Article critiques the Copyright Office's pre-publication report on AI training and its endorsement of a new, untested theory of market dilution under the fourth factor of fair use. The Copyright Office's view of market harm derived from new, *non-infringing* AI-generated works based simply on being in the same genre of works (e.g., romance novels or music) used to train the AI model is not only "uncharted territory" as the Copyright Office conceded, but also unconstitutional.

Finally, this Article explains how *technological* progress is equally as important as *creative* progress. The Progress Clause envisions the grant of "exclusive Right[s]" "[t]o promote the Progress of Science and useful Arts."[1] Both technological innovation and creative production redound to the public's benefit. Both serve national interest. Both should be weighed in the balance of fair use.

TABLE OF CONTENTS

1. U.S. CONST. art. I, § 8, cl. 8.

I. INTRODUCTION

"And now we face the latest technological frontier: artificial intelligence (AI). At its core, AI combines algorithms and enormous data sets to solve problems."

—*Chief Justice John Roberts*[2]

"In deep learning, there's no data like more data. The more examples of a given phenomenon a network is exposed to, the more accurately it can pick out patterns and identify things in the real world."

—*Kai-Fu Lee*[3]

The United States is ranked as the world leader in artificial intelligence (AI) in terms of innovation, investment, and implementation.[4] In 2024, it scored the highest score (100) on the Global AI Index, with a sizeable lead over the next closest rival, China (54).[5] Despite their disagreements, the Trump and Biden Administrations agreed on one thing: the United States is the world leader in AI and should aim to expand its lead.[6]

In 2023, President Biden's AI executive order instructed: America must "seize AI's promise and deepen the U.S. lead in AI innovation."[7] In 2025, President Trump echoed the same goal: "The United States must act decisively to retain leadership in AI and enhance our economic and national security."[8] President Trump's AI executive order declared: "It is the policy of the United States to

---

2. JOHN G. ROBERTS, JR., 2023 YEAR-END REPORT ON THE FEDERAL JUDICIARY 5 (2023), https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf [https://perma.cc/9AX3-AUCW].

3. KAI-FU LEE, AI SUPERPOWERS: CHINA, SILICON VALLEY, AND THE NEW WORLD ORDER 14 (First Mariner Books 2021) (2018).

4. *The Global AI Index*, TORTOISE (Sep. 19, 2024), https://www.tortoisemedia.com/intelligence/global-ai/#rankings [https://perma.cc/A5KW-WA4N].

5. *Id.*

6. *Artificial Intelligence for the American People*, TRUMP WHITE HOUSE ARCHIVES, https://trumpwhitehouse.archives.gov/ai/ [https://perma.cc/XLN9-JXZB] (last visited July 19, 2025); *Fact Sheet: Biden-Harris Administration Announces New AI Actions and Receives Additional Major Voluntary Commitment on AI*, BIDEN WHITE HOUSE ARCHIVES (July 26, 2024) [hereinafter *Biden AI Announcement*], https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2024/07/26/fact-sheet-biden-harris-administration-announces-new-ai-actions-and-receives-additional-major-voluntary-commitment-on-ai/ [https://perma.cc/KE5E-EQ2T].

7. *Biden AI Announcement*, *supra* note 6.

8. *Fact Sheet: President Donald J. Trump Takes Action to Enhance America's AI Leadership*, WHITE HOUSE (Jan. 23, 2025), https://www.whitehouse.gov/fact-sheets/2025/01/fact-sheet-president-donald-j-trump-takes-action-to-enhance-americas-ai-leadership/ [https://perma.cc/G7PK-V6X5].

sustain and enhance America's global AI dominance . . . to promote human flourishing, economic competitiveness, and national security."[9] AI is so central to the United States's interest that President Trump issued another executive order to prioritize AI education to students in kindergarten through twelfth grade, to "provide our Nation's youth with opportunities to cultivate the skills and understanding necessary to use and create the next generation of AI technology."[10]

The United States's lead in AI is, by no means, secure. China's government has a bold national plan to overtake the United States in AI by 2030.[11] China has already eclipsed the United States in terms of sheer volume of AI research.[12] China's government directly funds some Chinese AI companies and offers subsidies for companies that buy AI chips produced in China.[13] China's approach contrasts with the United States's approach to AI development, which relies heavily on private investment and venture capital.[14] According to the Information Technology & Innovation Foundation, "China is advancing rapidly in AI research and application, challenging the United States' dominance in this critical field."[15] Now, "[t]he question is not whether the United States can contain China in AI, but whether it can keep ahead."[16]

By late 2024, the answer was in doubt. In a watershed moment, "a small Chinese start-up called DeepSeek unveiled a new A.I. system that could match the capabilities of cutting-edge

9. Exec. Order No. 14,179, 90 Fed. Reg. 8741 (Jan. 23, 2025).

10. Exec. Order No. 14,277, 90 Fed. Reg. 17519 (Apr. 23, 2025).

11. *See* Kyle Chan et al., *China's Evolving Industrial Policy for AI*, RAND (June 26, 2025), https://www.rand.org/pubs/perspectives/PEA4012-1.html [https://perma.cc/4MVC-2A4V]; Mercy A. Kuo, *China's Bid to Lead the World in AI*, THE DIPLOMAT (July 6, 2024), https://thediplomat.com/2024/07/chinas-bid-to-lead-the-world-in-ai/ [https://perma.cc/G4ZN-YHYL]; Pablo Robles, *China Plans to Be a World Leader in Artificial Intelligence by 2030*, S. CHINA MORNING POST (Oct. 1, 2018), https://multimedia.scmp.com/news/china/article/2166148/china-2025-artificial-intelligence/index.html [https://perma.cc/68BT-AN7Q]; Daniel Araya, *Who Will Lead in the Age of Artificial Intelligence?*, BROOKINGS (Feb. 26, 2019), https://www.brookings.edu/articles/who-will-lead-in-the-age-of-artificial-intelligence/ [https://perma.cc/RHP6-X9YF].

12. Hodan Omaar, *How Innovative Is China in AI?*, ITIF (Aug. 26, 2024), https://itif.org/publications/2024/08/26/how-innovative-is-china-in-ai/ [https://perma.cc/S2BZ-LF7A].

13. *Id.*

14. *Id.*; *see* Nathan Bomey, *Charted: U.S. Is the Private Sector AI Leader*, AXIOS (July 9, 2024), https://www.axios.com/2024/07/09/us-ai-global-leader-private-sector [https://perma.cc/K9WT-Y7Q3]; Rana Foroohar, *Early Adoption of AI Will Boost US Growth*, FIN. TIMES (June 1, 2025) (U.S. private funding of AI was $109 billion, "nearly 12 times China's $9.3bn"), https://www.ft.com/content/339a7e8c-d7ba-499c-b02d-40a514d6bd8a [https://perma.cc/Z3NH-FV6E].

15. Omaar, *supra* note 12.

16. *Id.*; *see also* Austin Carr et al., *Is China Winning?*, BLOOMBERG BUSINESSWEEK, June 2025, at 40, 42 ("Wei Sun, an analyst at Counterpoint Technology Market Research, says the AI gap between the US and China is now measured in months, not years.").

chatbots from companies like OpenAI and Google."[17] DeepSeek did so, it said, using "only a fraction of the highly specialized computer chips that leading A.I. companies [in the United States] relied on to train their systems" due to the U.S. government's export restrictions on chips.[18] DeepSeek reportedly innovated how to train its models by "using a battery of engineering tricks—custom communication schemes between chips, reducing the size of fields to save memory, and innovative use of the mix-of-models approach."[19] Some skeptics, like Elon Musk, who has his own AI company, doubted DeepSeek did what it said on such little computing power or money.[20] But other prominent U.S. tech leaders see China starting to pull ahead.[21] Independent performance reviews even rated DeepSeek R1 as one of the best nonreasoning models.[22]

DeepSeek sent shock waves through the United States. The stock market lost $1 trillion in value in one day—with AI chip maker NVIDIA dropping $589 billion, the largest single-day drop

17. Cade Metz & Meaghan Tobin, *How Chinese A.I. Start-Up DeepSeek Is Competing with Silicon Valley Giants*, N.Y. TIMES (Jan. 27, 2025), https://www.nytimes.com/2025/01/23/technology/deepseek-china-ai-chips.html [https://perma.cc/48DB-EWUH].

18. *Id.*; *see also* Zeyi Yang, *How Chinese AI Startup DeepSeek Made a Model that Rivals OpenAI*, WIRED (Jan. 25, 2025, at 05:00 CT), https://www.wired.com/story/deepseek-china-model-ai/ [https://perma.cc/TV3K-UCL7].

19. Yang, *supra* note 18 (quoting Wendy Chang, software engineer). DeepSeek's methods drew scrutiny. OpenAI suspected that "DeepSeek extricated large volumes of data from OpenAI's tools to help develop its technology, using a process called distillation." Sam Schechner, *OpenAI Is Probing Whether DeepSeek Used Its Models to Train New Chatbot*, WALL ST. J. (Jan. 29, 2025, at 12:49 ET), https://www.wsj.com/tech/ai/openai-china-deepseek-chatgpt-probe-ce6b864e [https://perma.cc/92J9-2HJ6]. OpenAI CEO Sam Altman suggested that its proprietary approach may be "on the wrong side of history" and the company may "need to figure out a different open source strategy" to its AI models. *See* Deepa Seetharaman, *Sam Altman's Answer to DeepSeek Is Giving Away OpenAI's Tech*, WALL ST. J. (Jan. 31, 2025, 19:50 ET), https://www.wsj.com/tech/ai/sam-altmans-answer-to-deepseek-is-giving-away-openais-tech-d1a5a9ec?st=iKqbDG&reflink=desktopwebshare_permalink [https://perma.cc/X9HS-XTVL].

20. Max Zahn, *Elon Musk Launches His Own AI Company to Compete with ChatGPT,* ABC NEWS (July 13, 2023, at 10:58 CT), https://abcnews.go.com/Business/elon-musk-launches-ai-company-compete-chatgpt/story?id=101210078 [https://perma.cc/W22L-LVPA]; *see* Ben Sherry, *AI Leaders in the U.S. React to DeepSeek, Calling It 'Impressive' but Staying Skeptical*, INC. (Jan. 28, 2025), https://www.inc.com/ben-sherry/ai-leaders-in-the-u-s-react-to-deepseek-calling-it-impressive-but-staying-skeptical/91140125 [https://perma.cc/XM8A-WALS].

21. *See* Eric Schmidt & Selina Xu, *DeepSeek. Temu. TikTok. China Tech Is Starting to Pull Ahead.*, N.Y. TIMES (May 5, 2025), https://www.nytimes.com/2025/05/05/opinion/china-ai-deepseek-tiktok.html [https://perma.cc/6LMD-7FFJ].

22. *Independent Analysis of AI*, ARTIFICIAL ANALYSIS, https://artificialanalysis.ai/ [https://perma.cc/X572-N4S8] (last visited July 14, 2025); *see also* Liza Lin, Josh Chin & Raffaele Huang, *China Is Quickly Eroding America's Lead in the Global AI Race*, WALL ST. J. (July 1, 2025, at 23:00 ET), https://www.wsj.com/tech/ai/artificial-intelligence-us-vs-china-03372176?st=sis7bP&reflink=desktopwebshare_permalink [https://perma.cc/R6HC-BXAK] (showing global adoption of AI models from China has grown, providing stiff competition with AI models from United States).

for any company in U.S. history.[23] Microsoft CEO Satya Nadella called DeepSeek's new model "super impressive," and warned that "[w]e should take the developments out of China very, very seriously."[24] President Trump said DeepSeek's innovation was a "wake-up call for [U.S.] industries."[25] The 2025 Stanford AI Index report signaled the same alarm: "While the U.S. maintains its lead in quantity [of top AI models], Chinese models have rapidly closed the quality gap," and "China continues to lead in AI publications and patents."[26] The 2023 report of the Australian Strategic Policy Institute (ASPI) had even ranked China ahead of the United States in AI algorithms, as well as 37 of 44 other technologies.[27] ASPI concluded: "China has built the foundations to position itself as the world's leading science and technology superpower, by establishing a sometimes stunning lead in high-impact research across the majority of critical and emerging technology domains."[28] The fear that China will overtake the United States in AI reached Congress in 2025.[29]

The global AI arms race is on. Against that backdrop, the stakes in the copyright lawsuits against AI companies in the United States could not be higher. The copyright issue that confronts leading AI companies in the United States—from the Big Tech companies Google, Meta, Microsoft, NVIDIA, to the startup companies OpenAI, Anthropic, Stability AI, Midjourney, Perplexity, Cohere, Suno, and others (collectively, AI companies)—is whether their respective AI

---

23. *The $1 Trillion Shock: How China's DeepSeek Shook the Foundations of US Tech*, TIMES INDIA (Jan. 28, 2025, 22:35 IST), https://timesofindia.indiatimes.com/business/international-business/the-1-trillion-shock-how-chinas-deepseek-shook-the-foundations-of-us-tech/articleshow/117629763.cms [https://perma.cc/FT4S-VG34].

24. *DeepSeek Moment: AI Arms Race Disrupted by Potential Game-Changer from China*, CHAT GPT IS EATING THE WORLD (Jan. 27, 2025), https://chatgptiseatingtheworld.com/2025/01/27/deepseek-moment-ai-arms-race-disrupted-by-potential-game-changer-from-china/ [https://perma.cc/9C2Y-48AA].

25. Video posted by Acyn (@Acyn), X, Trump on DeepSeek: I view that as a positive. If it is fact and it is true, and nobody knows, but I view that as a positive. (Jan. 27, 2025, at 17:31 CT), https://x.com/acyn/status/1884021523016737021 [https://perma.cc/R83C-4696] (on file with the Houston Law Review).

26. STANFORD UNIV. HUM.-CENTERED A.I., ARTIFICIAL INTELLIGENCE INDEX REPORT 2025 3 (2025).

27. JAMIE GAIDA ET AL., AUSTRALIAN STRATEGIC POL'Y INST., ASPI'S CRITICAL TECHNOLOGY TRACKER: THE GLOBAL RACE FOR FUTURE POWER 1, 21 (2023), https://ad-aspi.s3.ap-southeast-2.amazonaws.com/2023-03/ASPIs%20Critical%20Technology%20Tracker_0.pdf?VersionId=ndm5v4DRMfpLvu.x69Bi_VUdMVLp07jw [https://perma.cc/FLX7-ZFAR].

28. *Id.* at 1.

29. *Video Sen. Ted Cruz Presiding over Hearing with Sam Altman. Stresses Importance of US Maintaining Lead over China in AI.*, CHAT GPT IS EATING THE WORLD (May 10, 2025), https://chatgptiseatingtheworld.com/2025/05/10/video-sen-ted-cruz-presiding-over-hearing-with-sam-altman-stresses-importance-of-us-mainting-lead-over-china-in-ai/ [https://perma.cc/3PKG-ZEP3].

models they developed, trained, and offered for public use are unlawful under U.S. copyright law. By mid-September 2025, fifty copyright lawsuits, which include many proposed class actions of potentially millions of class members, have been filed against these AI companies.[30] More lawsuits are likely. Other major U.S. companies offering generative AI, such as Musk's xAI, may face similar copyright lawsuits, especially if any of the current lawsuits are successful.[31]

If the AI companies are found liable for copyright infringement in the training of their AI models with copyrighted works, the potential remedies will be substantial. In the *New York Times*'s lawsuit against OpenAI and Microsoft, the statutory damages the *New York Times* can recover ranges from $7.5 billion to $300 billion in the general statutory range, but can increase up to *$1.5 trillion* if willful infringement by OpenAI is found as alleged by the *New York Times*.[32] To put OpenAI's potential liability in this one lawsuit into perspective, as successful as OpenAI is, it was valued in September 2025 at $500 billion—which suggests that it could go bankrupt if willful infringement is found.[33]

That's not all. The *New York Times* seeks complete "destruction . . . of *all GPT or other LLM models* and *training sets*

30. *See Updated Map of US Copyright Suits v. AI (Sept. 11 2025) Cases Hit 50.*, CHAT GPT IS EATING THE WORLD (Sep. 11, 2025), https://chatgptiseatingtheworld.com/2025/09/11/updated-map-of-us-copyright-suits-v-ai-sept-11-2025-cases-hit-51/ [https://perma.cc/C7X4-72VP]; *see also Master List of Lawsuits v. AI, ChatGPT, OpenAI, Microsoft, Meta, Midjourney & Other AI Cos.*, CHAT GPT IS EATING THE WORLD (Sep. 14, 2025) [hereinafter *Master List*], https://chatgptiseatingtheworld.com/2024/08/27/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/ [https://perma.cc/X5GD-RGNN] (listing U.S. copyright lawsuits).

31. *See* Zahn, *supra* note 20.

32. First Amended Complaint & Jury Trial Demand, at 6, 60, New York Times Co. v. Microsoft Corp., 777 F. Supp. 3d 283 (S.D.N.Y. 2025) (No. 1:23-cv-11195-SHS) [hereinafter New York Times' Complaint] (alleging ownership of 10 million registered works and defendants' infringement "[b]y building training datasets containing millions of copies of Times Works"). Please note: this estimate does not include a reduction in the total amount by treating the articles in one daily edition as a compilation, as should be the correct computation. *See* 17 U.S.C. § 504(c)(1)–(2). I am not privy to the precise number of compilations involved in the lawsuit.

33. *See* Berber Jin & Deepa Seetharaman, *OpenAI in Talks for Huge Investment Round Valuing It at Up to $300 Billion*, WALL ST. J. (Jan. 30, 2025, at 17:31 ET), https://www.wsj.com/tech/ai/openaiin-talks-for-huge-investment-round-valuing-it-up-to-300-billion-2a2d4327?reflink=desktopwebshare_permalink [https://perma.cc/KM8T-KFHT]; *cf.* Meghan Collins, *Napster Files for Bankruptcy*, CNN MONEY (June 3, 2002, at 13:25 ET), https://web.archive.org/web/20020614225631/http://money.cnn.com/2002/06/03/news/companies/napster_bankrupt/ [https://perma.cc/8EQ3-S46N] (describing how online music service Napster filed for bankruptcy after having been found to violate copyright law); Alex Fitzpatrick, *Aereo Just Disappeared for Good*, TIME (June 28, 2014, at 10:35 ET), https://time.com/2936238/aereo-off-supreme-court/ [https://perma.cc/7BXK-CNEB] (describing how streaming television service Aereo "decided to shut down" after the Supreme Court found its business practices violated copyright law).

that incorporate Times Works."[34] If granted, the remedy may mean that OpenAI would have to start all over. And the *Times*'s lawsuit against OpenAI is just the tip of the copyright iceberg. As of September 14, 2025, eighteen other copyright lawsuits have been filed against it around the world, most of them in the United States and most seeking class action certification.[35] And OpenAI is just one of many AI companies in the United States facing copyright lawsuits.[36]

The central legal question in these lawsuits is whether the respective defendant AI company's training of its AI model using the respective plaintiff's copyrighted works is fair use—or not. This novel legal question has sparked sharp, even bitter, disagreements among various industries, stakeholders, authors, legal scholars, commentators, and the public.[37] One side sees AI training as blatant theft, or outright "stealing" from artists and authors.[38] The other side sees AI training as a legitimate fair use to train AI models to develop new technologies with substantial non-infringing uses that can benefit society.[39] Both sides offer

34. New York Times' Complaint, *supra* note 32, at 68 (emphasis added).

35. *See Master List*, *supra* note 30; *Updated World Map of Copyright Suits v. AI (Sept. 11 2025) 71 Total*, CHAT GPT IS EATING THE WORLD (Sep. 11, 2025), https://chatgptiseatingtheworld.com/2025/09/11/updated-world-map-of-copyright-suits-v-ai-sept-11-2025-72-total/ [https://perma.cc/ET9H-XADU] (showing four copyright suits against OpenAI outside U.S.).

36. *Master List*, *supra* note 30.

37. *Compare Publishers Join with Worldwide Coalition to Condemn the Theft of Creative and Intellectual Authorship by Tech Companies for Generative AI Training*, ASS'N AM. PUBLISHERS (Oct. 22, 2024), https://publishers.org/news/publishers-join-with-worldwide-coalition-to-condemn-the-theft-of-creative-and-intellectual-authorship-by-tech-companies-for-generative-ai-training/ [https://perma.cc/W425-Z2PM] ("condemn[ing] the theft of creative and intellectual authorship" in AI training), *with* JOSHUA LEVINE & TIM HWANG, FOUND. FOR AM. INNOVATION, COPYRIGHT, AI, AND GREAT POWER COMPETITION 4, 7 (2025), https://cdn.sanity.io/files/d8lrla4f/staging/1e8b069431e03d9e8b23f03dede131ef5fdb16a9.pdf [https://perma.cc/HN7F-RNAR] (arguing fair use decisions in technology cases support fair use in AI training).

38. *See, e.g.*, Kate Mosse, *AI's Assault on Our Intellectual Property Must Be Stopped*, FIN. TIMES (Dec. 20, 2024), https://www.ft.com/content/f26bda64-1237-4188-8540-210367567089 [https://perma.cc/PU3D-D52B]; Mary Rasenberger, *As AI Is Embraced, What Happens to the Artists Whose Work Was Stolen to Build It?*, L.A. TIMES (June 18, 2024, at 03:02 PT), https://www.latimes.com/opinion/story/2024-06-18/artificial-intelligence-openai-media-manager-apple [https://perma.cc/4EM6-3EZX]; Kyle Chayka, *Is A.I. Art Stealing from Artists?*, NEW YORKER (Feb. 10, 2023), https://www.newyorker.com/culture/infinite-scroll/is-ai-art-stealing-from-artists [https://perma.cc/ZL6H-G7AU].

39. *See, e.g.*, Computer & Communications Industry Association, Comment Letter on Artificial Intelligence and Copyright, 2, 7, 9 (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-8740 [https://perma.cc/776C-LBM7] ("[U]ses of copyrighted materials as data in the creation and deployment of AI machine learning systems are fair uses."); Pamela Samuelson, Christopher Jon Sprigman, & Matthew Sag, Comment Letter on Artificial Intelligence and Copyright, 14 (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-8854 [https://perma.cc/LYY9-VEZD] ("Deriving uncopyrightable abstractions and associations from the training data and then using that knowledge to confect new digital artifacts is not just transformative, it is highly transformative," favoring fair use).

vastly different interpretations of the very same fair use cases of the Supreme Court and lower courts they say support their views. This novel legal question will soon be decided by the courts or juries in some of the fifty copyright lawsuits filed against AI companies. To complicate matters, in May 2025, after conducting a study, the Copyright Office issued a pre-publication version of its report elaborating its views on numerous issues related to fair use in AI training, but the status of that report remains uncertain given President Trump's firing of the Register of Copyrights Shira Perlmutter shortly after the pre-publication version's release.[40]

This Article crystallizes the important fair use principles for courts to consider in the AI litigation. The Article uses, as guideposts, the Supreme Court's two most recent cases on fair use, *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith* (*Warhol*)[41] and *Google LLC v. Oracle America, Inc.* (*Google*).[42] And it provides additional guidance drawn from *Sony Corp. of America v. University City Studios, Inc.* (*Sony*), an early but important fair use case involving a new technology in which the Court set forth the *Sony* safe harbor for technologies capable of substantial non-infringing uses.[43] Both doctrines, fair use and the *Sony* safe harbor, keep copyright from stifling technological innovation in the United States.

This Article makes four contributions. First, the Article recounts the history of AI training with datasets by researchers at universities to understand how and why this practice originated. The practice—what some call, invoking religious terms, AI's "original sin"—began in university research and later migrated to companies after the AI research showed promise.[44] The history of

40. U.S. COPYRIGHT OFF., COPYRIGHT AND ARTIFICIAL INTELLIGENCE PART 3: GENERATIVE AI TRAINING 32–35 (2025) [hereinafter PRE-PUBLICATION REPORT], https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf [https://perma.cc/VT89-JTTQ]; *see Pres. Trump Fires Register of Copyrights Shira Perlmutter Less than 24 Hours After She Issued Pre-Publication Report on AI Training & Fair Use More Favorable to Copyright Holders. Raises a Cloud of Uncertainty Over Report.*, CHAT GPT IS EATING THE WORLD (May 11, 2025), https://chatgptiseatingtheworld.com/2025/05/11/pres-trump-fires-register-of-copyright-shira-perlmutter-less-than-24-hours-after-she-issued-pre-publication-report-on-ai-training-fair-use-more-favorable-to-copyright-holders-raises-a-cloud-of/ [https://perma.cc/DHH4-G3N7].

41. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1266 (2023).

42. Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1190 (2021).

43. Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 442 (1984).

44. *See* The Daily, *A.I.'s Original Sin*, N.Y. TIMES (Apr. 16, 2024), https://www.nytimes.com/2024/04/16/podcasts/the-daily/ai-data.html [https://perma.cc/B658-QPNX]; JOHN WARNER, MORE THAN WORDS: HOW TO THINK ABOUT WRITING IN THE AGE OF AI 30 (2025) ("It's undeniable that generative AI has been born in sin and that it is already an ethical, moral, and environmental nightmare."); Pat Lawlor & Jerry Chang, *History of AI: How Generative AI Grew from Early Research*,

AI training in university research illuminates the important *technological* reason datasets consisting of copyrighted works were used: the use of larger and more diverse datasets propelled AI advances.[45] There was a long-felt but unmet need of researchers to figure out a way to develop AI dating back to the 1950s. For decades, it was unclear, if not doubtful, that AI research would lead to anything.[46] Even today, AI is still (fast) developing—as DeepSeek shows.[47] Indeed, it would be idiotic to think today's AI is the end-state for AI.

This history should inform the courts' resolution of fair use in the copyright lawsuits against AI companies. Courts should evaluate whether the use of copyrighted works in AI training at universities serves a fair use purpose—or not—by examining whether AI training serves a different or transformative purpose. For, if the use of copyrighted materials by university researchers to develop AI models is copyright infringement and not fair use, then, *a fortiori*, the fair use defense of AI companies, commercial entities, must fail. Conversely, if the courts find that such university-based AI training has a legitimate fair use purpose, then courts should reject broad arguments that use of copyrighted works in AI training by companies cannot serve a fair use purpose—e.g., because, as opponents argue, it purportedly is "not transformative" at all.[48]

Second, the Article shows how courts should follow *Warhol*'s teaching that courts must analyze fair use, *use-by-use*.[49] But *Warhol* must be read in conjunction with the Supreme Court's

---

QUALCOMM (Aug. 22, 2023), https://www.qualcomm.com/news/onq/2023/08/history-of-ai-how-generative-ai-grew-from-early-research [https://perma.cc/8ERC-Z9R8]; *see also* Karlo Broussard, *Using the Bible to Explain Original Sin*, CATH. ANSWERS (Jan. 10, 2023), https://www.catholic.com/magazine/online-edition/using-the-bible-to-explain-original-sin [https://perma.cc/CBY5-GET2] (explaining religious meaning of original sin).

45. *See infra notes* 264–86, 363–81 and accompanying text.

46. I explain the history of AI development and eventual recognition of scaling with larger datasets by researchers in Part IV, sections A and B. For a brief summary, *see* Lawlor & Chang, *supra* note 44.

47. *Cf.* Paul Sawers, *Meta's Yann LeCun Predicts 'New Paradigm of AI Architectures' Within 5 Years and 'Decade of Robotics*,*'* TECHCRUNCH (Jan. 23, 2025, at 07:28 PT), https://techcrunch.com/2025/01/23/metas-yann-lecun-predicts-a-new-ai-architectures-paradigm-within-5-years-and-decade-of-robotics/ [https://perma.cc/SL5C-9VH7] (discussing another area of AI development).

48. *See, e.g.*, Jacqueline C. Charlesworth, *Generative AI's Illusory Case for Fair Use*, 27 VAND. J. ENT. & TECH. L. 323, 359 (2025) ("Copying of protected works by generative AI systems has no similar claim to transformativeness. Works are copied in their entirety and mechanically encoded in the AI model without offering any search mechanism or other functional utility, let alone criticism or commentary.").

49. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1284 (2023).

recognition of fair uses in *Google* and *Sony*, which involved fair uses in the development and usage of new technologies.[50] Applying *Warhol*'s use-by-use approach to the controversies raised by AI, the Article shows how courts should examine each specific use of copyrighted materials alleged to be infringement or fair use. Use to train and develop an AI model is one use. But a different use occurs if AI produces outputs substantially similar to copyrighted works on which the AI model was trained, such as so-called regurgitated copies. Under *Warhol*, courts can find that uses in AI training serve a fair purpose but uses in AI outputs that are "regurgitated" or substantially similar copies do not. The "same copying may be fair when used for one purpose but not another."[51]

Third, this Article critiques the Copyright Office's pre-publication report on AI training and its endorsement of a new, untested theory of "market dilution" under Factor 4 of fair use.[52] The Copyright Office's view of market dilution of copyrighted works derived from the mere creation of new, non-infringing AI-generated works in the same genre or style of works (e.g., romance novels or music) used to train an AI model is not only "uncharted territory" as the Copyright Office conceded,[53] but also unconstitutional. Under the Progress Clause, copyright cannot be used to protect authors beyond "*their* respective [w]ritings," or from general market competition posed by non-infringing works.[54] Fair use is a First Amendment safeguard—it cannot be used to penalize the creation of non-infringing expression singling out one disfavored class of creators. Doing so also violates the First Amendment under which "*more speech*, not less, is the governing rule."[55]

Finally, this Article explains how *technological* progress is equally as important as *creative* progress. The training of AI models with copyrighted works can serve a fair use purpose to

---

50. *See id.* at 1277 (relying on Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183 (2021); Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984); and other Supreme Court fair use precedent). As the *Warhol* Court reminded, "[F]air use is a 'flexible' concept." *Id.* at 1274 (quoting *Google*, 141 S. Ct. at 1197).

51. *Id.* at 1277.

52. PRE-PUBLICATION REPORT, *supra* note 40, at 64–65.

53. *Id.*

54. *See* U.S. CONST. art. I, § 8, cl. 8 (emphasis added) ("The Congress shall have Power . . . To 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."); 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, 755–56 (2001) (describing this clause as "Progress Clause").

55. Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 361 (2010) (emphasis added); *see* Eldred v. Ashcroft, 537 U.S. 186, 219–20 (2003) (describing fair use as a First Amendment safeguard within copyright law).

develop a "highly creative and innovative tool," or "a new and transformative program,"[56] which is "consistent with that creative progress that is the basic constitutional objective of copyright itself" under the Progress Clause, which envisions the grant of exclusive rights as the means to promote the progress of science and useful arts.[57]

Part I distills important principles for Factor 1 of fair use ("the purpose and character of the use" of the plaintiff's work) from the Supreme Court's most recent decision of fair use, *Warhol*, which instructs that fair use must be analyzed on a use-by-use basis.[58] Part II distills important principles for fair uses related to developing or using new technologies under the Court's decisions in *Google* and *Sony*. The application of fair use to foster the development of new technologies is not only consistent with Congress's purpose in codifying the judge-made doctrine, but it also ultimately serves the constitutional goal of promoting progress in the United States.[59] Part III applies these fair use principles to AI training at both universities and AI companies. In either context, an AI developer's use of copyrighted works to train AI models to research, develop, create, and improve an AI model can serve a legitimate fair use purpose: namely, to create a new AI technology with public benefits. But, under *Warhol*, the outputs of AI generators, once deployed, must be analyzed on a use-by-use basis.[60] Some outputs, such as regurgitated or other substantially similar copies, may lack a fair use purpose and constitute infringement for which AI companies can be held liable if secondary liability is proven. Part IV responds to criticisms.

At the outset, it is important to underscore that a conclusion by the courts under Factor 1 that the use of copyrighted works to develop an AI model has a fair use purpose to create a new technology with substantial non-infringing uses does not mean the use of copies for AI training is automatically a fair use. The defendant's transformative purpose is just one factor that is balanced as "a matter of degree, and the degree of difference must be weighed against other considerations, like commercialism."[61] The remaining fair use factors, including the commerciality and the public benefit of use, as well as the potential

56. Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1203, 1209 (2021).

57. *Id.* at 1188; *see* Pollack, *supra* note 54, at 809.

58. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1271, 1284 (2023).

59. 17 U.S.C. § 107; U.S. CONST. art. I, § 8, cl. 8.

60. *See Warhol*, 143 S. Ct. at 1284.

61. *Id.* at 1273.

that such use results in cognizable market harm to the copyright holders' market for their works, including derivative works, must be weighed based on the evidence in each lawsuit.[62] But courts and juries should not ignore AI training's purpose of researching and developing new technologies—consistent with the Progress Clause's aim.

II. FAIR USE IS EXAMINED USE-BY-USE

The Supreme Court's decisions provide the starting point for examining fair use. Although Congress codified the judge-made doctrine of fair use in the Copyright Act of 1976,[63] courts continue to develop this "flexible" doctrine on a case-by-case basis. The Supreme Court's recent decision in *Warhol* made clear that, in analyzing fair use, courts must analyze the defendant's uses of a copyright owner's work on *a use-by-use basis*.[64] This Part explains *Warhol*'s use-by-use review of fair use.

A. Framing Warhol*: Analyzing Fair Use, Use by Use*

Legal commentators continue to parse the Supreme Court's *Warhol* decision as if reviewing a new artwork of renown—or perhaps a kind of legal Rorschach test.[65] One line of scholarship focuses on the Court's instruction to examine the question of fair use use-by-use.[66] To understand this approach, consider how the

62. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 11 F.4th 26, 37 (2d Cir. 2021). Given this fact-specific inquiry, requiring the evidence of the respective parties, this Article refrains from analyzing all four factors of fair use. The resolution of the forty plus copyright lawsuits will depend on the particular models and how they were trained by the respective defendants, not to mention the respective juries or judges that decide the issue at trial or on summary judgment.

63. H.R. REP. NO. 94-1476, at 66 (1976); 17 U.S.C. § 107 (codifying fair use doctrine and including four factors for courts to weigh in determining whether a use is fair). Due to page constraints, this Article focuses on the important first factor, "the purpose and character of the [defendant's] use" of the copyrighted work. 17 U.S.C. § 107.

64. *Warhol*, 143 S. Ct. at 1284.

65. *See, e.g.*, Neal Feigenson, *Say It with Pictures: Image and Text in* Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 76 ALA. L. REV. 79, 125 (2024); Shyamkrishna Balganesh & Peter S. Menell, *Going "Beyond" Mere Transformation:* Warhol *and Reconciliation of the Derivative Work Right and Fair Use*, 47 COLUM. J.L. & ARTS 413, 443 (2024); Michael D. Murray, *Generative AI Art: Copyright Infringement and Fair Use*, 26 SMU SCI. & TECH. L. REV. 259, 271 (2023).

66. *See, e.g.*, Pamela Samuelson, *Did the Solicitor General Hijack the* Warhol v. Goldsmith *Case?*, 47 COLUM. J.L. & ARTS 513, 545 (2024) ("*Warhol* agreed with OSG [Office of the Solicitor General] that fair use must be assessed on a use-by-use basis," but noting the full import of this approach is contestable); Timothy J. McFarlin, *Infringing Uses, Not Works*, 76 S.C. L. REV. 103, 122 (2024) (contending that *Warhol*'s use-by-use review of fair use requires courts to consider infringement based on an alleged unauthorized derivative work on a use-by-use basis); Sandra M. Aistars, *Copyright's Lost Art of Substantial Similarity*, 26 VAND. J. ENT. & TECH. L. 109, 155 (2023) (arguing that *Warhol* requires use-

Court framed the fair use analysis in *Warhol*. The case involved a copyright dispute involving two visual works.[67] In 1981, the photographer Lynn Goldsmith took black-and-white portraits of the musician Prince.[68] In 1984, *Vanity Fair* licensed one of Goldsmith's portraits of Prince to be used as an artistic reference for an artist to create a new artwork depicting Prince for its magazine, a use of Goldsmith's portrait licensed for one time only.[69] That year, *Vanity Fair* published the magazine containing the artist's rendition of Prince in a story about the musician.[70] Unbeknownst to Goldsmith, the commissioned artist was none other than Andy Warhol.[71] It turned out he made not just the artwork published in *Vanity Fair*, but an entire Prince Series consisting of four black-and-white sketches based on Goldsmith's portrait and, by using Warhol's silkscreen technique,[72] twelve color renditions of Prince's portrait, all with the same pose of Prince traced from Goldsmith's photograph.[73]

Goldsmith only discovered the other Warhol artworks years later, in 2016, after Condé Nast published a tribute magazine to Prince upon his untimely death.[74] On the cover was one of Warhol's other renditions of Prince—the so-called *Orange Prince* referring to its color—which Condé Nast had licensed for use from the Andy Warhol Foundation (AWF or Foundation), but not from Goldsmith.[75] That led to a lawsuit involving AWF and Goldsmith, who disagreed over whether Warhol's Prince Series involved a fair use of Goldsmith's portrait.[76] Goldsmith claimed that Warhol's

---

by-use analysis of fair use, but "the same logic does not apply to analyzing basic infringement," at least when a court finds no substantial similarity, which "applies to the work itself"); Jessica Silbey & Eva E. Subotnik, *What the* Warhol *Court Got Wrong: Use as an Artist Reference and the Derivative Work Doctrine*, 47 COLUM. J.L. & ARTS 353, 359 (2024) ("Without limiting *Warhol* to its facts, the Supreme Court decision appears to narrow the 'transformative use' test without overruling it, as well as establish a use-by-use fair use assessment of otherwise lawfully made works, which would be the first explicit articulation of such a rule."); Xiyin Tang, *Art After* Warhol, 71 UCLA L. REV. 870, 931 (2024) ("The Court had instead emphasized that fair use determinations must be made on a use-by-use basis—and, consequently, that its ruling did not reach the actual Warhol painting itself.").

67. *Warhol*, 143 S. Ct. at 1266.

68. *Id.* at 1267.

69. *Id.* (quoting license). For more about the common practice of artistic reference, see Silbey & Subotnik, *supra* note 66, at 382.

70. *Warhol*, 143. S. Ct. at 1267.

71. *Id.* at 1267–68.

72. *Id.* at 1294 (Kagan, J., dissenting) (describing Warhol's silkscreen process).

73. *Id.* at 1288 (images shown in Appendix).

74. *Id.* at 1268–69.

75. *Id.*

76. *Id.* at 1271.

Orange Prince, licensed by AWF to Condé Nast, infringed her copyright to the portrait of Prince on which Warhol had originally relied on for creating the works in the series.[77]

AWF revealed that, after "Warhol's death in 1987, portraits from the Prince Series have been sold or auctioned more than two dozen times," and "AWF transferred four works from the Prince Series to the Andy Warhol Museum in Pittsburgh."[78] Although the courts in the litigation never resolved whether any of these other uses of the Goldsmith portrait by the Warhol works were infringing, the courts were cognizant of the issue—perhaps worried about rendering the entire Warhol's Prince series illegal, even at the Warhol Museum, which draws 100,000 visitors annually.[79]

The Supreme Court stated these sales and uses of the other works were not before it.[80] The reason was not simply that Goldsmith disclaimed pursuing them.[81] Instead, the reason was more fundamental to the fair use analysis: "The fair use provision, and the first factor in particular, *requires an analysis of the specific 'use' of a copyrighted work that is alleged to be 'an infringement*.'"[82] The passage was presaged by the Court's earlier emphasis that "[h]ere, the *specific use* of Goldsmith's photograph alleged to infringe her copyright is AWF's licensing of Orange Prince to Condé Nast."[83] The Court belabored this point:

> Here, Goldsmith's copyrighted photograph has been used in *multiple ways*: After Goldsmith licensed the photograph to Vanity Fair to serve as an artist reference, Warhol used the photograph to create the Vanity Fair illustration and the other Prince Series works. Vanity Fair then used the photograph, pursuant to the license, when it published Warhol's illustration in 1984. Finally, AWF used the photograph when it licensed an image of Warhol's Orange Prince to Condé Nast in 2016. Only that last use, however, AWF's commercial licensing of Orange

77. *See id.* at 1267–69, 1271.

78. *See* The Andy Warhol Found. for the Visual Arts, Inc.'s Memorandum of L. in Support of Its Motion for Summary Judgment at 26, Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 382 F.3d 312 (2019) (No. 1:17-cv-02532-JGK).

79. *Warhol*, 143 S. Ct. at 1264; *The Andy Warhol Museum*, HAMILTON-SELWAY, https://hamiltonselway.com/andy-warhol-museum-need-know/ [https://perma.cc/GJF5-6N2N] (last visited July 19, 2025).

80. *See Warhol*, 143 S. Ct. at 1278.

81. *Id.* at 1278 n.9 ("AWF sought a declaratory judgment that would cover the original Prince Series works, but Goldsmith has abandoned all claims to relief other than her claim as to the 2016 Condé Nast license and her request for prospective relief as to similar commercial licensing.").

82. *Id.* at 1277 (emphasis added).

83. *Id.* at 1273 (emphasis added).

> Prince to Condé Nast, is alleged to be infringing. *We limit our analysis accordingly. In particular, the Court expresses no opinion as to the creation, display, or sale of any of the original Prince Series works*.[84]

In a footnote, the majority rejected the dissent's assumption that "*any and all uses* of an original work entail the same first-factor analysis based solely on the content of a secondary work."[85] The majority found such an assumption "contradicts the fair use statute and this Court's precedents."[86]

The Court cited both *Campbell* and *Sony* for applying this use-by-use review.[87] In *Campbell*, a case involving the rap band 2 Live Crew's parody of the song "Oh, Pretty Woman," the Court contrasted the use of a copyrighted work "to advertise a product, even in a parody," with "the sale of a parody for its own sake, let alone one performed a single time by students in school."[88] Likewise, in *Sony*, a case involving home taping of broadcast shows to watch later, the Court contrasted the recording of TV "for a commercial or profit-making purpose" with "private home use."[89]

Table 1 below indicates a range of uses by Warhol, the AWF, and others, listed in chronological order. Had Goldsmith challenged the other uses of her work by Warhol or the Foundation, they would have been analyzed separately. The third, shaded row indicates the only use of Goldsmith's work at issue in the Supreme Court.

84. *Id.* at 1277–78 (emphasis added) (footnote omitted).
85. *Id.* at 1278 n.10 (emphasis added).
86. *Id.*
87. *Id.* at 1277.
88. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 571, 585 (1994).
89. Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 449–51 (1984).

Table 1. Various Uses of Goldsmith's Portrait by Warhol, Foundation, and Others.[90]

User	**Use**
Andy Warhol	Creation of Own Work in 1984. Use Goldsmith photograph to create sixteen new Prince artworks, at least one under license between Goldsmith and *Vanity Fair*.
Andy Warhol	Public distribution of copy in 1984. Use for license of one Prince artwork for publication in *Vanity Fair* in 1984 under license between *Vanity Fair* magazine and Goldsmith.
Foundation At issue in *Andy Warhol Foundation v. Goldsmith*.	Public distribution of copy in 2016. Use for license of *Orange Prince* artwork for publication on magazine cover in 2016 under license between *Vanity Fair* magazine and Foundation.
Andy Warhol or Foundation	Public distribution of Warhol artworks (times unspecified). Use in any sale and public distribution of original silkscreen prints of Prince Series.
Andy Warhol, Foundation, or Andy Warhol Museum	Public display of Warhol artworks. Use in any specific instance of public display of Prince Series, such as in a nonprofit museum.
Andy Warhol, Foundation, or Andy Warhol Museum	Other public distribution of copies of Warhol artworks. Use in any specific instance of making copies and publicly distributing Prince Series, such as in "a for-profit book commenting on 20th-century art."

90. *Warhol*, 143 S. Ct. at 1266–68, 1273, 1278, 1288, 1291.

Art collectors of Prince Series artworks	Public display of Warhol artworks. Use in any specific instance of public display of Prince Series, such as in a nonprofit museum.
Art collectors of Prince Series artworks	Other public distribution of copies of Warhol artworks. Use in any specific instance of making copies and publicly distributing Prince Series.

B. Corollary: The Same Copying May Be Fair when Used for One Purpose but Not Another

The Court's discussion of *Campbell* and *Sony* illuminates a corollary of the use-by-use analysis of fair use: "The same copying may be fair when used for one purpose but not another."[91] Although the Court abstained from reviewing the other uses of Goldsmith's portrait that Warhol and the Foundation made, this corollary left open the possibility that some of Warhol's or the Foundation's uses were fair uses.[92] In concurrence, Justice Gorsuch suggested it:

> [W]hile our interpretation of the first fair-use factor does not favor the Foundation in this case, it may in others. If, for example, the Foundation had sought to display Mr. Warhol's image of Prince in a nonprofit museum or a for-profit book commenting on 20th-century art, the purpose and character of that use might well point to fair use.[93]

Judge Jacobs took a similar view in the court of appeals below.[94]

The concern stemmed from the fact that "[t]he sixteen original works have been acquired by various galleries, art dealers, and the Andy Warhol Museum."[95] If Warhol's Prince Series were held to be infringing derivative works at Warhol's creation, then the galleries, art dealers, and the Andy Warhol Museum all might be committing copyright infringement.

That possibility still lurks. In the settlement of the case, Goldsmith "confirmed that because of the expiration of the statute

91. *Id.* at 1263.

92. *Id.* at 1291 (Gorsuch, J., concurring).

93. *Id.*

94. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 11 F.4th 26, 54–55 (2d Cir. 2021) (Jacobs, J., concurring).

95. *Id.* at 54.

of limitations she is no longer advancing any of those claims," which AWF understood to include "any claims for relief for the original creation of the Prince Series because of the expiration of the statute of limitations."[96] Given the separate accrual rule for copyright law's statute of limitations,[97] even the settlement does not foreclose future claims of infringement against the Prince Series.

C. The Importance of Defendant's Purpose: Did the Defendant Have a Further Purpose or Different Character of Use?

The next important principle to draw from *Warhol* is the need to determine if the defendant had a "further purpose or different character" of use than the author of the work.[98] Indeed, the Court used this term eight times in its analysis.[99] And it used "different purpose" two times.[100] Although the Court didn't distinguish between "further" and "different" purpose, the ordinary meaning of the terms denote slightly different concepts.

"Further" means "additional" or "extending beyond."[101] "Different" means "not the same" or "dissimilar."[102] Although the two words overlap in meaning, "further" conveys something that "different" does not. A further purpose encompasses a scenario in which the secondary user of a work has one purpose similar to the author's reason for creating or disseminating the work but has an *additional* purpose *beyond* the author's, a scenario expressly contemplated by the Court in *Warhol*.[103]

For example, in *Google*, Google had the exact same purpose in using the same Java declaring code to call up certain tasks repeated in different computer programs, but it also had the *additional* purpose of creating an entirely new product, or

96. *See* Final Judgment at ¶¶ 2–3, Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 382 F.3d 312 (2019) (No. 1:17-cv-02532-JGK).

97. Separate accrual means that "when a defendant commits successive violations, the statute of limitations runs separately from each violation." Petrella v. Metro-Goldwyn-Mayer, Inc., 572 U.S. 663, 671 (2014).

98. *Warhol*, 143 S. Ct. at 1277.

99. *Id.* at 1262–63, 1275–77, 1299, 1311.

100. *Id.* at 1274, 1289.

101. *Further*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/further [https://perma.cc/YG8R-QQ6X] (last visited July 15, 2025).

102. *Different*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/different [https://perma.cc/9V9V-PVQN] (last visited July 29, 2025).

103. *Warhol*, 143 S. Ct. at 1274–75 ("Whether a use shares the purpose or character of an original work, or instead has a further purpose or different character, is a matter of degree."). The Court described this inquiry as, not subjective, but instead "objective," examining "what use was made, *i.e.,* what the user does with the original work." *Id.* at 1284.

computing platform for smartphones.[104] That is why the Supreme Court said Google had "the same reason" that the author of Java did, but the Court went "further" to find Google had the additional, or further, purpose of creating an innovative new tool.[105] Google's use of Java declaring code was not completely different from its intended purpose, but it nonetheless had the "further" purpose to create a new technology that went beyond anything Sun Microsystems did. This goal of developing a new technology was a transformative fair use purpose.[106]

The Court has not discussed "further purpose" and "different character" as two separate concepts, instead typically lumping them together. But, a few times, the Court has focused on character alone, such as in describing the "commercial character,"[107] "educational character,"[108] or "parodic character" of the defendant's use.[109] Yet the Court has also described a "parodic purpose," which suggests that purpose and character might overlap or refer to the same general concept.[110] Even in this context, they arguably refer to distinct concepts. A secondary user might have a parodic purpose but so poorly execute it that the character of the user's work does not appear like a parody—or contains only a tiny parodic element that can be arguably perceived.

In any event, if the defendant has the same purpose or one substantially similar to the author's, the defendant lacks a "further purpose or different character."[111] *Warhol* clarified that, in analyzing the defendant's purpose of use of the copyrighted works, a court examines how different or similar it is to the purpose of the original work.[112] The Court concluded: "Taken together, these two elements—that Goldsmith's photograph and AWF's 2016 licensing of Orange Prince share *substantially the same purpose*, and that AWF's use of Goldsmith's photo was of a *commercial nature*—counsel against fair use, absent some other

104. *See* Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1203 (2021).
105. *See id.*
106. *See id.*
107. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 572, 579 (1994); *Warhol*, 143 S. Ct. at 1280.
108. *Campbell*, 510 U.S. at 585.
109. *Id.* at 582.
110. *See id.* at 589.
111. *Id.* at 579.
112. *Warhol*, 143 S. Ct. at 1274.

justification for copying."[113] The substantially same purpose was licensing the respective visual work of Goldsmith's or Warhol's for the cover of a magazine.[114] The same purpose of use likely signals substitution of the copyright holder's work: "The use of an original work to achieve a purpose that is the same as, or highly similar to, that of the original work is more likely to substitute for, or 'supplan[t],' the work."[115] Instead of licensing Goldsmith's portrait of Prince, a magazine can license Warhol's portrait, thereby substituting for Goldsmith's work being used in the same magazine.

The Court rejected AWF's argument that the use in Warhol's work weighs in favor of fair use because it "has a new meaning or message" and is therefore transformative in purpose, favoring fair use.[116] Said the Court: "*Campbell* cannot be read to mean that § 107(1) weighs in favor of any use that adds some new expression, meaning, or message. Otherwise, 'transformative use' would swallow the copyright owner's exclusive right to prepare derivative works."[117] The Court was concerned that a broad assertion of transformative use in creating new works under fair use would undermine the copyright holder's right to make derivative works.[118] The Court thus viewed the examination of purpose under Factor 1 as "a *matter of degree*, and the degree of difference must be balanced against the commercial nature of the use."[119]

The Court therefore rejected the defendant's creation of new expression that includes a copied portion of the plaintiff's work as automatically favoring fair use: "Although new expression may be relevant to whether a copying use has a sufficiently distinct purpose or character, it is not, without more, dispositive of the first factor."[120] The Court cautioned that its reasoning "does not mean, however, . . . derivative works borrowing heavily from an original cannot be fair uses."[121] It pointed to Warhol's *Soup Cans* artwork, which (famously) depicts a Campbell's soup can for a different

113. *Id.* at 1280 (emphasis added).
114. *Id.* at 1280–81.
115. *Id.* at 1274 (alteration in original).
116. *Id.* at 1281–82.
117. *Id.* at 1282.
118. *Id.*
119. *Id.* at 1277 (emphasis added).
120. *Id.* at 1273.
121. *Id.* at 1280.

purpose ("an artistic commentary on consumerism") than the trademark purpose of the can's advertising logo.[122]

In sum, *Warhol* instructs that each of the defendant's uses should be analyzed on a use-by-use basis under fair use. If the defendant's use has a "further purpose or different character" than the author did in disseminating the work, then such use tends to weigh in favor of fair use in the balance of the four factors. But the purpose and character of use is "a matter of degree . . . weighed against other considerations."[123]

III. TECHNOLOGICAL FAIR USES

Part III traces the case law that recognizes fair uses to develop new technologies, including the Supreme Court's important decision in *Google v. Oracle*.[124] Courts considering fair use defenses in the AI litigation should not read *Warhol* and *Google* as mutually exclusive precedents requiring courts to choose one or the other to follow. Instead, these precedents, which were decided close in time, should be read and applied together.

A. Congress Expected Fair Use Would Resolve Disputes over Uses Related to New Technologies

The Supreme Court has decided six cases of fair use under the Copyright Act.[125] Two cases involved a novel question of fair use in the context of new technologies.[126] The Supreme Court had the opportunity to consider fair use in two other new technology cases, but resolved the disputes over file-sharing software and the copy

122. *Id.* at 1281.

123. *Id.* at 1273.

124. Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1203 (2021).

125. *See Warhol*, 143 S. Ct. at 1287 (holding that the licensing for a magazine cover of a Warhol work, which was based on a portrait taken by Goldsmith, had substantially similar purpose of portrait, factoring against fair use); *Google*, 141 S. Ct. at 1190, 1194 (upholding jury verdict that Google's copying of Java declaring code for use in Android operating system was fair use); Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579, 582–83, 594 (1994) (recognizing that defendant's copying of parts of song to make a parody of it was a transformative purpose, favoring fair use); Stewart v. Abend, 495 U.S. 207, 236–38 (1990) (continued exploitation of movie or derivative work made from magazine story during renewal term was not fair use); Harper & Row v. Nation Enters., 471 U.S. 539, 542 (1985) (rejecting fair use defense of magazine that copied and published parts of purloined copy of President Ford's memoirs to "scoop" the sale of the book); Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 449–51 (1984) (holding that time-shift recording broadcast TV shows for later viewing at home was fair use).

126. *See Google*, 141 S. Ct. at 1190–91 (use of Java declaring code in Android operating system); *Sony*, 464 U.S. at 419–20 (use of VCR to record broadcast TV programs).

machine without addressing the issue[127] or deciding it by an equally divided court.[128] Before explaining the Court's teachings in *Google* and *Sony*, it is crucial to understand why it is no surprise that the Court has faced technology-related fair use cases.[129]

When Congress codified the judge-made doctrine of fair use in the Copyright Act of 1976, Congress saw the utility of fair use as a way for the courts to address whether new uses of copyrighted works precipitated by "technological change" should be recognized as fair uses.[130] As the Federal Circuit explained, "[t]he legislative history of section 107 suggests that courts should adapt the fair use exception to accommodate new technological innovations."[131] Citing the House Report to the 1976 Act, the *Google* Court elaborated: "[C]ourts are to 'adapt the doctrine [of fair use] to particular situations on a case-by-case basis' and in light of 'rapid technological change.'"[132] "In a word, we have understood the provision to set forth general principles, the application of which requires judicial balancing, depending upon relevant circumstances, including 'significant changes in technology.'"[133]

This technology-accommodating purpose of fair use avoids the recurring predicament of treating every new technological use as infringement absent an amendment by Congress. In enacting the 1976 Act, Congress sought to avoid the inherent inefficiencies of prior copyright acts whose provisions were technology-specific, thereby requiring Congress to enact amendments for major

127. *See* Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd., 545 U.S. 913, 919–20 (2005) (file sharing software). The parties trained their arguments on the *Sony* safe harbor and what proof of substantial non-infringing uses of a technology satisfied *Sony*'s requirement. Had the Court decided that issue, it would have considered file-sharing's capability for non-infringing uses, including fair use. *See id.* at 933–34; *id.* at 945 (Ginsburg, J., concurring) (contending that defendants' software service would not qualify for *Sony* safe harbor, and noting "[h]ere, there has been no finding of any fair use and little beyond anecdotal evidence of noninfringing uses"); *id.* at 949, 952–54 (Breyer, J., concurring) (contending that defendants' software service would qualify for *Sony* safe harbor).

128. Williams & Wilkins Co. v. United States, 487 F.2d 1345, 1348 (Ct. Cl. 1973), *aff'd by an equally divided court*, 420 U.S. 376, 376 (1975).

129. *See* Pamela Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, 71 UCLA L. REV. 1484, 1494 (2024); Edward Lee, *Technological Fair Use*, 83 S. CAL. L. REV. 797, 805–07 (2010).

130. H.R. REP. NO. 94-1476, at 66 (1976) ("The bill endorses the purpose and general scope of the judicial doctrine of fair use, but there is no disposition to freeze the doctrine in the statute, especially during a period of rapid technological change."); S. REP. NO. 94-473, at 62 (1975).

131. Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 843 (Fed. Cir. 1992).

132. Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1198 (2021) (second alteration in original) (quoting H.R. REP. NO. 94-1476, at 66 (1976)).

133. *Id.* at 1197 (quoting Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 430 (1984)).

changes caused by new technologies that went beyond the literal terms of the acts.[134]

This approach of U.S. copyright law to accommodate technological changes through fair use is an American doctrine.[135] It differs greatly from the approach of continental European and other civil law countries, which do not recognize fair use at all, instead adopting specific and much narrower copyright exceptions.[136] This global divide is now the subject of contentious debate with some proponents seeking the adoption of fair use in other countries, partly for the flexibility it offers and its ability to resolve copyright disputes raised by disruptions caused by new technologies without requiring an amendment by the legislature.[137] Taking sides in this debate is not our task. Suffice it to say, fair use originated as an American doctrine—and was codified by Congress with technological change in mind.

The relative importance of technology-related fair use decisions in the Supreme Court's fair use jurisprudence is understandable. Indeed, it is an *essential* component of the copyright system Congress established. These technology-related cases are not the run-of-the-mill copyright lawsuits. Instead, they often involve fundamental questions about not only the scope of copyright (which every fair use defense asks), but also, more importantly, the need to balance the United States's interest in promoting technological innovation—which has profound national economic importance.

For example, in 2020, around the time the Supreme Court decided *Google v. Oracle*, Google's Android had 40.27% of the U.S. market share and 73.06% of the global market share in smartphones.[138] Similarly, in 1985, shortly after the Supreme Court's decision in *Sony Corp. v. Universal City Studios*, about 20% of U.S. households owned VCRs; by 1989, 65% of households did, comprising

134. *See* Edward Lee, *AI and the Sound of Music*, 134 YALE L.J.F. 187, 198–99 (2024); Brad A. Greenberg, *Rethinking Technology Neutrality*, 100 MINN. L. REV. 1495, 1510–11 (2016).

135. *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 576 (1994) ("[T]he doctrine was recognized by the American courts . . . .").

136. *See* Naama Daniel, *Lost in Transit: How Enforcement of Foreign Copyright Judgments Undermines the Right to Research*, 38 AM. U. INT'L L. REV. 87, 104 (2023) ('[C]ivil law states tend to have confined lists of exceptions to copyright, and their courts tend to interpret them narrowly . . . .").

137. *See, e.g.*, Christophe Geiger & Elena Izyumenko, *Towards a European "Fair Use" Grounded in Freedom of Expression*, 35 AM. U. INT'L L. REV. 1, 6–7 (2019); Peter K. Yu, *Fair Use and Its Global Paradigm Evolution*, 2019 U. ILL. L. REV. 111, 134–35.

138. *iPhone vs Android Statistics*, BACKLINKO (Aug. 21, 2025), https://backlinko.com/iphone-vs-android-statistics [https://perma.cc/HSY7-G7BT].

62 million VCRs.[139] The VCR gave rise to the new market for watching movies on the VCR, an unexpected financial boon for Hollywood movie studios through the sales of movie videos.[140] Home viewing of TV shows led to today's practices of watching "on demand" and eventually "binge-watching."[141] The Court's decision in both cases had significance for millions of Americans who used and made popular the technology at issue.

This is not to suggest that, whenever a new technology is involved, a fair use must be found. Instead, it is to make clear the role that fair use plays in striking a balance between copyright holders' legitimate interests and accommodating technological innovation in the United States. As discussed below, along with the *Sony* safe harbor, fair use is central to preventing copyright from stifling technological innovation. Copyright is not intended to give authors' rights to control technologies with substantial non-infringing uses or their non-infringing outputs, as explained next.[142]

B. How Google *and* Sony *Promote and Protect Technological Progress for the United States*

1. Fostering Technology Through Fair Use. The Supreme Court's two technology-accommodating fair use decisions, *Sony* and *Google*, provide exemplars for understanding two basic ways in which new technologies can use copyrighted works. We can add the two circuit court fair use decisions that the Supreme Court favorably cited in *Google*: the Ninth Circuit's decisions in *Sega Enterprises Ltd. v. Accolade, Inc.* and *Sony Computer Entertainment Inc. v. Connectix Corp.*[143] These cases can be referred to as the "reverse-engineering" or "intermediate use" cases involving a defendant's making copies of copyrighted works for internal, nonpublic use during the process of

---

139. AUGUST E. GRANT & JENNIFER H. MEADOWS, COMMUNICATION TECHNOLOGY UPDATE 35 tbl. 2.11 (10th ed. 2006); Alexi Horowitz-Ghazi, *How the VCR Began America's Love of On-Demand Content*, NPR (Aug. 6, 2016, at 17:05 ET), https://www.npr.org/2016/08/06/489002713/how-the-vcr-began-americas-love-of-on-demand-content [https://perma.cc/D6ND-2PFY].

140. *See* JAMES LARDNER, FAST FORWARD: HOLLYWOOD, THE JAPANESE, AND THE ONSLAUGHT OF THE VCR 316–17 (1987).

141. *See generally* Andrew Moodie, *A Short History of Watching Films at Home*, REVIEWSPHERE (Mar. 26, 2020), https://www.reviewsphere.org/news/a-short-history-of-watching-films-at-home/ [https://perma.cc/YQS4-P49L] (summarizing the public's transition from TV to VHS to DVD to streaming as the primary form of media consumption); Kelly West, *Unsurprising: Netflix Survey Indicates People Like to Binge-Watch TV*, CINEMABLEND (Dec. 13, 2013), https://www.cinemablend.com/television/Unsurprising-Netflix-Survey-Indicates-People-Like-Binge-Watch-TV-61045.html [https://perma.cc/D8AY-243T].

142. *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 441 (1984).

143. Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 598–99 (9th Cir. 2000); Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1527–28 (9th Cir. 1992).

getting access to the unprotected elements of the works, including, in *Sega*, to create new, non-infringing works.[144] In addition, the *Warhol* Court relied on *Authors Guild v. Google, Inc.*, a case involving Google's fair use in copying millions of books to create and operate Google Book Search, a technology enabling in-text search of books.[145]

All but *Sony* involve uses of copyrighted works to develop or create a new technology or a new program (putting aside the likely possibility that *Sony* tested its device's capability to record TV shows).[146] By contrast, *Sony* is a case involving the output of a new technology, a VCR, in enabling consumers to make copies of broadcast TV shows for later viewing.[147]

As shown in Table 2 below, we can categorize these cases in terms of their respective uses of copyrighted works: (1) technology development by developers or (2) technology usage by consumers. *Technology development* refers to the process of researching, developing, testing, and ultimately creating a new technology or a new program. *Technology usage* refers to how the technology is used, such as consumers' use of the technology after a public launch.[148]

---

144. *See id.* For other applications of these precedents, see also Assessment Techs. of WI, LLC. v. WIREdata, Inc., 350 F.3d 640, 645 (7th Cir. 2003) ("[I]f the only way WIREdata could obtain public-domain data about properties in southeastern Wisconsin would be by copying the data in the municipalities' databases as embedded in Market Drive, so that it would be copying the compilation and not just the compiled data only because the data and the format in which they were organized could not be disentangled, it would be privileged to make such a copy . . . ."); Bateman v. Mnemonics, Inc., 79 F.3d 1532, 1539 n.18 (11th Cir. 1996) (agreeing with *Sega*).

145. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1286 (2023) (citing Authors Guild v. Google, Inc., 804 F.3d 202, 215 (2d Cir. 2015)); *Authors Guild*, 804 F.3d at 208–09, 211.

146. *See* Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1193 (2021) (copied Java declaring code for use in new smartphone operating system); *Connectix*, 203 F.3d at 598–99 (copied operating system of game console to make an interoperable program); *Sega*, 977 F.2d at 1514 (same); *Authors Guild*, 804 F.3d at 215 (copied books to create a full-text search function).

147. *See Sony*, 464 U.S. 417, 419–20 (use of VCR to make copies for broadcast TV shows for later viewing). We might include the *Williams & Wilkins Co.* case decided by the Court of Claims under the old 1909 Copyright Act: the court held that the photocopies of scientific journal articles that the libraries for National Institutes of Health (NIH) and National Library of Medicine (NLM) made were fair uses. Williams & Wilkins Co. v. United States, 487 F.2d 1345, 1348–50, 1353 (Ct. Cl. 1973), *aff'd by an equally divided court*, 420 U.S. 376 (1975). The decision, like *Sony*, involves the output of a then-new technology. An equally divided Supreme Court upheld the decision. *Id.*

148. My use of "technology development" and "technology usage" to describe two stages is similar to the Copyright Office's description of the "development" of an AI model and its later "deployment" in an AI system. *Cf.* PRE-PUBLICATION REPORT, *supra* note 40, at 2, 4, 21–22 (describing "development" as creating generative AI models and "deployment" as tailoring and promoting generative AI models for varied purposes). For simplicity, I use the term "technology development" to encompass all stages of developing the technology prior

In the last column, Table 2 indicates whether the defendant's use of copyrighted work served a fair use or "further" purpose, either in a transformative or non-transformative way. Drawing from a law review article written by Judge Pierre Leval, the Supreme Court has described a defendant's transformative use of copyrighted work in this way: the use "adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message."[149] The Supreme Court also used "transformative" to describe Google's new Android platform, "a highly creative and innovative tool for a smartphone environment."[150]

Google v. Oracle, the Second Circuit's decision in *Authors Guild*, and the Ninth Circuit's decisions in *Sega* and *Connectix* all involved a technology company's fair use of copies of copyrighted works based on the defendant's transformative or different purpose to create a new technology or program for the public's use. In *Google v. Oracle* and *Authors Guild*, the defendant's copying went beyond intermediate copying confined to the development of the technology, as was the case in *Sega* and *Connectix*: Google's Android operating system contained copied elements of Java SE used by consumers and app developers, and Google Books Search had copies of entire books used and stored within Google's database.[151] In the latter case, Google's copying was extensive: it involved millions of physical books that Google scanned to create its database thereby enabling the new function of full-text searching to find relevant books.[152] The *Warhol* Court relied on the *Authors Guild* analysis of fair use, albeit without discussing the technology.[153]

---

to its public launch, including the deployment of a model in a system. I intend the term "technology usage" to refer to use of the AI technology by consumers or users after the technology is released. *Cf.* Ana Raquel Costa-Brito et al., *Home-Based Exercise Interventions Delivered by Technology in Older Adults: A Scoping Review of Technological Tools Usage*, INT'L J. MED. INFORMATICS, Jan. 2024, at 1, 1–2 (review of studies that examine technology usage levels in real world).

149. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) (citing Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990)).

150. *Google*, 141 S. Ct. at 1203, 1209 (calling Google's new Android platform "a new and transformative program").

151. *See id.* at 1203; *Authors Guild*, 804 F.3d at 221 (Google Book Search); *Sega*, 977 F.2d at 1514–15; *Connectix*, 203 F.3d at 601.

152. *See Authors Guild*, 804 F.3d at 208, 221.

153. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1276 (2023) (citing and quoting *Authors Guild*, 804 F.3d at 214).

Table 2. Technology Fair Use Decisions Decided or Favorably Cited by Supreme Court.

Case	**Technology development:** Use of Copyrighted Works to Create New Technology?	**Technology usage:** Use of Copyrighted Works in Public Use of Technology?	**Factor 1:** Use of Copyrighted Works Had Further Purpose or Different Character?
Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984).	No.	Yes, by users of VCR for personal time-shift recordings.	No, time-shifted copies of free TV shows was for watching them.[154]
Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183 (2021).	Yes, use of Java declaring code for Android operating system to facilitate computer programmers' ability to write apps for Android.	Yes, declaring code was part of Android OS and can be used by programmers writing Android apps.	Yes, "use of the Sun Java API seeks to create new products," i.e., "a highly creative and innovative tool for a smartphone environment."[155]
Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992).	Yes, in reverse engineering of OS to find uncopyrightable element necessary for interoperability of new game created by defendant.	No.	Yes, "intermediate copying of computer code as an initial step *in the development of a competing product.*"[156]

154. Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 420–21 (1984).
155. *Google*, 141 S. Ct. at 1191–92, 1203.
156. *Sega*, 977 F.2d at 1514–15, 1518 (emphasis added).

Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).	Yes, in reverse engineering of OS to find uncopyrightable element necessary for game emulator to make plaintiff's games run on computers.	No.	Yes, "*creates a new platform*, the personal computer, on which consumers can play games designed for the Sony PlayStation."[157]
Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).	Yes, to create a database to enable within-book search of published books and to enable text data mining (TDM) analysis of frequency of use of words in entire corpus of books.	Yes, copies stored in database and snippets of books shown in response to user's searches.	Yes, copies in database serve further purpose of searching within-text of all books in database to find relevant sources.[158]

157. Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 601–02, 606 (9th Cir. 2000) (emphasis added).

158. Authors Guild v. Google, Inc., 804 F.3d 202, 207 (2d Cir. 2015).

We can apply this same framework (technology development–technology usage–Factor 1 of fair use) to other technology-accommodating fair use cases from the lower courts, shown in Table 3 below. These cases involve both (i) technology development and (ii) technology usage in which copies of works are stored by the defendant in a large database of content scraped from the internet,[159] copied from physical books owned by universities and libraries,[160] collected from students and used without their permission in a technology to detect plagiarism,[161] and copied from legal briefs filed in litigation to "creat[e] an interactive legal research tool" offered by Lexis and Westlaw, respectively.[162]

This second group of cases involves situations in which the defendant made massive amounts of copies of works of expression, including works within the core of copyright protection under Factor 2 of fair use.[163] They are similar to the Google Book Search case involving the Authors Guild the *Warhol* Court relied on, as noted above in the discussion of Table 2.[164] These cases all involved the creation of a technology that copied numerous copyrighted works and stored them in an internal database of the defendant for certain transformative purposes listed in the last column in Table 3, without publicly disseminating substitutional copies of the works.[165] The technologies all did something different from redistributing copies. The technologies innovated.

159. *See* Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1155–56, 1168 (9th Cir. 2007) (Google image search); Kelly v. Arriba Soft Corp., 336 F.3d 811, 822 (9th Cir. 2003) (Arriba Soft image search); Field v. Google, Inc., 412 F. Supp. 2d 1106, 1122–23 (D. Nev. 2006) (Google search of cache copies of websites).

160. *See* Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 90, 101 (2d Cir. 2014) (HathiTrust Digital Library); *see also Authors Guild*, 804 F.3d at 207, 225 (Google Book Search).

161. *See* A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 634–35 (4th Cir. 2009).

162. White v. W. Publ'g Corp., 29 F. Supp. 3d 396, 398–99 (S.D.N.Y. 2014).

163. *See supra* notes 151–61; 17 U.S.C. § 107.

164. *See supra* notes 159–61.

165. The cases in Table 3 are listed chronologically, by highest court. *See* Kelly v. Arriba Soft Corp., 336 F.3d 811, 822 (9th Cir. 2003) (Arriba Soft image search); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1168 (9th Cir. 2007) (Google image search); *iParadigms*, 562 F.3d at 634 (antiplagiarism detection for student papers); *HathiTrust*, 755 F.3d at 101 (HathiTrust Digital Library of books); Field v. Google, Inc., 412 F. Supp. 2d 1106, 1123 (D. Nev. 2006) (cached copy of websites for Google search); *White*, 29 F. Supp., at 399 (Westlaw and Lexis search tools).

Table 3. Lower Courts' Finding Fair Use in Other Technology Cases.

Case	**Technology development:** Use of Copyrighted Works to Create New Technology?	**Technology usage:** Use of Copyrighted Works in Public Use of Technology?	**Factor 1:** Use of Copyrighted Works Had Further Purpose or Different Character?
Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003).	Yes, to create a searchable database of online images.	Yes, copies stored in database and outputs show thumbnail images.	Yes, copies in database serve further purpose of searching online images to find relevant ones.[166]
Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007).	Yes, to create a searchable database of online images.	Yes, copies stored in database and outputs show thumbnail images of reduced resolution.	Yes, copies in database serve further purpose of searching online images to find relevant ones.[167]
A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009).	Yes, to create a searchable database of student papers.	Yes, copies stored in database but no direct quotations.	Yes, copies in database serve further purpose of finding potential plagiarism in student papers.[168]

166. *Kelly*, 336 F.3d at 815.
167. *Perfect 10*, 508 F.3d at 1155.
168. *iParadigms*, 526 F.3d at 634.

Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, (2d Cir. 2014).	Yes, to create a database to enable within-book search of published books. Secondary use to store digital copies for preservation.	Yes, copies stored in database. No snippets of books shown, but (i) page numbers in books term is found; (ii) access to full books to people with print-disability.	Yes, copies in database serve further purpose of searching within-text of all books in database to find relevant sources.[169]
Field v. Google, Inc., 412 F. Supp. 2d 1106 (D. Nev. 2006).	Yes, to create a searchable database of cached copies of Internet websites.	Yes, copies stored in database and "cached" static copy of website publicly accessible.	Yes, copies in database serve further purpose of allowing static view of snapshot of website, useful when website is down.[170]
White v. West Pub. Corp., 29 F. Supp. 3d 396 (S.D.N.Y. 2014).	Yes, to create a searchable database of copies of legal briefs to "creat[e] an interactive legal research tool."	Yes, copies stored in databases of West and Lexis.	Yes, copies in database serve further purpose of "creating an interactive legal research tool."[171]

169. *HathiTrust*, 755 F.3d at 97, 101, 103.
170. *Field*, 412 F. Supp. 2d at 1118, 1120.
171. White v. W. Pub. Corp., 29 F. Supp. 3d 396, 399 (S.D.N.Y. 2014).

Because fair use is a flexible doctrine, decided on a case-by-case basis, we should refrain from divining hard-and-fast or rigid rules from these decisions.[172] At a minimum, however, these fair use decisions recognize that uses of copyright works—ranging from computer programs to creative works within the core of copyright protection—to develop and deploy a new technology *can* serve a fair use purpose.[173] As summarized in Tables 2 and 3 above, often the technology is deployed in such a way that the public output of any copies used to develop the technology is quite limited (e.g., a snippet or a thumbnail), but even such uses may involve, internal to the technology in a non-public manner, stored copies of the *entire* works by the technology company.[174]

Developing a technology doesn't automatically serve a fair use purpose. Fair use is fact-specific and must be decided on a case-by-case basis. Technologies that do nothing more than *publicly disseminate* or redistribute exact copies of existing works (whether in whole or in part), even in new ways, such as Napster, Aereo, and TVEyes, are unlikely to involve a fair use purpose; such public dissemination of copies is no different from and may substitute for the public distribution or public performance of the works.[175]

Comparing the last column in Tables 2 and 3 with the last column in Table 4 below highlights this critical distinction. Table 4 shows the technology cases in which fair use defenses were rejected,[176] as well as *ABC v. Aereo*, which did not formally

172. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1274, 1277 (2023) ("Because those principles apply across a wide range of copyrightable material, from books to photographs to software, fair use is a 'flexible' concept, and 'its application may well vary depending on context.'").

173. *See, e.g.*, Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1203, 1209 (2021); Authors Guild v. Google, Inc., 804 F.3d 202, 207 (2d Cir. 2015).

174. *See, e.g.*, Kelly v. Arriba Soft Corp., 336 F.3d 811, 815 (9th Cir. 2003).

175. *See infra* note 176.

176. The cases in Table 4 are separated between cases that lack the defendant's own storage of copies in a searchable database for the operation of a technology and then by cases with such a database. Within each category, the cases are first listed in order of highest court and then chronologically. *See* Am. Broad. Cos. v. Aereo, Inc., 573 U.S. 431, 436–37 (2014) (technology service's streaming TV programs online contemporaneous to their broadcast was infringement); Infinity Broad. Corp. v. Kirkwood, 150 F.3d 104, 106 (2d Cir. 1998) (technology service's retransmission of radio broadcasts over the phone was not fair use); A&M Recs., Inc. v. Napster, Inc., 239 F.3d 1004, 1012, 1014–15 (9th Cir. 2001) (music file sharing service's facilitation of unauthorized sharing of music files for purposes of temporary sampling and space-shifting of music files were not fair use); Capitol Recs., LLC v. ReDigi, Inc., 910 F.3d 649, 659–60, 662–64 (2d Cir. 2018) (technology service's facilitation of resale of digital copies of music files from iTunes was not fair use or allowed under first-sale doctrine); United States v. Am. Soc'y of Composers, Authors & Publishers, 599 F. Supp. 2d 415, 423, 426, 433 (S.D.N.Y. 2009) (wireless carrier's use of music for

involve a fair use defense even though *Sony*'s recognition of fair use in home time-shift recordings of broadcast TV was discussed by the district court.[177] As the second column in Table 4 indicates (by the word "No"), the defendant had no purpose of developing a technology to justify the use of the copyrighted works. The top portion of Table 4 lists these cases that lacked use of copyrighted works to develop the technology itself (beyond merely another way to redistribute or disseminate copies of the works), from *Aereo* to *United States v. ASCAP*.[178] Not surprisingly, after finding the respective uses were not transformative in purpose (because users of each technology enjoyed the plaintiffs' works for the same purpose as the plaintiffs offered the works to the public), the courts rejected fair use defenses.[179]

"ringtones and ringback tones" previews were not fair use); Buena Vista Home Ent., Inc. v. Video Pipeline Inc., 342 F.3d 191, 198–99 (3d Cir. 2003) (video service's unauthorized "clip previews" of movies for use by home video retailers was not fair use); Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 175, 180 (2d Cir. 2018) (technology service's provision of create a search-able database of recorded TV and radio broadcasts to find relevant content was not fair use); Hachette Book Grp., Inc. v. Internet Archive, 115 F.4th 163, 174 (2d Cir. 2024) (online library of books); UMG Recordings, Inc. v. MP3.com, Inc., 92 F. Supp. 2d 349, 350, 352 (S.D.N.Y. 2000) (technology service's facilitation of listening to music online through space-shifting based on proof of authorized CD was not fair use); Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 543, 557 (S.D.N.Y. 2013) (searchable database of news clippings).

177. Am. Broad. Cos., Inc. v. Aereo, Inc., 874 F. Supp. 2d 373, 393 (S.D.N.Y. 2012), *aff'd sub nom.*, WNET, Thirteen v. Aereo, Inc., 712 F.3d 676 (2d Cir. 2013), *rev'd sub nom.*, Am. Broad. Cos., Inc. v. Aereo, Inc., 573 U.S. 431 (2014), *on remand*, 2014 WL 5393867 (S.D.N.Y Oct. 13, 2014); *see also* Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121, 133 (2d Cir. 2008) (Cablevision did not engage in infringement or public performance by enabling customers to use its remote DVR for recording of shows); Fox Broad. Co. v. Dish Network LLC, 747 F.3d 1060, 1069 (9th Cir. 2014).

178. *See infra* Table 4; *supra* note 176 (cases from *Infinity Broad.* through *Am. Soc'y of Composers, Authors & Publishers*).

179. *See infra* Table 4; *supra* note 176.

Table 4. Courts' Rejection of Fair Use in Technology Cases.

Case	**Technology development:** Use of Copyrighted Works to Create New Technology?	**Technology usage:** Use of Copyrighted Works in Public Use of Technology?	**Factor 1** Use of Copyrighted Works Had Further Purpose or Different Character?
American Broad. Cos. v. Aereo, Inc., 573 U.S. 431 (2014).	No.	Yes, service enabled users to record TV shows using remote personal antennas and recording offered by online service.	[No court decision on fair use.][180]
Infinity Broad. Corp. v. Kirkwood, 150 F.3d 104 (2d Cir. 1998).	No.	Yes, retransmission of radio broadcasts over telephone.	No, service just "sell[s] access to unaltered radio broadcasts."[181]
A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004 (9th Cir. 2001).	No.	Yes, file-sharing copies online.	No, file sharing of music files "does not transform the copyrighted work."[182]
Capitol Recs., LLC v. ReDigi, Inc., 910 F.3d 649 (2d Cir. 2018).	No.	Yes, service makes copies of music files to facilitate resales of them online, while attempting to ensure deletion of seller's copy.	No, service enables "resale of digital music files, which resales compete with sales of the same recorded music by the rights holder."[183]

180. *Am. Broad. Cos., Inc.*, 573 U.S. at 436, 450.
181. *Infinity Broad. Corp.*, 150 F.3d at 106.
182. A&M Recs., Inc. v. Napster, Inc., 239 F.3d 1004, 1015 (9th Cir. 2001).
183. Capitol Recs., LLC v. ReDigi, Inc., 910 F.3d 649, 652–53, 661 (2d Cir. 2018).

United States v. Am. Soc'y of Composers, Authors & Publishers, 599 F. Supp. 2d 415 (S.D.N.Y. 2009).	No.	Yes, wireless service provider planned to offer previews of ringtones of copyrighted music without license.	No, wireless carrier's preview of ringtones served same purpose.[184]
Buena Vista Home Ent., Inc. v. Video Pipeline Inc., 342 F.3d 191 (3d Cir. 2003).	No.	Yes, service made "clip previews" of Disney movies sold to retail websites selling home videos.	No, clip previews substituted for Disney's movie trailers.[185]
Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169 (2d Cir. 2018).	Yes, to create a searchable database of TV and radio broadcasts.	Yes, user views up to ten minutes of recordings relevant to search topic.	Yes, but only modestly in allowing clients to time shift and to view what "they want at a time and place that is convenient.[186]
Hachette Book Grp., Inc. v. Internet Archive, 115 F.4th 163 (2d Cir. 2024).	Yes, to create a searchable database of online library of books and other works, some of which are copyrighted.	Yes, during pandemic emergency, user given access to digital books "without regard to the corresponding number of physical books in . . . possession."	No, service made "digital copies of the Works and distributes those copies to its users in full, for free."[187]

184. United States v. Am. Soc'y of Composers, Authors & Publishers, 599 F. Supp. 2d 415, 421, 426–27 (S.D.N.Y. 2009).

185. Buena Vista Home Ent., Inc. v. Video Pipeline, Inc., 342 F.3d 191, 194, 202 (3d Cir. 2003).

186. Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 175, 177–78, 188 (2d Cir. 2018).

187. Hachette Book Grp., Inc. v. Internet Archive, 115 F.4th 163, 176, 181, 185 (2d Cir. 2024).

UMG Recordings v. MP3.com, 92 F. Supp. 2d 349 (S.D.N.Y. 2000).	Yes, to create an online service for people to listen to music for which they "prove[d]" they owned a corresponding CD.	Yes, service purchased "thousands of popular CDs in which plaintiffs held the copyrights, and, without authorization, copied their recordings onto its computer servers so as to be able to replay the recordings for its subscribers."	No, service "simply repackages those recordings to facilitate their transmission through another medium."[188]
Associated Press v. Meltwater U.S. Holdings, 931 F. Supp. 2d 537 (S.D.N.Y. 2013).	Yes, to create a searchable database for a news clipping service to provide clients with excerpts of news article.	Yes, user receives 300-word excerpts of news articles relevant to searches, including email feeds.	No, service "copies AP content in order to make money directly from the undiluted use of the copyrighted material" as a substitute of original works.[189]

Table 4's bottom half (below the divider) lists those cases that involved uses of copyrighted works for both technology development and usage. The technology development involved, in most cases, the defendant's creation of a searchable database of copyrighted works as an integral part of the technology's operation.[190] But these decisions rejected the fair use defense, not

188. UMG Recordings v. MP3.com, 92 F. Supp. 2d 349, 350-51 (S.D.N.Y. 2000).

189. Associated Press v. Meltwater U.S. Holdings, 931 F. Supp. 2d 537, 545, 552 (S.D.N.Y. 2013).

190. *See Video Pipeline Inc.*, 342 F.3d at 195, 198 (video service's unauthorized "clip previews" of movies); *Fox News Network*, 883 F.3d at 174–75 (technology service's provision of a searchable database of recorded TV and radio broadcasts to find relevant content was not fair use); *Hachette Book Grp., Inc.*,115 F.4th at 176 (online library of books); *UMG Recordings, Inc.*, 92 F. Supp. 2d at 350–51 (technology service's facilitation of listening to music online through space-shifting based on proof of authorized CD was not fair use); *Associated Press*, 931 F. Supp. 2d at 543–44, 546 (news excerpts). Granted, some critics may

due to the creation of a database to enable a search capability, but instead, due to how the database was used to create *outputs* that merely substituted for the copyrighted works—respectively, movie clip previews (Video Pipeline), ten minutes of recorded TV and radio broadcasts (TVEyes), the entirety of books (Internet Archive), the entirety of songs (MP3.com), and news excerpts (Meltwater).[191] None of these cases involved a transformative purpose in the output of the respective technology; instead, the courts found the outputs were merely substitutional copies of the copyright holder's works.[192] Thus, these cases stand for the basic principle that creating a searchable database of copyrighted works to find relevant information contained in works does not guarantee a fair use when the outputs from the search are so extensive they effectively substitute for the original works. They stand in stark contrast with the far more limited, non-substitutional outputs of the technology in the cases in Tables 2 and 3.

All the cases in Table 4, show that, when the use of a technology involves dissemination or redistribution of copies of copyrighted works that do not have a transformative purpose but, instead, are commercial substitutes for the plaintiffs' works, courts have rejected fair use. Or, as the Second Circuit put it, quoting Judge Leval's influential law review article on fair use, "a use of copyrighted material that 'merely repackages or republishes the original' is unlikely to be deemed a fair use."[193] To the extent the technology or service was new, it was new only in devising a different way of disseminating copies of works.

disagree with the distinction I draw based on the integration of a searchable database in these cases versus the other cases of technologies or services in Table 4 that merely redistribute copies without such databases. But I think the distinction is justified, given the other searchable database fair use cases in Tables 2 and 3.

191. *See supra* Table 4; *supra* note 190.

192. *See supra* Table 4. I classify TVEyes as nontransformative in purpose, despite the Second Circuit's characterization of the purpose as somewhat transformative: the technology has a "transformative purpose of enhancing efficiency: it enables TVEyes's clients to view all of the Fox programming that . . . discussed a particular topic of interest to them" and "it eliminates the clients' need even to view entire programs, because the ten most relevant minutes are presented to them." *Fox News Network*, 883 F.3d at 177–78.

193. *See supra* Table 4; Infinity Broad. Corp. v. Kirkwood, 150 F.3d 104, 108 (quoting Leval, *supra* note 149, at 1111).

By contrast, the technology cases summarized in Tables 2 and 3 above all involved fair use purposes, different from the copyright holder's purpose. All, except for *Sony*, have transformative purposes in either technology development or usage, or both, with only limited use of the works in the outputs when the technology is publicly deployed (e.g., snippet view, thumbnail images, limited amount of declaring code for programmers).[194] These technologies all offered new functionalities (e.g., search or finding information) that did not involve disseminating substitutional copies. They all did something *different*.

Because fair use is fact specific and eschews bright-line rules, it would be foolish—and legal error—to try to devise a formula for determining fair use. That said, it's not surprising that Tables 2 through 4 show a clear pattern: (i) when the defendant's use of copyrighted works had the *further* purpose of developing a new technology that did not result in substitutional copies of the works being offered to the public, the defendant prevailed in establishing fair use.[195] But (ii) when the defendant lacked such a purpose or offered the public merely a technology or service that resulted in substitutional copies of the works being disseminated, the defendant lost.[196]

In short, the latter technology cases in Table 4 involved infringing outputs, whereas the former in Table 2 and 3 did not. Instead, these successful fair use cases involved transformative uses that yielded new technologies and computer programs that offered new functionalities—often ones that simply did not exist before.

One contrary view is worth dissecting. The Authors Guild argued that *Google v. Oracle* should be limited to uses of computer programs, which typically receive a thinner scope of protection due to their functional nature.[197] In rejecting a fair use defense by ROSS Intelligence, Judge Stephanos Bibas took a similar view, distinguishing *Google* as a case involving computer code, not works involving "written words."[198] Judge Bibas also distinguished *Google* as a case where "[t]he copying was *necessary* for

194. *See supra* notes 146–47, 165 (citing technological fair use cases).

195. *See supra* notes 146–47, 165.

196. *See supra* note 190.

197. The Author's Guild, Comment Letter on Artificial Intelligence and Copyright (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-9036 [https://perma.cc/FX5Q-V7TV].

198. Thomson Reuters Enter. Ctr. GMBH v. ROSS Intel., Inc., 765 F. Supp. 3d 382, 398, 401 (D. Del. 2025).

competitors to innovate."[199] As a result, *Google* was "inapt" and Judge Bibas ruled that ROSS Intelligence's use of Westlaw's headnotes for case reports was not transformative, even though the use was to train its AI model so it can learn to identify parts of uncopyrightable judicial opinions that are responsive to human questions, without the AI model generating any outputs containing Westlaw's copyrighted expression.[200]

Limiting *Google v. Oracle* to computer code is mistaken. First, it creates, contrary to the Supreme Court's constant admonition, a bright-line rule for fair use, which is decided on a case-by-case basis.[201] The Supreme Court's fair use jurisprudence counsels against such a rigid approach. Instead, these factors are matters of degree, weighed in relation to each other, as the *Warhol* Court reminded.[202] The type of copyrighted work is already weighed under Factor 2, the nature of the copyrighted work. Thus, to treat works that do not involve computer code as a disqualifying element under Factor 1, as Judge Bibas did, when considering the defendant's purpose of use to create a new technology improperly double counts—or conflates—the nature of the copyrighted work in Factor 2 as a part of the analysis of the defendant's purpose in Factor 1.[203]

It also creates a false dichotomy between computer code and "written words." Both are forms for expression—indeed, Congress included computer programs *within* the definition of literary works.[204] The advent of "vibe coding"—in which humans can write computer programs simply with natural language instructions to AI generators—further shows the false dichotomy between computer code and "written words."[205] And, from the perspective of the Progress Clause and fostering innovation in the United States, the dichotomy between computer code and words makes no sense. The United States has a national interest in fostering the

199. *Id.* at 398.

200. *Id.* at 398–400.

201. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1274 (2023); Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1196–97 (2021); Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577–78 (1994).

202. *See Warhol*, 143 S. Ct. at 1273.

203. *See Thomson Reuters*, 765 F. Supp. 3d at 398; 17 U.S.C. § 107.

204. *See* 17 U.S.C. § 101 (defining of "literary works" to include "works, other than audiovisual works, expressed in words, numbers, or other verbal or numerical symbols or indicia").

205. *See* Shalini Harkar, *What Is Vibe Coding?*, IBM (Apr. 8, 2025), https://www.ibm.com/think/topics/vibe-coding [https://perma.cc/S9V7-CLFG] ("Vibe coding is a fresh take in coding where users express their intention using plain speech and the AI transforms that thinking into executable code.").

development of new technologies regardless of the type of work copied, provided the balance of factors ultimately favors fair use.

Taking a restrictive view of Factor 1—and limiting technological development as a legitimate purpose of use only for computer programs, as Judge Bibas did in *Thomson Reuters*—will thwart innovation in the United States. Indeed, none of the fair uses listed in Tables 2 and 3 above to create new technologies related to search (of websites, images, and books), plagiarism detection, and Westlaw's own research tool (including legal briefs within its database) would be allowed. Judge Bibas failed to cite, much less discuss, most of these other technological fair use decisions, which all involved copyrighted works beyond computer programs.[206] They are far more consistent with the flexible approach to fair use the Supreme Court has adopted. Moreover, Judge Bibas gave insufficient weight to the expected output of ROSS Intelligence's legal research tool: mere quotes from uncopyrightable judicial opinions, *not* any copyrightable expression of Westlaw.[207]

That left Judge Bibas to devise a market harm under Factor 4 based not on the harm to the "market for or value of the copyrighted work,"[208] but instead the more general harm to the broad market for legal-research platforms.[209] But if legal-research platform here refers to Westlaw's entire research service people use, such platform is a functional system or method of operation that is not copyrightable.[210] Although Judge Bibas also posited a potential derivative market in "data to train legal AIs," he provided no analysis on why that's a cognizable market for copyright holders to control.[211] Two other federal judges, in subsequent fair use decisions involving AI training, rejected the very argument that the market for data for AI training should be

206. *See Thomson Reuters*, 765 F. Supp. 3d at, 398 (failing to discuss technological fair use decisions beyond *Oracle*, *Sega*, and *Connectix*). Judge Bibas did rely on *Authors Guild* in analyzing Factors 2 and 3 but not Factors 1 and 4; he did not discuss the technology involved in that case, Google Book Search, or why the decision was not helpful in analyzing Factors 1 and 4. *Id.* at 397–400. A number of decisions have discussed technologies that copied works beyond computer programs. *See, e.g.*, Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 419–20, 449–50 (1984); Williams & Wilkins Co. v. United States, 487 F.2d 1345, 1348, 1362 (Ct. Cl. 1973); *see also Campbell*, 510 U.S. at 571–73, 594.

207. *See* Thomson Reuters Enter. Ctr. GMBH v. ROSS Intel., Inc., 694 F. Supp. 3d 467, 483 (D. Del. 2023), *vacated in part*, 765 F. Supp. 3d 382, 398 (D. Del. 2025).

208. 17 U.S.C. § 107.

209. *Thomson Reuters*, 765 F. Supp. 3d. at 400.

210. *See* 117 U.S.C. § 102(b); Baker v. Selden, 101 U.S. 99, 102 (1879) (noting that systems cannot be protected by copyright, but instead protection "can only be secured by a patent").

211. *Thomson Reuters*, 765 F. Supp 3d. at 400.

controlled by copyright holders, with one judge noting the inherent circularity problem in deeming every new licensing opportunity a market that copyright holders are entitled to control.[212] Although ROSS's legal research tool could compete with Westlaw, competition among businesses is the norm in a free market. Indeed, "[t]he heart of our national economic policy long has been faith in the value of competition."[213]

2. Fostering Technology Through the Sony *Safe Harbor*. The *Sony* safe harbor—which shields the development of technologies with substantial non-infringing uses from secondary liability—complements this technology-accommodating role of fair use.[214] One way of thinking about this complementary role: fair use facilitates especially the development of new technologies, while the *Sony* safe harbor ensures that people's use of these new technologies does not lead to unwarranted or excessive liability imposed on the technology developer. Fair use and the *Sony* safe harbor are flip sides of the same coin. Both accommodate innovation.

In *Sony*, the Court analogized the movie studios' claims of copyright liability based on Sony's sale of the VCR to a patentee's broad assertion of patent "to extend his monopoly beyond the limits of his specific grant."[215] A copyright was not meant to give the holder a right to stop the development of a new technology. Accordingly, the Court devised the *Sony* safe harbor from the staple article of commerce doctrine codified in the Patent Code, which shields from secondary liability "a staple article or commodity of commerce suitable for substantial non-infringing

212. Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *17 (N.D. Cal. June 23, 2025) ("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."); Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *16 (N.D. Cal. June 25, 2025) ("Therefore, to prevent the fourth factor analysis from becoming circular and favoring the rightsholder in every case, harm from the loss of fees paid to license a work for a transformative purpose is not cognizable.").

213. Nat'l Soc'y of Pro. Eng'rs. v. United States, 435 U.S. 679, 695 (1978) (quoting Standard Oil Co. v. FTC, 340 U.S. 231, 248 (1951)); *see* Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 607–08 (9th Cir. 2000) ("For this reason, some economic loss by Sony as a result of this competition does not compel a finding of no fair use. Sony understandably seeks control over the market for devices that play games Sony produces or licenses. The copyright law, however, does not confer such a monopoly."); *Bartz*, 2025 WL 1741691, at *17 ("This [increase in non-infringing works generated by AI] is not the kind of competitive or creative displacement that concerns the Copyright Act. The Act seeks to advance original works of authorship, not to protect authors against competition.").

214. *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 436–37 (1984).

215. *Id.* at 441–42.

use."[216] The reason was compelling: there is a "public interest in access to that article of commerce."[217] "For that reason, in contributory infringement cases arising under the patent laws the Court has always recognized the critical importance of not allowing the patentee to extend his monopoly beyond the limits of his specific grant."[218] Patent law sets a very high bar for secondary liability based on distribution of articles of commerce: "These cases deny the patentee any right to control the distribution of unpatented articles unless they are 'unsuited for *any* commercial non-infringing use.'"[219]

The *Sony* Court held that one of the unauthorized uses of the VCR—consumers' recordings of broadcast shows to watch later—was a permissible fair use, or a substantial non-infringing use.[220] In *Grokster*, the Court described the *Sony* safe harbor as a doctrine that "leaves breathing room for innovation and a vigorous commerce."[221] The *Sony* safe harbor helps navigate the "tradeoff" copyright law confronts: "The more artistic protection is favored, the more technological innovation may be discouraged . . . ."[222]

So does fair use. As the *Google* Court explained:

> [Fair use] can focus on the legitimate need to provide incentives to produce copyrighted material while examining the extent to which yet further protection *creates unrelated or illegitimate harms in other markets* or *to the development of other products*. In a word, it can carry out its basic purpose of providing a context-based check that can help to keep a copyright monopoly within its lawful bounds.[223]

The Court held that Google's unlicensed use of Oracle's Java code served a fair use purpose "to create new products."[224] And the purpose of the new product (Android) was different from the copyright holder's: "Google's basic purpose was to create a *different* task-related system for a *different* computing environment (smartphones) and to create a platform—the Android platform—

216. *Id.* at 440–41 (citing 35 U.S.C. § 271(c)).

217. *Id.* at 440.

218. *Id.* at 441.

219. *Id.* (citing Dawson Chem. Co. v. Rohm & Haas Co., 448 U.S. 176, 198 (1980)) (emphasis added).

220. *Id.* at 442.

221. Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd., 545 U.S. 913, 932–33 (2005).

222. *Id.* at 928.

223. Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1198 (2021) (emphasis added).

224. *See id.* at 1203.

that would help achieve and popularize that objective."[225] In short, Google's purpose of use of Oracle's copyrighted work was to create "a new and transformative program."[226]

3. Fostering Technologies Promotes Progress in the United States. The *Google* Court explained the importance of technological innovation to the Progress Clause: [Google's] use was consistent with that creative "progress" that is the basic constitutional objective of copyright itself.[227] Developing a new technology was not just a fair use purpose as a matter of law, the purpose served the constitutional goal of the Progress Clause—by facilitating the creation of a new computing platform, which in turn spurred the creation of new applications.[228] Society benefits.[229]

One must not forget that the Progress Clause seeks to promote *national* progress. That explains why fair use cases involving new technologies often implicate much broader, national concerns beyond the parties' interests. As the *Sony* Court explained:

> The 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. When technological change has rendered its literal terms ambiguous, the Copyright Act must be construed in light of this basic purpose.[230]

A typical copyright lawsuit is not likely to affect the national economy and innovation. But a copyright lawsuit against an entire class of new technology might. As the Supreme Court concluded in *Sony*, "[s]uch a rule would block the wheels of commerce."[231]

Promoting progress is also potentially international in scope. Although not a fair use decision, *Eldred v. Ashcroft* recognizes that the Copyright Act may permissibly take into account how U.S. copyright law operates alongside the copyright laws and

225. *Id.* at 1205 (emphasis added).

226. *Id.* at 1209.

227. *Id.* at 1203 (citing Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 349–50 ("The primary objective of copyright is not to reward the labor of authors, but '[t]o promote the Progress of Science and useful Arts'" (quoting U.S. CONST., art. I, § 8, cl. 8))).

228. *See id.*

229. *See id.* at 1206 (discussing evidence of a public benefit that a jury heard).

230. *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 432 (1984) (emphasis added) (citations omitted).

231. *Id.* at 441 (citation omitted).

protections offered by other countries.[232] For example, in the *Eldred*, the Supreme Court concluded that Congress had a rational basis in extending all existing and future copyright terms by twenty years in order to avoid U.S. authors being disadvantaged in foreign countries that applied the rule of the shorter term.[233] Quoting an article written by Shira Perlmutter, who later served as the Register of Copyrights, the Court noted: "[M]atching th[e] level of [copyright] protection in the United States [to that in the EU] can ensure stronger protection for U.S. works abroad and avoid competitive disadvantages vis-à-vis foreign rightholders."[234] Considering these international dimensions of copyright laws enabled the United States to "'play a leadership role' in the give-and-take evolution of the international copyright system."[235] Thus, because the Trump Administration declared a national priority to maintain the U.S. as the global leader in AI,[236] courts should be mindful of that stated national interest when rendering a decision that could jeopardize the United States' position in the world.

C. Importance of Fair Use to the Information Economy

Before turning to AI, one final point is worth emphasizing. We live today in the information economy: information itself is harnessed in new and better ways to drive productivity.[237] Indeed, the internet itself is an example of how communications technologies can transform work and society, becoming an arguably essential part of modern daily life.[238] The internet and

232. Eldred v. Ashcroft, 537 U.S. 186, 205–06 (2003).

233. *See id.*

234. *Id.* at 206 (alterations in original) (quoting Shira Perlmutter, *Participation in the International Copyright System as a Means to Promote the Progress of Science and Useful Arts*, 36 LOY. L.A. L. REV. 323, 330 (2002)).

235. *Id.* at 206 (quoting Perlmutter, *supra* note 234, at 332).

236. *See* Exec. Order No. 14,179, *supra* note 9.

237. *See* YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM 3 (2006) ("What characterizes the networked information economy is that decentralized individual action—specifically, new and important cooperative and coordinate action carried out through radically distributed, nonmarket mechanisms that do not depend on proprietary strategies—plays a much greater role than it did, or could have, in the industrial information economy."); CARL SHAPIRO & HAL R. VARIAN, INFORMATION RULES: A STRATEGIC GUIDE TO THE NETWORK ECONOMY 9 (1999) ("Improved information infrastructure has vastly increased our ability to store, retrieve, sort, filter, and distribute information, thereby greatly enhancing the value of the underlying information itself.").

238. *See, e.g.*, Yajie Liu et al., *Does the Development of the Internet Improve the Allocative Efficiency of Production Factors? Evidence from Surveys of Chinese Manufacturing Firms*, 66 STRUCTURAL CHANGE AND ECON. DYNAMICS 161, 162 (2023).

other digital technologies enabled the growth and collection of "big data."[239] And now, with that big data, developers have been able to develop and construct impressive AI models that would have seemed science fiction only a decade ago.

In this information economy, where new technologies attempt to harness big data, fair use serves an important role. Because much of big data collected is often copyrighted,[240] fair use can provide a way for courts to balance the competing interests of protecting copyright holders and fostering innovation in new technologies that offer society new ways of doing things.

IV. USE-BY-USE ANALYSIS OF USES OF WORKS IN AI TRAINING AND USAGE

Part IV discusses how courts should evaluate the issues related to uses of copyrighted works in the AI litigation involving a fair use defense. Courts should first identify, as *Warhol* instructs, the "specific 'use[s]'" involved and then analyze them use-by-use.[241] Although the types of AI models and generators in the litigation vary, we can divide the uses of copyrighted works into two general phases: (i) *technology development*: a research and development phase involving the training of an AI model in creating and developing the model so it works; and (ii) *technology usage*: a public use phase during which an AI generator is made available for public use during which the AI generates outputs at the direction of users' prompts, some of which outputs may be alleged to be infringing.[242] Each use should be analyzed individually.

239. *See* Myron P. Gutmann, Emily Klancher Merchant, & Evan Roberts, *"Big Data" in Economic History*, 78. J. ECON. HIST. 268, 269–70 (2018) (describing the history of the growth and collection of big data).

240. *See Fair Use and the Training of AI Models on Copyrighted Works*, BITLAW, https://www.bitlaw.com/ai/AI-training-fair-use.html [https://perma.cc/7U22-SG7L] (last visited July 19, 2025) ("Because the vast majority of text-based knowledge created in the last century—including books, journalism, source code, and artistic expression—is protected by copyright law, these AI systems inevitably encounter the challenge of training on copyrighted content. In fact, the scale of data required for effective AI training arguably makes the use of copyrighted works nearly unavoidable.").

241. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1277 (2023).

242. Granted, even after an AI generator is launched, it may continue to be refined and developed. *See Pricing*, OPENAI: PLATFORM, https://platform.openai.com/docs/pricing [https://perma.cc/T94Z-FBSE] (last visited July 15, 2025) (listing latest models available).

A. The Origin of Training AI Models at Universities

The lawsuits against AI companies that allege copyright infringement typically focus on the training of the respective AI models with datasets containing millions, if not billions, of works, a large percentage of which is copyrighted.[243] A significant body of legal scholarship has analyzed whether this AI training is a fair use.[244] Legal scholars divide.[245] My goal is to add to this debate—with the goal of assisting courts that confront this issue—by discussing points that the existing literature largely ignores.[246]

1. Technology Development: The Further Purpose and Different Character of Using Works in AI Training. I start with an examination of the history and practice of using copyrighted works in AI training in AI research at U.S. universities. As explained below, this examination provides important context for understanding the specific purpose of using copyrighted works in AI training.

a. The Importance of Examining Whether AI Training at Universities Serves a Fair Use Purpose—or Not. Some commentators may wonder why I am writing about university-based AI research when no university has been sued in the AI litigation. But I believe it is important to examine the question of fair use in AI

243. *See, e.g.*, New York Times' Complaint, *supra* note 32, at 6.

244. *Compare* Matthew Sag, *Copyright Safety for Generative AI*, 61 HOU. L. REV. 295, 304–05 (2023) (arguing that AI training involves "nonexpressive use" of work: "the rationale for allowing for-profit and academic researchers to derive valuable data from other people's copyrighted works is a necessary implication of the fundamental distinction between protectable original expression and unprotectable facts, ideas, abstractions, and functional elements"); Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 770–79 (2021) (arguing that AI training involves the model's identification of uncopyrightable elements in training works and should be recognized as "fair learning," weighing in favor of fair use); *with* Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. L. REV. J.L. & ARTS 45, 49–50 (2017) (casting doubt on fair use in AI training because "certain applications of machine learning challenge the doctrine of non-expressive use by recasting the analysis of the two most important factors of fair use: the purpose of the use, and its effect on the market for the works used").

245. *See supra* note 244.

246. A notable exception is Deven R. Desai & Mark Riedl, *Between Copyright and Computer Science: The Law and Ethics of Generative AI*, 22 NW. J. TECH. & INTELL. PROP. 55, 72–76 (2024) (arguing the goals of academic research and commercial research in AI are different: "[T]he model [developed in academic research] need not be deployed at scale with millions of users paying to use it").

training for research and development of AI models by researchers at not-for-profit universities. Why? For three important reasons.

First, to understand why this practice of AI training occurs in the first place, we must understand the history of AI research and development—and the many failures in AI research. AI training did not start at for-profit tech companies. It started at universities, predominantly in the United States and Canada, to test the basic idea that AI could be developed by trying to mimic the human brain, an idea dating back to the 1950s.[247] The practice of training AI models on larger and larger datasets—a method referred to as "scaling up" or "scaling"—originated at universities, not tech companies, and, at some point, the datasets included copyrighted works that were not licensed for training by the copyright owners.[248] If this practice is blatant theft of copyrighted works, then this "original sin" occurred in academia, not Silicon Valley.

Second, examining whether AI training using copyrighted works at universities has a legitimate fair use purpose provides an important point of comparison that will sharpen the courts' fair use analysis. In *Warhol* and other cases, the Supreme Court employed this comparative method in analyzing fair use by considering other examples beyond the defendant's case—for example, the majority opinion analyzed the purpose of Warhol's famous *Soup Cans* artwork as a way to help the Court to "draw[]" "distinction[s]" with the particular work at issue.[249] This same comparative approach to fair use will sharpen courts' analysis of fair use in AI training.

Third, Factor 4 of fair use requires courts to consider the potential market harm on the copyright holder's work "not only the extent of market harm caused by the particular actions of the alleged infringer, but also '*whether unrestricted and widespread*

247. For a historical account of AI's development, see CADE METZ, GENIUS MAKERS: THE MAVERICKS WHO BROUGHT AI TO GOOGLE, FACEBOOK, AND THE WORLD 3–4, 15–22 (2021) (describing early days of AI research and attempt to mimic human brain).

248. *See infra* notes 267–68, 275–81, 357–60, 370–72, 376–80 and accompanying text; Peter Henderson et al., *Foundation Models and Fair Use*, J. MACH. LEARNING & RSCH., Sep. 2023, at 1, 1–2, 5–6, 8–9 ("Researchers, at least in the United States, have long relied on the legal doctrine of *fair use* to avoid liability from using copyrighted data."); Timothy B. Lee & James Grimmelmann, *Why the New York Times Might Win its Copyright Lawsuit Against OpenAI*, ARS TECHNICA (Feb. 20, 2024, 08:05 CT), https://arstechnica.com/tech-policy/2024/02/why-the-new-york-times-might-win-its-copyright-lawsuit-against-openai/ [https://perma.cc/T5CD-6SVG] ("A lot of early AI research was done in an academic setting; the law specifically mentions teaching, scholarship, and research as examples of fair use. As a result, the machine-learning community has traditionally taken a relaxed attitude toward copyright. Early training sets frequently included copyrighted material.").

249. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1281 (2023).

conduct of the sort engaged in by the defendant . . . would result in a substantially adverse impact on the potential market' for the original."[250] The widespread conduct of using copyrighted works in datasets for AI training already occurs at both universities and industry.[251] Thus, courts do not even have to speculate in conducting this required examination of the widespread conduct of using large datasets consisting of copyright works. At the same time, in a similar case involving the use of a copyrighted work to develop a new technology, the *Google* decision also took "into account the public benefits the copying will likely produce" when analyzing the potential market harm under Factor 4.[252] Accordingly, courts should compare the public benefits with the "dollar amounts likely lost" by the copyright holders.[253] Once we understand that AI training to develop new AI models at universities is part of the "*widespread conduct of the sort engaged in by the defendant*," both the market harm and the public benefits of such AI training come into sharper relief.[254]

Legal scholars on both sides of the debate have all but ignored this vital question of AI training with copyrighted works at universities.[255] Part of the reason may be that none of the

250. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 590 (1994) (quoting 3 M. NIMMER & D. NIMMER, NIMMER ON COPYRIGHT § 13.05[A][4], 13–102.61 (1993)) (emphasis added).

251. *See* Lee & Grimmelmann, *supra* note 248.

252. Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1191, 1206 (2021).

253. *Id.* at 1206.

254. *Campbell*, 510 U.S. at 590 (quoting 3 M. NIMMER & D. NIMMER, NIMMER ON COPYRIGHT § 13.05[A][4], 13–102.61 (1993) (emphasis added). In recounting the history of this practice, I do not assert it should be recognized as a custom. *See, e.g.*, Wall Data Inc. v. L.A. Cnty Sheriff's Dep't., 447 F.3d 769, 778 (9th Cir. 2006) ("[F]air use is appropriate where a 'reasonable copyright owner' would have consented to the use, i.e., where the 'custom or public policy' at the time would have defined the use as reasonable."); *see also* Lloyd L. Weinreb, *Fair's Fair: A Comment on the Fair Use Doctrine*, 103 HARV. L. REV. 1137, 1140 (1990) (discussing the role of fairness and customary practice in the fair use analysis). The Supreme Court has not discussed custom as a relevant consideration in the fair use analysis. I do not do so here. *See also* Jennifer E. Rothman, *Why Custom Cannot Save Copyright's Fair Use Defense*, 93 VA. L. REV. IN BRIEF 243, 244–45 (2008), https://virginialawreview.org/articles/why-custom-cannot-save-copyrights-fair-use-defense/ [https://perma.cc/NFQ2-Y3PF] (rejecting reliance on "custom" as a normative indicator of fair use). Instead, what is important is identifying the *purpose* of this practice, why it originated and spread in AI research from universities to AI companies. Understanding that common purpose is critical.

255. *Compare* Lemley & Casey, *supra* note 244, at 745–46 (concluding that AI training on copyrighted works may qualify as fair use when conducted "for a commercial purpose"); Elizabeth Spica, *Public Interest, the True Soul: Copyright's Fair Use Doctrine and the Use of Copyrighted Works to Train Generative AI Tools*, 33 TEX. INTELL. PROP. L.J. 67, 80–81 (2024) (finding that ingestion of copyrighted works to train AI tools is transformative fair use, though analyzing only commercial applications); *and* Art Neill, James Thomas & Erika Lee, *A Framework for Applying Copyright Law to the Training of Textual Generative Artificial Intelligence*, 32 TEX. INTELL. PROP. L.J. 225, 236–37 (arguing that commercial use

copyright lawsuits against AI involves a university, although some of the researchers at some of the defendant companies, such as Yann LeCun who works for Meta and New York University, are affiliated with universities.[256] Another reason may be that some people may simply assume that AI training at universities either is a fair use or at least serves a fair use purpose of research at a not-for-profit educational institution.[257] Even the Authors Guild, in its comment submitted to the Copyright Office's study, acknowledged that AI training for noncommercial purposes might provide support for a defendant's argument of a fair use under Factor 1.[258] Likewise, the lawyer for the *New York Times* in its lawsuit against OpenAI and Microsoft took a similar view, suggesting that "training models for research purposes or training models in a university," such as when OpenAI "present[ed] itself as being this nonprofit benefit entity developing artificial intelligence for the benefit of humanity" might weigh differently under Factor 1 than a commercial use.[259] But others doubt that even noncommercial AI training is a fair use.[260] And the fear of being sued for copyright infringement may deter AI researchers at universities from using copyrighted works in research.[261]

Examination of fair use for university AI research helps to sharpen the analysis. For, if such noncommercial AI training is not

---

of copyrighted works to train OpenAI's models is likely protected by fair use); *with* Celeste Shen, Note, *Fair Use, Licensing, and Authors' Rights in the Age of Generative AI*, 22 NW. J. TECH. & INTELL. PROP. 157, 166–69 (2024) (arguing that commercial AI training is not transformative, causes market harm, and should not qualify as fair use); *and* Charlesworth, *supra* note 48, at 359 (critiquing broad fair use defenses in the commercial AI training context and rejecting expansive applications). *But see* Desai & Riedl, *supra* note 246, at 72–74 (discussing AI training at universities).

256. Yann LeCun (@yann-lecun), LINKEDIN, https://www.linkedin.com/in/yann-lecun [https://perma.cc/K6G9-JJZC]; *see* Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *1036 (N.D. Cal. June 25,2025) (copyright suit against Meta for its use of books to train AI model).

257. *See, e.g.*, Daniel Jeffries (@Dan_Jeffries1), X (Dec. 28, 2023, at 03:27 CT), https://x.com/Dan_Jeffries1/status/1740303405254377808 [https://perma.cc/37US-GC5K].

258. *See* The Author's Guild, *supra* note 197.

259. *See* Transcript for Oral Argument on Motion to Dismiss, at 41–42, N.Y. Times Co. v. Microsoft Corp., No. 23-cv-11195 (SHS), 2025 WL 1009179 (S.D.N.Y. Apr. 4, 2025).

260. *See* Am. Ass'n of Indep. Music & Recording Indus. Ass'n of Am., Inc., Comment Letter on Artificial Intelligence and Copyright (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-8833 [https://perma.cc/3ZNK-DPC9].

261. *See Researchers Say AI Copyright Cases Could Have Negative Impact on Academic Research*, GA. TECH. (Nov. 21, 2024), https://research.gatech.edu/researchers-say-ai-copyright-cases-could-have-negative-impact-academic-research [https://perma.cc/3423-RW9Q]. A related challenge is that some datasets may be decreasing in size due to the objection of websites and copyright holders. *See* Kevin Roose, *The Data That Powers A.I. Is Disappearing Fast*, N.Y. TIMES (July 19, 2024), https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html [https://perma.cc/H72A-TC35].

a valid fair use purpose, then, *a fortiori*, the commercial AI training by for-profit companies is not valid, either. In such case, the university researchers would all be committing massive copyright infringement, and their universities would likely be liable for contributory infringement by providing the "site and facilities" where such infringement occurred, with at least willful blindness to such activity.[262] But, if such university-based AI training has a valid fair purpose, then it provides a reason to reject broad arguments that AI training cannot serve a fair use purpose because it putatively has no different purpose than the copyright holders' purpose. In either case, courts would gain much better insight to what makes AI training serve—or not serve—a legitimate fair use purpose.

Any analysis of the purpose and character of using copyrighted works to train AI models in the AI litigation should address this question. The answer is vital to AI research—and the future of how AI research is conducted in the United States, both at companies and universities.

b. The History of AI Training at Universities and the Essential Role of Scaling the Size of Datasets. AI didn't develop overnight. AI development started in the 1950s, spanning nearly seventy-five years.[263] This research was beset with many years during which funding for AI researchers dried up due to deep skepticism of the entire project.[264] These periods were so dire they are called "AI winter[s]."[265]

The school of AI research that drew the most skepticism was the one that, surprisingly, has sparked the emergence of generative AI today: the so-called connectionist school that espouses the goal of trying to develop AI through artificial neural

262. *See* Fonovisa, Inc. v. Cherry Auction, Inc., 76 F.3d 259, 264–65 (9th Cir. 1996) ("[P]roviding the site and facilities for known infringing activity is sufficient to establish contributory liability.").

263. *See* METZ, *supra* note 247, at 3; *A Brief History of Artificial Intelligence*, NAT'L INST. OF JUST. (Sep. 30, 2018), https://nij.ojp.gov/topics/articles/brief-history-artificial-intelligence?utm_source [https://perma.cc/NEP3-MNV8].

264. *See* METZ, *supra* note 247, at 3; *AI Winter: The Highs and Lows of Artificial Intelligence*, HIST. OF DATA SCI. (Sep. 1, 2021), https://www.historyofdatascience.com/ai-winter-the-highs-and-lows-of-artificial-intelligence/ [https://perma.cc/DN5U-2YCR].

265. *See* METZ, *supra* note 247, at 34; *2 AI Winters and 1 Hot AI Summer*, ENTEFY (July 31, 2023), https://www.entefy.com/blog/2-ai-winters-and-1-hot-ai-summer/ [https://perma.cc/YGB7-WUWB].

networks that mimic the human brain.[266] As late as the 1990s, this connectionist school drew great skepticism, if not outright scorn. As Cade Metz recounts:

> The idea of a neural network dated back to the 1950s, but the early pioneers had never gotten it working as well as they had hoped. By the new millennium, most researchers had given up on the idea, convinced it was a technological dead end . . . . When submitting research papers to academic journals, those who still explored the technology would often disguise it as something else, replacing the words "neural network" with language less likely to offend their fellow scientists.[267]

The primary method of developing these neural networks is to expose the AI models to vast amounts of data, typically from human-produced content, which may include both copyrighted and public domain works.[268] During this process of deep learning, an AI model analyzes or dissects the large dataset into many discrete, tiny elements (or "tokens" for large language models) to identify the patterns and relationships among the elements in the entire dataset, assigning weights or numerical values representing statistical representations of the relationships among the many elements.[269] This deep learning by AI is unsupervised by humans—the model is left to its own devices, so to speak, to figure out the patterns and relationships in the data.[270]

As Yann LeCun, Yoshua Bengio, and Geoffrey Hinton described the general process of deep learning in *Nature* in 2015:

> Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection and

266. *See* Shannon Flynn, *The Difference Between Symbolic AI and Connectionist AI,* RE-WORK (Sep. 24, 2020), https://blog.re-work.co/the-difference-between-symbolic-ai-and-connectionist-ai/ [https://perma.cc/X9GM-VT35].

267. METZ, *supra* note 247, at 3.

268. *See* Flynn, *supra* note 266; Adam Buick, *Copyright and AI Training Data—Transparency to the Rescue?*, 20 J. INTELL. PROP. L. & PRAC. 182, 183–84 (2025) ("These models require immense quantities of data, with the largest training datasets comprising millions of text documents, images, audio samples, or other forms of content. Most of this material is protected by copyright . . . ."(footnote omitted)).

269. *See* Ruhma Khawaja, *Demystifying Embeddings 101: The Foundation of Large Language Models*, DATA SCI. DOJO (Aug. 17, 2023), https://datasciencedojo.com/blog/embeddings-and-llm/ [https://perma.cc/YF2P-P4KK]; Jon Evans, *Attention Is All You Need to Understand*, GRADIENT ASCENDANT (Apr. 19, 2023), https://aiascendant.substack.com/p/attention-is-all-you-need-to-understand [https://perma.cc/2BT5-EB7E].

270. *What Is Unsupervised Learning?*, GOOGLE CLOUD, https://cloud.google.com/discover/what-is-unsupervised-learning [https://perma.cc/XQ9Q-6HQ2].

> classification. . . . The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure.
>
> Deep learning is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has turned out to be very good at discovering intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government.[271]

Today, this idea of AI training may already seem banal. But it was a pioneering, indeed radical (to some, probably crazy), idea that was contrary to the prevailing school of AI research, the symbolic school, which took a different approach in having humans write logic-based principles that were supposed to govern the AI model.[272] Part of the reason for the first AI winter in the 1970s was a skeptical critique of neural network research at the time by Marvin Minsky and Seymour Papert, who suggested that it may be a "sterile" path of research. For decades, AI research plodded along with no clear approach as the one likely to yield a successful artificial intelligence.[273] It would be a profound mistake to underestimate or ignore the significance of the discovery of AI deep learning through large datasets. In 2024, Hinton was awarded the Nobel Prize in physics for his insight in developing "a method that can autonomously find properties in data, and so perform tasks such as identifying specific elements in pictures."[274]

AI research began to advance in the early 2000s when researchers had greater access to content or data available online.[275] These advances were driven by research from the

---

271. Yann LeCun, Yoshua Bengio & Geoffery Hinton, *Deep Learning*, 521 NATURE 436, 436 (2015), https://hal.science/hal-04206682/document [https://perma.cc/EJQ2-FL3U].

272. *See* Flynn, *supra* note 266.

273. *See* MARVIN MINKSY & SEYMOUR PAPERT, PERCEPTRONS: AN INTRODUCTION TO COMPUTATIONAL GEOMETRY 232–33 (1969) (attacking single-layer perceptrons and research of neural networks at the time); *see also* METZ, *supra* note 247, at 34 (Minsky and Papert book "pushed most researchers away from connectionism"); Will Knight, *What Marvin Minsky Still Means for AI*, MIT TECH. REV. (Jan. 26, 2016), https://www.technologyreview.com/2016/01/26/163622/what-marvin-minsky-still-means-for-ai/ [https://perma.cc/MDC9-RGHB] ("The book has been blamed for directing research away from this area of research for many years."); Istvan S. N. Berkeley, *The Curious Case of Connectionism*, 2 OPEN PHIL. 190, 193–94 (2019) ("This effectively put the bulk of connectionist research (with a few notable exceptions) into a state of hibernation until the mid-1980s when the second phase of connectionism took off.").

274. *The Nobel Prize in Physics 2024: Press Release*, THE NOBEL PRIZE (Oct. 8, 2024), https://www.nobelprize.org/prizes/physics/2024/press-release/ [https://perma.cc/VN2T-RUJG].

275. *See* METZ, *supra* note 247, at 66–79.

connectionist approach, using data to train AI models by letting the models "learn" on their own. In 2006, in a seminal paper by Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh, which has been cited 23,198 times, the AI researchers used the Modified National Institute of Standards and Technology (MNIST) dataset of hand-written digits in "60,000 training images and 10,000 test images" to successfully create a generative AI model.[276] Because the MNIST created and shared the dataset for research, it did not raise a copyright issue.[277] But the researchers' training AI models with large datasets validated this method of deep learning, spurring researchers to use this method of AI training with large datasets.[278]

For example, in 2009, Rajat Raina, Anand Madhavan, and Andrew Ng published an important paper recognizing how graphical processors (GPUs) can enhance unsupervised learning.[279] They advised: "[S]caling up existing DBN [deep belief network] and sparse coding models to use more parameters, or *more training data*, *might produce very significant performance benefits*."[280] In 2012, Ng, while working at Stanford and for Google Brain, teamed up with Google researchers to show that this "scaling up"—meaning increasing the size of—"the core components involved in training deep networks: the dataset, the model, and the computational resources" worked effectively in unsupervised learning by a computer vision model from unlabeled

276. *See* Geoffrey E. Hinton, Simon Osindero & Yee-Whye Teh, *A Fast Learning Algorithm for Deep Belief Nets*, 18 NEURAL COMPUTATION 1527, 1535–36, 1540, 1546 (2006); (cited 23,198 times according to Google Scholar, as of Sep. 16, 2025).

277. *See MNIST Database*, WIKIPEDIA, (Aug. 26, 2025, at 20:53 UTC) https://en.wikipedia.org/wiki/MNIST_database [https://perma.cc/3DG5-U35U].

278. *See* LeCun, Bengio & Hinton, *supra* note 271, at 440 ("When deep convolutional networks were applied to a data set of about a million images from the web that contained 1,000 different classes, they achieved spectacular results, almost halving the error rates of the best competing approaches."). This is not to rule out the possibility that AI researchers won't reach a point of diminishing returns with increasingly massive datasets. *See New Study Finds Bigger Datasets Might Not Always Be Better for AI Models*, TECHXPLORE (Nov. 13, 2023), https://techxplore.com/news/2023-11-bigger-datasets-ai.html [https://perma.cc/U3XA-8DTE]. AI researchers may also figure out shortcuts or alternatives, such as "distillation" from other AI generators. *See* Miles Kruppa & Deepa Seetharaman, *Why 'Distillation' Has Become the Scariest Word for AI Companies*, WALL ST. J. (Jan. 30, 2025, at 08:00 ET), https://www.wsj.com/tech/ai/why-distillation-has-become-the-scariest-wordfor-ai-companies-aa146ae3?gaa_at=eafs&gaa_n=ASWzDAgBfkv8nEXofDDA-nPve7Aov_NlTAwUXOl0D8d7-xCHYKHNLwaDH9XfBJ8opLo%3D&gaa_ts=68b20a3e&gaa_sig=8S4qnCIYSOk5Pw4GE3tXI [https://perma.cc/XU3B-25XV].

279. *See* Rajat Raina, Anand Madhavan & Andrew Ng, *Large-Scale Deep Unsupervised Learning Using Graphics Processor*, *in* PROCEEDINGS OF THE TWENTY-SIX INT'L CONF. MACH. LEARNING 873, 875 (2009).

280. *Id.* at 873 (emphasis added).

data consisting of an image copied from each of 10 million YouTube videos.[281]

These successes in AI research show the size of the training datasets matters. As datasets got larger, the AI models improved. As LeCun, Bengio, and Hinton correctly predicted in 2015: "We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of *increases in the amount of available computation and data*."[282]

Training data thus became crucial for AI research and development. To advance AI development, researchers compiled and shared online datasets for AI training. Today, various online repositories, such as Kaggle, Google Dataset Search, Harvard Dataverse, and UCI Machine Learning Repository, facilitate the sharing of datasets.[283] Likewise, some nonprofits have been established specifically for the purpose of serving as a repository for datasets, although the copyright lawsuits against AI companies have raised questions even about some of these nonprofits.[284] One dataset tracker now lists 13,612 datasets for AI training.[285]

It is unclear how extensively AI researchers considered, if at all, the potential copyright issues surrounding the collection and use of large datasets that included numerous copyrighted works. Some researchers have recently taken the view that "[r]esearchers, at least in the United States, have long relied on

---

281. *See* Quoc V. Le et al., Building High-Level Features Using Large Scale Unsupervised Learning 1–2 (July 12, 2012) (unpublished manuscript) (accepted for the 29th International Conference on Machine Learning), https://arxiv.org/pdf/1112.6209 [https://perma.cc/TK6T-HELA]; *Andrew Ng*, WIKIPEDIA (July 30, 2025), https://en.wikipedia.org/wiki/Andrew_Ng [https://perma.cc/4XUU-WJ7N].

282. *See* LeCun, Bengio & Hinton, *supra* note 271, at 436 (emphasis added); *see also* Desai & Reidl, *supra* note 246, at 77 ("In LLM research, the method itself mandates larger datasets.").

283. *See, e.g.*, *Datasets*, KAGGLE, https://www.kaggle.com/datasets [https://perma.cc/2DK2-94SX] (last visited July 19, 2025); *Dataset Search*, GOOGLE, https://datasetsearch.research.google.com/ [https://perma.cc/FQ9P-ARBU] (last visited Aug. 3, 2025); *Dataverse*, HARVARD DATAVERSE, https://dataverse.harvard.edu/ [https://perma.cc/UMJ9-SJ7A] (last visited Aug. 13, 2025); *Welcome to the UC Irvine Machine Learning Repository*, UC IRVINE, https://archive.ics.uci.edu/ [https://perma.cc/84WM-HSP6] (last visited Aug. 3, 2025).

284. *See, e.g.*, *Large-Scale Artificial Intelligence Open Network*, LAION, https://laion.ai/ [https://perma.cc/7RLR-Y73B] (last visited July 13, 2025); *About*, ELEUTHERAI, https://www.eleuther.ai/about [https://perma.cc/K4QA-UV5C] (last visited Aug. 3, 2025); *Common Crawl Maintains a Free, Open Repository of Web Crawl Data That Can Be Used by Anyone*, COMMON CRAWL, https://commoncrawl.org/ [https://perma.cc/698K-ZUBB] (last visited July 17, 2025); *see* Kate Knibbs, *Publishers Target Common Crawl in Fight over AI Training Data*, WIRED (June 13, 2024, at 11:21 CT), https://www.wired.com/story/the-fight-against-ai-comes-to-a-foundational-data-set/ [https://perma.cc/8A44-MRVN].

285. *See Datasets*, PAPERS WITH CODE, https://paperswithcode.com/datasets [https://perma.cc/7X9F-X4BW] (last visited July 15, 2025).

the legal doctrine of fair use to avoid liability from using copyrighted data."[286] Even if this statement accurately reflects the sentiments of American researchers, there is very little discussion of any copyright concerns in the AI research literature. Perhaps the dearth of copyright discussion is understandable given that AI researchers are not typically lawyers and may have only a vague understanding of copyright law.

For example, in 2008, Gary B. Huang, Marwan Mattar, Tamara Berg, and Erik Learned-Miller compiled, used, and shared a "Labelled Faces in the Wild" database of 13,233 face images, while citing a non-exhaustive list of face images databases compiled by others.[287] The paper indicates that the images were extracted from "news articles on the web," meaning they were likely copyrighted images.[288] The researchers compiled and openly shared their dataset of images with others "to make research performed with the database as consistent and comparable as possible."[289] Their dataset is still publicly available.[290]

ImageNet is an even bigger dataset of images. Indeed, it is one of the most famous datasets of images in the annals of AI research, consisting of over 14 million images scraped from the internet.[291] But the compiler of the ImageNet database admits on its website that it "does not own the copyright of the images."[292] For that reason, ImageNet is shared with others "only for non-commercial research and educational purposes."[293] And the website imposes a host of other conditions on the use of the dataset, including an agreement by other researchers to following condition:

---

286. Henderson et al., *supra* note 248, at 2 (emphasis omitted).

287. *See* Gary B. Huang et al., Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments 1, 11, 14 (Sep. 16, 2008) (unpublished manuscript) (submitted at Workshop on Faces in 'Real-Life' Images), https://inria.hal.science/file/index/docid/321923/filename/Huang_long_eccv2008-lfw.pdf [https://perma.cc/3ZLY-6KUW].

288. *Id.* at 3.

289. *Id.* at 1, 13.

290. *See Labelled Faces in the Wild (LFW) Dataset*, KAGGLE, https://www.kaggle.com/datasets/jessicali9530/lfw-dataset [https://perma.cc/4YT2-6EVP] (last visited Aug. 3, 2025).

291. *See ImageNet: A Pioneering Vision for Computers*, HIST. OF DATA SCI. (Aug. 27, 2021), https://www.historyofdatascience.com/imagenet-a-pioneering-vision-for-computers/ [https://perma.cc/Z2LX-9RGG]; Dave Gershgorn, *The Data That Transformed AI Research—and Possibly the World*, QUARTZ (July 20, 2022), https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world [https://perma.cc/B7ER-TMU2].

292. *About ImageNet*, IMAGENET, https://image-net.org/about.php [https://perma.cc/Q4P3-Z9QS] (last visited July 14, 2025).

293. *Download*, IMAGENET, https://image-net.org/download.php [https://perma.cc/B6LR-YS7Q] (last visited July 20, 2025).

> Researcher . . . shall defend and indemnify . . . Princeton University, and Stanford University, including their employees, Trustees, officers and agents, against any and all claims arising from Researcher's use of the Database, including but not limited to Researcher's use of any copies of copyrighted images that he or she may create from the Database.[294]

ImageNet's more prominent copyright concern on its website probably reflects the popularity and importance of the dataset. It is worth briefly summarizing the origin of the dataset and its importance to AI research. While a professor of computer science at the University of Illinois Urbana-Champaign in 2006, Fei-Fei Li had the key insight that the research for computer vision (in which AI programs can recognize things in images or media) was suffering from a lack of enough data.[295] In Li's view, "the key to better model performance could be bigger datasets that reflected the diversity of the real world."[296] To that end, while a professor at Princeton, Li began an ambitious project of collecting millions of images downloaded from Google image search[297]—more than 14 million images[298]—and enlisting people on Amazon's Mechanical Turk to label the images (with terms from the Wordnet synonym sets), so the images are paired with textual descriptions in the dataset now titled ImageNet.[299] The ImageNet dataset was used as a part of an important annual competition among researchers, known as the ImageNet Large Scale Visual Recognition Challenge, or ILSVRC.[300] Researchers developed computer programs using neural networks to identify as many as the objects in the images as possible. The ILSVRC and the ImageNet dataset sparked rapid improvement in computer vision within just five years, which "[m]any see . . . as the catalyst for the AI boom the world

294. *Id.* This copyright indemnification clause also existed in 2014. *See Download FAQ*, IMAGENET, https://web.archive.org/web/20140215073453/http://image-net.org/download-faq [https://perma.cc/YVZ5-NSY8] (last visited Aug. 23, 2025) (indemnity clause in 2014).

295. *See AlexNet and ImageNet: The Birth of Deep Learning*, PINECONE, https://www.pinecone.io/learn/series/image-search/imagenet/ [https://perma.cc/UX6E-6DEL] (last visited July 19, 2025).

296. *Id.*

297. *See* FEI-FEI LI, THE WORLDS I SEE: CURIOSITY, EXPLORATION, AND DISCOVERY AT THE DAWN OF AI, 153, 160, 163 (2023) (describing process of creating ImageNet).

298. *See* Nico Klingler, *ImageNet Dataset: Evolution & Applications*, VISO.AI (Oct. 11, 2024), https://viso.ai/deep-learning/imagenet/ [https://perma.cc/MH7P-G7HD].

299. *See* LI, *supra* note 297, at 158–60, 171–73; *About ImageNet*, *supra* note 292.

300. Jason Brownlee, *A Gentle Introduction to the ImageNet Challenge (ILSVRC)*, MACH. LEARNING MASTERY (July 5, 2019), https://machinelearningmastery.com/introduction-to-the-imagenet-large-scale-visual-recognition-challenge-ilsvrc/ [https://perma.cc/XWZ6-45JY].

is experiencing today."[301] As Hinton said in 2023, "[s]he was the first computer vision researcher to truly understand the power of big data and her work opened the floodgates for deep learning."[302]

In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton used the ImageNet dataset for the AI training discussed in their pathbreaking paper in 2012,[303] also known as the AlexNet paper.[304] The researchers showed that using larger datasets to train AI models, instead of the small datasets that had been used in prior studies, improved the AI model's abilities: "objects in realistic settings exhibit considerable variability, so to learn to recognize them *it is necessary to use much larger training sets*."[305] Using larger datasets was necessary to develop better AI models.[306] As the AI researchers concluded: "All of our experiments suggest *that our results can be improved* simply *by waiting for faster GPUs and bigger datasets* to become available."[307]

Their advance in AI training is considered a major breakthrough for AI—what one commentator called the "[b]irth of deep learning."[308] The AlexNet paper is one of the most cited in computer science,[309] now with more than 182,000 citations.[310] Their AI model showed that it "can outperform traditional image classification methods when *trained on large datasets*."[311] Yet, none of this advancement in AI could have occurred if Krizhevsky, Sutskever, and Hinton had been required to obtain copyright licenses for all 1.2 million images they used to train their AI model, which they also shared publicly.[312] And,

---

301. Gershgorn, *supra* note 291.

302. Geoffrey Hinton (@geoffreyhinton), X (Nov. 7, 2023, at 11:06 CT), https://x.com/geoffreyhinton/status/1721937095860633663 [https://perma.cc/NZW4-LRGV].

303. Alex Krizhevsky, Ilya Sutskever & Geoffrey Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, COMMC'NS ASSOC. COMPUTING MACH., June 2017, at 84, 85 (originally published in the *Proceedings of the 25th International Conference on Neural Information Processing Systems* in 2012).

304. *See* METZ, *supra* note 247, at 96.

305. Krizhevsky, Sutskever & Hinton, *supra* note 303, at 84, 90 (emphasis added).

306. *Id.* at 85, 90.

307. *Id.* at 85 (emphasis added).

308. *See AlexNet and ImageNet*, *supra* note 295.

309. *See* METZ, *supra* note 247, at 97.

310. *See* Alex Krzhevsky, GOOGLE SCHOLAR, https://scholar.google.com/citations?user=xegzhJcAAAAJ&hl=en [https://perma.cc/KD4C-X23X] (last visited Sep. 16, 2025) (listing citations for AlexNet paper).

311. Tim Mucci, *The History of AI*, IBM (emphasis added), https://www.ibm.com/think/topics/history-of-artificial-intelligence [https://perma.cc/FAA2-FKFJ] (last visited July 17, 2025).

312. *See Cuda-Convnet*, GOOGLE CODE ARCHIVE, https://code.google.com/archive/p/cuda-convnet/ [https://perma.cc/9UMX-9EK8] (last visited Aug. 16, 2025); Personal Website by Alex Krizhevsky, UNIV. OF TORONTO, https://www.cs.toronto.edu/~kriz/ [https://perma.cc/AY6D-8XXC] (last visited Aug. 16, 2025).

without their breakthrough in deep learning, it seems doubtful that AI would be as far advanced as it is today.

Other major AI advances commonly involved large datasets consisting of copyrighted works often scraped from the internet.[313] Indeed, it's likely that many leading AI researchers in the United States at one point used a dataset that included copyrighted works for which they did not obtain licenses to use.[314]

Consider the BookCorpus dataset introduced by researchers from the University of Toronto and MIT in 2015.[315] The researchers copied and collected a "corpus of 11,038 books from the web," albeit providing little, if any, information about their provenance.[316] The BookCorpus dataset or derivatives from it were later used by other researchers in major projects yielding significant advances in AI, including the Google researchers' BERT model,[317] a joint project between University of Washington

313. *See, e.g.*, Quoc V. Le et al., *supra* note 281, at 1, 4 (dataset of 10 million images "downloaded from the Internet"); Ian J. Goodfellow et al., Generative Adversarial Nets 1, 6, 9 n.19 (June 10, 2014) (unpublished manuscript) (accepted at the Advances in Neural Information Processing Systems Proceedings) https://arxiv.org/pdf/1406.2661 [https://perma.cc/6P97-D4CC] (datasets included CIFAR-10 "tiny images" dataset). *See generally* Alex Krizhevsky, Learning Multiple Layers of Features from Tiny Images (Apr. 8, 2009) (Master's thesis, University of Toronto) (on file with the University of Toronto) (explaining provenance of 80 million images researchers at MIT and New York University collected from Internet).

314. *See About ImageNet*, *supra* note 292; *see, e.g.*, Kaiming He et al., *Deep Residual Learning for Image Recognition*, *in* 29TH IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 770, 773 (2016), https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf [https://perma.cc/KJ3J-NYQS] (using ImageNet dataset); Karen Simonyan & Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition 1, 8 (Apr. 10, 2015) (unpublished manuscript) (workshop papers at the 2015 Internation Conference on Learning Representations), https://arxiv.org/pdf/1409.1556 [https://perma.cc/X69N-5GNH] (using ImageNet dataset from 2012).

315. *See* Yukun Zhu et al., *Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books*, *in* 2017 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION 19, 21 (2017), https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Zhu_Aligning_Books_and_ICCV_2015_paper.pdf [https://perma.cc/X568-QNTG].

316. *Id.* at 21. According to Hugging Face, there were 2,930 duplicates in the dataset, leaving 7,185 unique books. *See Dataset Card for BookCorpus*, HUGGING FACE, https://huggingface.co/datasets/bookcorpus/bookcorpus [https://perma.cc/TTZ8-LGEW] (last visited July 19, 2025).

317. *See* Jacob Devlin et al., *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, *in* 2019 CONFERENCE ON THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 4171, 4174 (2019), https://aclanthology.org/N19-1.pdf [https://perma.cc/KBY5-J2SW].

researchers and Facebook AI called the RoBERTa model, OpenAI's GPT model,[318] and the XLNet model.[319]

The BookCorpus later drew controversy, however. In 2016, the Authors Guild criticized Google's commercial use of the books for its own AI technology.[320] Then, in 2021, two researchers, Jack Bandy and Nicholas Vincent, went further and questioned whether the entire BookCorpus dataset violated the copyrights of the books authors whose works were copied from the Smashwords platform without permission or licenses.[321] Bandy and Vincent's sharp critique raised an important question. But their article didn't go far enough. The copyright question they raised applies not just to BookCorpus, but to *every* dataset ever compiled by AI researchers with copyrighted works that were not licensed from the copyright holders.

A copyright lawsuit against OpenAI and Microsoft now alleges that BookCorpus was one of several pirated datasets that OpenAI used to train its AI models.[322] But, if BookCorpus is considered infringement, ImageNet,[323] the Pile compilation of the nonprofit EleutherAI,[324] the datasets of image-text pairs compiled by the German nonprofit

318. *See* Yinhan Liu et al., Roberta: A Robustly Optimized BERT Pretraining Approach 2–3 (2019) (unpublished manuscript) (submitted as a conference paper for the 2020 ICLR), https://arxiv.org/pdf/1907.11692 [https://perma.cc/QR7H-2T2K]; Alec Raford et al., Improving Language Understanding by Generative Pre-Training, OPENAI (June 11, 2018), https://openai.com/index/language-unsupervised/ [https://perma.cc/286V-NYYT].

319. *See* Zhilin Yang, XLNet: Generalized Autoregressive Pretraining for Language Understanding 6 (2019) (unpublished manuscript) (accepted for the 33rd Conference on Neural Information Processing Systems), https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf [https://perma.cc/LC9G-UNLD].

320. Richard Lea, *Google Swallows 11,000 Novels to Improve AI's Conversation*, GUARDIAN (Sep. 28, 2016, at 05:00 EDT), https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation [https://perma.cc/CV9E-3HEK].

321. *See* Jack Bandy & Nicholas Vincent, Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus 1, 6, 8 (2021) (unpublished manuscript) (accepted at the 35th Conference on Neural Information Processing Systems), https://openreview.net/pdf?id=Qd_eU1wvJeu [https://perma.cc/8DQK-5U6M] ("BookCorpus likely violates copyright restrictions for many books . . . ."); *see also* Lea, *supra* note 320.

322. *See* Complaint at 41–47, Denial v. OpenAI, Inc., No. 3:25-CV-05495 (N.D. Cal. June 30, 2025).

323. *See About ImageNet*, *supra* note 292.

324. *See* Leo Gao et al., The Pile: An 800GB Dataset of Diverse Text for Language Modeling 1, 5 (Dec. 31, 2020) (unpublished manuscript), https://arxiv.org/pdf/2101.00027 [https://perma.cc/2Q74-3D3B].

LAOIN,[325] Common Crawl,[326] and every other dataset containing unlicensed copyrighted works are, too. And the entire edifice on which all university-based AI research in the United States was built—datasets—may come crashing down like a house of cards.[327] The AI lawsuits have already raised a cloud over the very practice of collecting large datasets for AI research. The original BookCorpus is no longer available, although others have reportedly cloned it.[328] An even larger and more controversial dataset of 196,640 books—called "[B]ooks 3" compiled by Shawn Presser in 2020 ostensibly for purpose of AI training[329]—has garnered even more opprobrium.[330] In 2023, the website that hosted it removed it due to a claim of copyright infringement from a Danish group.[331] So-called "shadow librar[ies]" consisting of unauthorized copies of books have elicited condemnation

325. *See About*, LAION, https://laion.ai/about/ [https://perma.cc/AX8X-UCBY] (last visited July 15, 2025); Christoph Schuhmann, *AI as a Superpower: LAOIN and the Role of Open Source in Artificial Intelligence*, MLCON (June 21, 2023), https://mlconference.ai/blog/ai-as-a-superpower-laion-and-the-role-of-open-source-in-artificial-intelligence [https://perma.cc/47BK-FSVH].

326. *See* Alistair Barr & Kali Hays, *The New York Times Got Its Content Removed from One of the Biggest AI Training Datasets. Here's How It Did It.*, BUS. INSIDER (Nov. 8, 2023, at 10:00 CT), https://www.businessinsider.com/new-york-times-content-removed-common-crawl-ai-training-dataset-2023-11 [https://perma.cc/WC6R-XZNS].

327. Not all datasets available online developed from or for university researchers. The Library Genesis, also known as "LibGen," dataset is reputed to be a "shadow library" of millions of pirated books, articles, and other works that were assembled by a group located in Russia. *See Library Genesis*, WIKIPEDIA, https://en.wikipedia.org/wiki/Library_Genesis [https://perma.cc/T3XS-6ZM3] (last visited July 18, 2025); Kate Knibbs, *Meta Secretly Trained Its AI on a Notorious Piracy Database, Newly Unredacted Court Docs Reveal*, WIRED (Jan. 9, 2025, at 17:33 CT), https://www.wired.com/story/new-documents-unredacted-meta-copyright-ai-lawsuit/ [https://perma.cc/YE3Y-JCYQ]. Meta's apparent use of this notorious dataset, using torrent file sharing, has become a flash point in the copyright lawsuit brought by book authors. *See Will Meta's Alleged "Seeding" of Pirated Copies via BitTorrent Come Back to Bite Meta & Mark Zuckerberg?*, CHATGPT IS EATING THE WORLD (Jan. 12, 2025), https://chatgptiseatingtheworld.com/2025/01/12/will-metas-alleged-seeding-of-pirated-copies-via-bittorrent-come-back-to-bite-meta-mark-zuckerberg/ [https://perma.cc/JEJ4-8KLE].

328. *See* Steven van de Graaf, *Replicating the Toronto BookCorpus Dataset—a Write-Up*, MEDIUM (Dec. 6, 2019), https://medium.com/data-science/replicating-the-toronto-bookcorpus-dataset-a-write-up-44ea7b87d091 [https://perma.cc/D79K-MH5U].

329. *See* Shawn Presser (@theshawwn), X (Oct. 25, 2020, at 03:32 CT), https://x.com/theshawwn/status/1320282149329784833?s=61&t=jQbmCk1JqL7depzFWJNuPA [https://perma.cc/8LLQ-G9GT].

330. *See* Katie Knibbs, *The Battle Over Books3 Could Change AI Forever*, WIRED (Sep. 4, 2023, at 6:00 CT), https://www.wired.com/story/battle-over-books3/ [https://perma.cc/U89N-BRJS]; Alex Reisner, *These 183,000 Books Are Fueling the Biggest Fight in Publishing and Tech*, ATLANTIC (Sep. 25, 2023), https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/ [https://perma.cc/H4BA-DETR].

331. *See* Rizwan Choudhury, *Anti-Piracy Group Shuts Down Book3, A Popular Dataset for AI Models*, INTERESTING ENG'G (Aug. 20, 2023, at 09:03 ET), https://interestingengineering.com/innovation/anti-piracy-group-shuts-down-books3-a-popular-dataset-for-ai-models [https://perma.cc/ATF8-ZAB5].

from the book authors in the various copyright lawsuits they have filed against AI companies.[332]

2. Factor 1: The Further Purpose and Different Character of Researchers Using Copyrighted Works to Research, Develop, Train, and Improve New AI Models at Universities. With this important history underlying the practice of training AI models on large datasets in mind, we can examine Factor 1 of fair use as applied to university-based AI research, where this practice began. How should courts frame "the purpose and character of the use" of copyrighted works at universities to train AI models under Factor 1 of fair use?[333] Do university AI researchers have a further purpose or different character in their use of the copyrighted works in datasets collected and used without permission of the copyright owners?

Based on the Supreme Court's last two fair use decisions, *Warhol* and *Google*, the Court appears less inclined to recognize or coin new legal concepts for the fair use analysis. Although the two decisions retain the concept of "transformative" the Court adopted (from Judge Leval's law review article) in *Campbell*,[334] the Court did not add any new fair use terminology but hewed closely to the facts and examined whether the defendant's use of the copyrighted works had a different or further purpose than the copyright holder's purpose.[335]

Following the Court's lead, we can describe the purpose and character of the use of copyrighted works AI researchers to train AI models as follows:

> **Purpose of using works to train AI models**: to develop a new AI model. AI researchers at universities use copyrighted works in large datasets for the further purpose to research, develop, and create AI models that work and that can perform better or with new functionalities, including generative capabilities to create new, non-infringing content. This practice of using larger datasets, including with copyrighted works, in AI training was a

332. *See* Barr & Hays, *supra* note 326; Ella Creamer, *'Meta Has Stolen Books': Authors to Protest in London Against AI Trained Using 'Shadow Library,'* GUARDIAN (Apr. 3, 2025, at 03:00 ET), https://www.theguardian.com/books/2025/apr/03/meta-has-stolen-books-authors-to-protest-in-london-against-ai-trained-using-shadow-library [https://perma.cc/LT3A-RL4C]; *see Library Genius*, *supra* note 327.

333. 17 U.S.C. § 107(1).

334. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994); Leval, *supra* note 149, at 1111.

335. *See* Andy Warhol Found. v. Goldsmith, 143 S. Ct. 1258, 1262, 1274–75 (2023); Google LLC v. Oracle Am., Inc. 141 S. Ct 1183, 1188, 1197 (2021).

> key insight leading university researchers identified that propelled the advances in AI development.
>
> **Character of using works to train AI models**: computational analysis. AI researchers' use of copyrighted works in AI training involves the AI model's computational analysis of the entire corpus of training datasets: the model dissects the entire corpus of datasets into tiny elements, and the model assigns weights, or numerical values, representing the model's assessment of the statistical relationships among the elements of the entire corpus of datasets in a process of deep learning.

For the time being, I have not included the noncommercial nature of university-based research because I want to show Factor 1 favors fair use even without weighing the noncommercial nature of this activity.

Both the purpose and character of this use of copyrighted works in AI training are different from the copyright holder's purpose and character of use in exploiting their works by offering them for human enjoyment. This use of works in AI training serves a "further purpose," the *transformative purpose* to research, develop, and create the new AI model, a new technology.[336] Section 107 recognizes "research" as a purpose of fair use.[337] Because AI is still fast developing and has many unknowns, AI research is an important purpose to advance the development of AI, which is consistent with the national policy of the United States set forth by President Trump.[338] Even more fundamentally, AI training with copyrighted works serves the transformative purpose of developing and creating a new technology, new AI models with better or more advanced capabilities than the existing state of the art, including now reasoning models.[339] These AI models simply do not exist without training through large datasets—and they improve as the datasets get larger.

Likewise, the *character* of use of works in AI training involves something different from human enjoyment of the works: the AI model's computational analysis of the entire corpus of datasets.

---

336. *See* Google LLC, 141 S. Ct. at 1188, 1202–03 ("[I]n determining whether a use is 'transformative,' we must go further and examine the copying's more specifically described 'purpose[s]' and 'character.'") (second alternation in original) (citing 17 U.S.C. § 107(1)).

337. 17 U.S.C. § 107.

338. *See* Exec. Order No. 14,179, *supra* note 9.

339. *Cf.* Rina Diane Caballar & Cole Stryker, *What Is Reasoning in AI?*, IBM (Mar. 14, 2025), https://www.ibm.com/think/topics/ai-reasoning [https://perma.cc/R3WK-C38S] (explaining that advances in AI training have produced reasoning models with more advanced capabilities based on training from datasets).

The model dissects the entire corpus of datasets into tiny elements ("tokens" for Large Language Models (LLMs) and pixels for diffusion models) and the model assigns weights reflecting the model's assessment of the statistical relationships among the elements of the entire corpus of datasets in a deep learning process.[340] Through this learning, the model is capable of generating responses to human queries or prompts.[341] After training, the AI models do not store the materials in the datasets on which they were trained—they are not giant databases, putting aside the issue of inadvertent memorization by an AI model.[342] Instead, they are *artificial intelligence*, a new technology that has escaped successful human development for decades.

To borrow the Supreme Court's reasoning in *Google v. Oracle*, another new technology fair use case, the AI training "seeks to create new products," a "highly creative and innovative tool," and "a new and transformative program."[343] Such use of copyrighted works is "consistent with that creative 'progress' that is the basic constitutional objective of copyright itself" of the Progress Clause.[344] In *Google*, the Court recognized that Google's creating of a new computing platform had a different, transformative purpose than the plaintiff Oracle, whose Java declaring code Google copied.[345] That same reasoning should apply here, perhaps with even greater force given that developing AI is a recognized national policy.[346] One uses copyrighted works to develop new technologies that advance innovation—and ultimately progress in the United States.

---

340. *See* Computer & Communications Industry Association, *supra* note 39, at 10; LeCun, Bengio & Hinton, *supra* note 271; Hinton, Osindero & Teh, *supra* note 276, at 1528, 1540; *see also* Samuelson, *supra* note 129, at 1539 (describing fair use decisions for "computational analyses" of large databases to enable people to search the works in the database); *Id.* at 1547–48 (arguing that AI training is "very similar to the computational search and research tool fair use cases," with the difference that "generative AI developers make copies of works to train computational models to produce new content, not just to display parts of existing contents").

341. *See How Does AI Answer Questions?*, MAXTRAIN (May 7, 2024), https://maxtrain.com/2024/01/09/how-does-ai-answer-questions/ [https://perma.cc/7VJC-H47E].

342. *See Our Approach to Data and AI*, OPENAI (May 7, 2024), https://openai.com/index/approach-to-data-and-ai/ [https://perma.cc/JP77-USP2] ("AI models learn from relationships in information to create something new; they don't store data like a database. When we train language models, we take trillions of words, and ask a computer to come up with an equation that best describes the relationship among the words and the underlying process that produced them. After the training process is complete, the AI model does not retain access to data analyzed in training.").

343. Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1203, 1209 (2021).

344. *Id.* at 1203.

345. *Id.* at 1203, 1209.

346. *See supra* notes 9–10 and accompanying text.

To return to the analytic framework elaborated in Part II, we can plot the further purpose and different character of AI developers using copyrighted works to train AI models, as shown in Table 5 below. As apparent from the Table, AI training follows the pattern of other successful technology fair uses: the use of copyrighted works was for the further purpose of creating a new technology—and it did so in a way without (re)distributing copies of the works during the public's use of the technology, putting aside the problem of occasional regurgitations (an issue discussed later).

Table 5. The Further Purpose and Different Character of Use to Create AI Model.

Facts	**Technology development:** Use of Copyrighted Works to Create New Technology?	**Technology usage:** Use of Copyrighted Works in Public Use of Technology?	**Factor 1:** Use of Copyrighted Works Had Further Purpose or Different Character?
AI model trained on copyrighted works in training dataset.	Yes, to create, research, develop, and train new AI model through deep learning.	No, copies of works from training datasets are not stored in AI model. AI developer does not intend any "memorized" copies, but if they are regurgitated, should be analyzed as separate use in the output of AI.	Yes, use of works has further purpose to create, research, develop, and train AI model. The character of use by the AI model involves computational analysis, assigning weights to patterns and relationships identified among tiny elements. At universities, the research is typically noncommercial.

Some critics may object to this conclusion of a transformative purpose in AI training by analogizing to the example of students learning from copyrighted textbooks.[347] If students must pay for their textbooks, so should AI developers, they argue.[348] Before the hearing on the summary judgment motions in *Kadrey v. Meta*, Judge Chhabria drew on this analogy in posing a question using the hypothetical of a professor downloading a pirated copy of a book to give to a "brilliant student" who "use[s] the knowledge to do something transformative."[349]

Analogizing student learning to AI or machine learning may have superficial appeal. But I think it oversimplifies the two types of learning—and trivializes the national interest in promoting technological innovation in the United States, especially now in AI. The key difference is that AI learning is necessary for the creation of a new technology, the AI model itself, whereas student learning does not necessarily yield anything for the public's benefit, especially not a technology that can be utilized by potentially millions of Americans. Put differently, technological innovation in the United States does not hinge on whether a student can get free access to a copyrighted book. But innovation in AI may be greatly affected, if not hindered, if AI researchers have no fair use defense for their use of copyrighted works. As Judge Chhabria explained in *Kadrey v. Meta*, "[b]y creating a tool that anyone can use, Meta's copying has the potential to exponentially multiply creative expression in a way that teaching individual people does not."[350]

Such AI research and development results in technological innovation for the United States and advances progress in ways that educating students does not. If successful, the new technology can directly benefit millions of Americans and add to national productivity and the economy, not to mention the United States' standing in the world. The sheer scale of the datasets needed to train foundational AI models makes it far more difficult, if not impossible, to seek licenses for all the training materials. By

347. *See* Robert Brauneis, *Copyright and the Training of Human Authors and Generative Machines*, 47 COLUM. J.L. & ARTS 1, 14, 43 n.169 (2025).

348. *See id.* at 14.

349. *See Expressing Skepticism of Both Sides' Arguments, Judge Chhabria Poses 12 Questions on Fair Use & Pirated Books*, CHATGPT IS EATING THE WORLD (Apr. 30, 2025) [hereinafter *Expressing Skepticism*], https://chatgptiseatingtheworld.com/2025/04/30/expressing-skepticism-of-both-sides-arguments-judge-chhabria-poses-12-questions-on-fair-use-pirated-books/ [https://perma.cc/UV3U-E5W8].

350. Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *10 (N.D. Cal. June 25, 2025).

contrast, a typical university course has a far more limited scope of materials—which can be readily purchased from bookstores or licensed from the Copyright Clearance Center.[351] Giving students free access to course materials may result in their getting an A on the exam. Giving AI researchers free access to datasets may result in new advances in AI that redound to the public's benefit—and the United States' global standing.

Often, the model is not just a new technology in the sense of the latest version of a smartphone, but a pioneering technology that had not even existed just three years ago. Training is an essential step in the AI model's creation, effectiveness, and performance. This overall purpose—to research, develop, and create AI models that work and that can perform better or with new functionalities—should be considered a fair use purpose of technology development,[352] with the rest of the fair use factors still to be weighed based on the evidence presented in the respective cases. Using larger, more diverse datasets can also serve the additional purpose of de-biasing an AI model, mitigating the chance the AI model contains biases due to being trained on less diverse or unrepresentative datasets (i.e., dataset biases).[353] Training on small datasets may create more problems.[354]

---

351. *See e.g.*, *Copyright: Academic Copying and Student Course Packets*, YALE UNIV., https://ogc.yale.edu/ogc/copyright-academic-copying-and-student-course-packets [https://perma.cc/Z2VJ-RAG9] (last visited Aug. 30, 2025).

352. It is important to underscore that having a fair use purpose does not guarantee a fair use was made. For example, the purpose of creating a Harry Potter lexicon was a legitimate fair use purpose, but the lexicon J.K. Rowling took issue within litigation went too far in its copying of Rowling's works in the lexicon originally planned for publication. *See* Warner Bros. Entm't v. RDR Books, 575 Supp. 2d 513, 541, 547, 551 (S.D.N.Y. 2008). Yet, afterwards, the publisher was able to publish a cleaned-up version of the Lexicon, consistent with its transformative purpose. *See generally* STEVE VANDER ARK, THE LEXICON: AN UNAUTHORIZED GUIDE TO HARRY POTTER FICTION AND RELATED MATERIALS (2009).

353. *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 621 (2018); *see also* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 681 (2016) ("[I]f data mining draws inferences from a biased sample of the population, any decision that rests on these inferences may systematically disadvantage those who are under- or overrepresented in the dataset."); Judy Hoffman et al., One-Shot Adaptation of Supervised Deep Convolutional Models 1–2 (2014) (unpublished manuscript), https://arxiv.org/pdf/1312.6204 [https://perma.cc/MCY2-4E6C] (study using "the 1.2 million labeled images available in the 2012 ImageNet 1000-way classification dataset . . . to train the model in" "show[ed] that a generic supervised deep CNN model trained on a large dataset reduces, but does not remove, dataset bias") (workshop paper accepted to the International Conference on Learning Representations 2014); Desai & Reidl, *supra* note 246, at 80 (discussing how larger and more diverse datasets can reduce bias: "[H]aving a larger book dataset is good for one's model because more data likely improves the model's performance on a pure predictive performance metric *and* allows the model to incorporate a larger range of voices").

354. *See* Guo-Jun Qi & Jiebo Luo, *Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods*, 44 IEEE TRANSACTIONS ON PATTERN ANALYSIS & MACH. INTEL. 2168, 2168, 2181–82 (2022) (discussing "small data challenges" in AI training).

Granted, sometimes, the models may "memorize" some of the materials from their training for reasons that AI developers apparently do not understand enough to eliminate.[355] To the extent AI generators produce any regurgitated and "memorized" copies,[356] those outputs should be considered on a use-by-use basis under *Warhol* and may constitute infringement in some cases, an issue discussed later below.

B. Training AI Models by AI Companies

The history of training AI models developed by AI companies followed the same path as AI training at universities: scaling up the size of datasets was essential to improving the capabilities and performance of the AI models.

1. Technology Development: The History of AI Training by AI Companies and the Essential Role of Scaling the Size of Datasets. AI companies did not reinvent the AI training wheel. The method of training AI models with large datasets migrated to for-profit companies, both AI startups and Big Tech companies.[357] Often, AI researchers who started in academia were hired by these companies or even founded them. During the decade starting in 2010, Big Tech (e.g., Google, Meta, Microsoft) lured leading and promising university-based AI researchers, such as Hinton, LeCun, and Ng, who, in some cases, retained their university affiliations while also working for Big Tech companies.[358] AI companies built on the existing state of the art—and employed the same method of AI training by scaling with large datasets, an insight that university-based researchers had identified in the prior decade as necessary to improve AI model performance.[359] As researchers at OpenAI explained in 2016: "[D]eep learning . . . *must be scaled* to be truly impressive—a small neural network is a proof of concept, but a big neural network actually solves the problem and is useful."[360]

---

355. *See* Jiaheng Wei et al., Memorization in Deep Learning: A Survey 8–9 (June 6, 2024) (unpublished manuscript), https://arxiv.org/html/2406.03880v1 [https://perma.cc/UT6Q-T3B4].

356. For more on memorization in AI models, see *id*. at 7–12.

357. *See* Lee & Grimmelmann, *supra* note 248.

358. *See* METZ, *supra* note 247, at 84–89, 120–21, 196–97.

359. *See supra* notes 247–48 and accompanying text.

360. *Infrastructure for Deep Learning*, OPENAI (Aug. 29, 2016), https://openai.com/index/infrastructure-for-deep-learning/ [https://perma.cc/AGY3-ZRC8] (emphasis added); *see also Generative Models*, OPENAI (June 16, 2016), https://openai.com/index/generative-models/ [https://perma.cc/3QZG-ZSUT] ("Generative models are a rapidly advancing area of research. As we continue to advance these models and scale up the training and the datasets, we can expect to eventually generate samples that depict entirely plausible images or videos.").

In a seminal article first published in 2017, a team of Google researchers published "Attention Is All You Need," which discussed the major breakthrough of the transformer, or architecture of an AI model that allows it to rely "entirely on self-attention to compute representations of its input and output."[361] The Google researchers relied on the publicly available "WMT 2014 English-German dataset consisting of about 4.5 million sentence pairs," and also "the significantly larger WMT 2014 English-French dataset consisting of 36M sentences."[362] Both datasets include content extracted from the Common Crawl dataset.[363] Common Crawl is a nonprofit organization that crawls the internet and compiles in a dataset "raw web page data, metadata extracts, and text extracts" from websites.[364] At a minimum, the text extracts from websites likely include a good amount of copyrighted content scraped from the internet. For that reason, after the recent advent of AI, many major publishers and content producers have sent notices to Common Crawl to stop using their copyrighted content.[365] Given the inclusion of Common Crawl, some of the extracted content used by Google researchers was likely copyrighted.

Scaling up the size of datasets used to train AI models continued at OpenAI and other AI companies. Because Google shared its knowledge about transformers by posting its research paper online, a common practice of sharing knowledge even among AI companies, OpenAI developers took the idea and ran with it to develop their breakthrough technology for ChatGPT.[366] In 2020, researchers at OpenAI published an important paper describing

361. Ashish Vaswani et al., Attention Is All You Need 1–2 (Aug. 12, 2023) (unpublished manuscript) (accepted at the 31 Conference on Neural Info. Processing Sys.), https://arxiv.org/pdf/1706.03762 [https://perma.cc/2L8C-4MCV].

362. *Id.* at 7.

363. *See* Rishi Singh, *Cleaning and Normalizing WMT Dataset for French to English*, OPENNMT (June 16, 2020, at 14:48 CT), https://forum.opennmt.net/t/cleaning-and-normalizing-wmt-dataset-for-french-to-english/3785 [https://perma.cc/YT3P-69LC]; Ondřej Bojar et al., *Findings of the 2014 Workshop on Statistical Machine Translation*, Workshop on Stat. Mach. Translation, June 26–27, 2014, at 12, 14.

364. *See Overview*, COMMON CRAWL, https://commoncrawl.org/overview [https://perma.cc/XPA4-4PN6] (last visited July 29, 2025).

365. *See* Knibbs, *supra* note 284.

366. *See How Google Gave the Key Breakthrough Technology for ChatGPT to OpenAI*, CHAT GPT IS EATING THE WORLD (Aug. 15, 2023), https://chatgptiseatingtheworld.com/2023/08/15/how-google-gave-the-key-breakthrough-technology-for-chatgpt-to-openai/ [https://perma.cc/NQ4N-XL26].

their success in developing its GPT-3 model,[367] which was "a language model with 175 billion parameters, making it one of the largest and most sophisticated AI models" at the time.[368] As summarized in the article, which has been cited more than 50,000 times, OpenAI's LLM was able "to generate human-like text, engage in conversations, write code, translate languages and generate creative writing based on natural language prompts."[369]

The OpenAI researchers harnessed the same method Hinton and others had recognized before: scaling up the training of the AI model by using larger and more diverse datasets.[370] The paper included a list of the datasets used. Figure 1 shows the breakdown of the datasets:

Figure 1. Datasets Used to Train OpenAI's GPT-3 Model.[371]

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

OpenAI researchers demonstrated that "scaling up language models greatly improves task-agnostic, few-shot performance," based on AI training that involves only a few examples of what the AI model is supposed to learn but otherwise leaves the model on its own unsupervised learning.[372] OpenAI's use of datasets consisting of books—called Books1 and Books2—is the main reason it faces six lawsuits filed by book authors and the Authors

367. Tom B. Brown et al., *Language Models Are Few-Shot Learners*, 34 NEURAL INFO. PROCESSING SYS., at 1, 9 (2020); *see also* Alec Radford et al., *Language Models are Unsupervised Multitask Learners*, OPENAI, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [https://perma.cc/VH4F-QQ7S] (last visited July 29, 2025) ("Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText.").

368. Mucci, *supra* note 311.

369. *Id.*; GOOGLE SCHOLAR, "Language Models are Few-Shot Learners", https://scholar.google.com/scholar?hl=en&as_sdt=0%2C44&q=Language+Models+are+Few-Shot+Learners&btnG= [https://perma.cc/3UCM-HNL6] (last visited Sep. 17, 2025) (cited 54,147 times as of September 2025).

370. *See* Brown et al., *supra* note 367, at 6.

371. *Id.* at 9.

372. *See id.* at 1–3.

Guild.[373] How OpenAI collected and used the two books datasets has become a flashpoint in the lawsuits.[374] During the litigation, OpenAI disclosed that it had destroyed the two books datasets and no longer uses them for training.[375]

Meta employed the same "scaling up" method of AI training.[376] As it explained in 2023: "The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, *by training on more tokens than what is typically used*."[377] Indeed, Meta said it followed the approach used by OpenAI in the 2020 paper mentioned above.[378] And, not surprisingly, Meta also included the Books3 dataset among the datasets to train LLaMA,[379] for which it has now received similar condemnation from book authors as OpenAI has faced. In the Meta lawsuit, evidence indicates that Meta developers concluded, from its testing, that the Library Genesis books dataset was necessary to match the state-of-the-art benchmarks set by other companies, including OpenAI.[380] One independent study conducted by Berkeley researchers found that training an AI model with the Books3 dataset (also used by Meta) resulted in "significant improvements in LLM name cloze accuracy

373. *See Master List*, *supra* note 30 (listing lawsuits against OpenAI, including those filed by book authors).

374. *See* Darius Rafieyan & Hasan Chowdhury, *OpenAI Destroyed a Trove of Books Used to Train AI Models. The Employees Who Collected the Data Are Gone*, BUS. INSIDER (May 7, 2024, at 16:28 CT), https://www.businessinsider.com/openai-destroyed-ai-training-datasets-lawsuit-authors-books-copyright-2024-5 [https://perma.cc/ZJS7-WUXB].

375. *See id.* Adding to the controversy was the death of the former OpenAI employee Suchir Balaji, who reportedly was involved in the collection of some of the datasets used by OpenAI but later became a whistleblower who criticized OpenAI for committing copyright infringement. *See* Cade Metz, *Former OpenAI Researcher Says the Company Broke Copyright Law*, N.Y. TIMES (Oct. 23, 2024), https://www.nytimes.com/2024/10/23/technology/openai-copyright-law.html [https://perma.cc/WMG3-8LAQ]; Emily Mae Czachor, *OpenAI Whistleblower Suchir Balaji Dead at Age 26; Family Seeks Answers as Death Ruled Suicide*, CBS NEWS (Dec. 22, 2024, at 16:36 ET), https://www.cbsnews.com/news/suchir-balaji-openai-whistleblower-dead-california/ [https://perma.cc/K2TU-6345].

376. *See* Hugo Touvron et al., *LlaMA: Open and Efficient Foundation Language Models*, META AI (Feb. 24, 2023), https://ai.meta.com/research/publications/llama-open-and-efficient-foundation-language-models/ [https://perma.cc/ND3G-2XPQ].

377. *Id.* (emphasis added).

378. *See id.* ("Our training approach is similar to the methods described in previous work . . . .(citing Brown et al., 2020; Chowdhery et al., 2022)); *see also supra* notes 367–70 and accompanying text (discussing OpenAi's paper regarding its GPT-3 model).

379. *See* Knibbs, *supra* note 327.

380. *See Evidence of Meta's Use of LibGen Dataset and Seeding Torrents to Share Files. Wanted to Compete with OpenAI and Mistral.*, CHATGPT IS EATING THE WORLD (Feb. 6, 2025), https://chatgptiseatingtheworld.com/2025/02/06/evidence-of-metas-use-of-libgen-dataset-and-seeding-torrents-to-share-files-wanted-to-compete-with-openai-and-mistral/ [https://perma.cc/9HWR-ND9K].

on books available within the Books3 dataset compared to those not present in these data."[381] Using the cloze test of omitting words from sentences for the AI model to fill in, the researchers found that "direct access to the full text of a book [in] Books3 significantly enhances performance on the name cloze task."[382] Beyond OpenAI and Meta, the use of scaling was widely accepted among AI researchers as a preferred technique for training.[383]

The use of "pirated" books, including from so-called shadow libraries online, is one of the most contested issues in the AI litigation involving book authors.[384] Judge Alsup even suggested in dicta that such use is categorically infringing, no matter if immediately used to train AI models without establishing a permanent library.[385] I'll return to this controversy in the last Part since it raises a distinct set of issues. For now, it is important to recognize that starting with the collection of BookCorpus by AI researchers at universities, AI developers

381. Stella Jia & Abhishek Nagaraj, *Cloze Encounters: The Impact of Pirated Data Access on LLM Performance* 1, 7, 10, 17 (Nat'l Bureau of Econ. Rsch., Working Paper No. 33598, 2025).

382. *Id.* at 3–4.

383. *See, e.g.*, Hao Li et al., *On the Scalability of Diffusion-Based Text-To-Image Generation*, *in* 2024 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 9400, 9406, https://ieeexplore.ieee.org/document/10655871 [https://perma.cc/8AHR-48B6]; Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models*, *in* 36 NEURAL INFO. PROCESSING SYS. 8, 10 (2022), https://papers.nips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf [https://perma.cc/UZW7-ESCU]; Aditya Ramesh et al., *Zero-Shot Text-to-Image Generation*, PROCS. OF MACH. LEARNING (2020), https://proceedings.mlr.press/v139/ramesh21a/ramesh21a.pdf [https://perma.cc/TQX9-ZPPY] ("We find that scale can lead to improved generalization, both in terms of zero-shot performance relative to previous domain-specific approaches, and in terms of the range of capabilities that emerge from a single generative model."); Tom Henighan et al., Scaling Laws for Autoregressive Generative Modeling 5 (Nov. 6, 2020) (unpublished manuscript), https://arxiv.org/pdf/2010.14701 [https://perma.cc/K354-YKXX]; Jared Kaplan et al., Scaling Laws for Neural Language Models 7 (Jan. 23, 2020) (unpublished manuscript), https://arxiv.org/pdf/2001.08361 [https://perma.cc/5A77-46RQ]; *see* CHRISTOPHER SUMMERFIELD, THESE STRANGE NEW MINDS: HOW AI LEARNED TO TALK AND WHAT IT MEANS 108–09 (2025) (explaining breakthrough in AI through scaling).

384. *Compare* Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *11 (N.D. Cal. June 23, 2025) (finding that Anthropic's use of shadow library to build a permanent library was a use separate from training and not fair use) *with* Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *9–12 (N.D. Cal. June 25, 2025) (finding that Meta's use of shadow libraries was for the ultimate transformative purpose of training its AI model while noting that "if Meta's act of downloading propped up these libraries or perpetuated their unlawful activities—for instance, if they got ad revenue from Meta's visits to their websites—then that could affect the 'character' of Meta's use" but plaintiffs failed to present such evidence).

385. *See Bartz*, 2025 WL 1741691, at *11 ("Such piracy of otherwise available copies is inherently, irredeemably infringing even if the pirated copies are immediately used for the transformative use and immediately discarded."). This part was dicta because Judge Alsup treated Anthropic's building of a library with pirated books as a separate use. *See id.* ("But this order need not decide this case on that rule. Anthropic did not use these copies only for training its LLM. Indeed, it retained pirated copies even after deciding it would not use them or copies from them for training its LLMs ever again.").

recognized that incorporating books in scaling their datasets yielded better results from AI models.[386] Perhaps that should not be surprising given that many published books probably embody much higher quality and more polished text, or use of language.

After copyright holders sued these companies starting in 2023, it is less common for the AI companies to disclose their datasets.[387] However, OpenAI's GPT-4 model reportedly has over 1.8 trillion parameters, roughly ten times the size of GPT-3.[388] Parameters represent units that the model has assigned in its analysis of the training data.[389] Using more training data should roughly correlate with more parameters.[390] Google's and Meta's large language models

386. *See, e.g.*, Zhu et al., *supra* note 315, at 1, 7, 9 ("Books provide us with very rich, descriptive text that conveys both fine-grained visual details (how people or scenes look like) as well as high-level semantics (what people think and feel, and how their states evolve through a story)."); Devlin et al., *supra* note 317, at 5 ("It is critical to use a document-level corpus rather than a shuffled sentence-level corpus . . . in order to extract long contiguous sequences."); Liu et al., *supra* note 318, at 3 ("BERT-style pretraining crucially relies on large quantities of text [which] . . . can result in improved end-task performance."); *Bartz*, 2025 WL 1741691, at *2 ("As Anthropic trained successive LLMs, it became convinced that using books was the most cost-effective means to achieve a world-class LLM."); *Kadrey*, 2025 WL 1752484, at *5 ("Books are good data for training LLMs' memories because, in the words of one of Meta's expert witnesses, they are 'long but consistent,' maintaining a particular style and coherent structure. They are also high quality in the sense that they generally are well written and use proper grammar (especially compared to text from the internet, which varies widely on these metrics).").

387. *See* Ana Andrijevic, *Generative AI and Transparency of Databases and Their Content, from a Copyright Perspective*, CTR. FOR ART LAW (May 21, 2024), https://itsartlaw.org/2024/05/21/generative-ai-and-transparency-of-databases-and-their-content-from-a-copyright-perspective/ [https://perma.cc/B2TZ-G4QY] ("Over the last few years, AI companies have become more cautious about disclosing the databases used to train their AI models, as illustrated, for instance, by Meta (for the training of Llama) and OpenAI (for the training of GPTs, i.e. Generative Pre-trained Transformers)." (footnotes omitted)); Buick, *supra* 268, at 184 ("Many major AI developers have shifted from detailed explanations of the training data used to train a particular model to single sentence descriptions.").

388. *See* Sarah Chudleigh, *GPT-3 vs GPT-4 | What's the Difference?*, BOTPRESS (Jan. 10, 2025), https://botpress.com/blog/gpt-3-vs-gpt-4-whats-the-difference [https://perma.cc/63Y3-CNGZ].

389. *See* Josh Howarth, *Number of Parameters in GPT-4 (Latest Data)*, EXPLODING TOPICS (June 17, 2025), https://explodingtopics.com/blog/gpt-parameters [https://perma.cc/39JJ-KL6X].

390. *See* Veronika Samborska, *Scaling Up: How Increasing Inputs Has Made Artificial Intelligence More Capable*, OUR WORLD IN DATA (Jan. 19, 2025), https://ourworldindata.org/scaling-up-ai [https://perma.cc/5LJQ-RRGM].

also reportedly have over 1 trillion parameters.[391] By one estimate, Anthropic's Claude 3 may have 2 trillion.[392]

These examples should not be read to indicate AI development will continue on this same path. Meta's LeCun believes that LLMs are almost obsolete, for example.[393] Other methods, such as distillation, and other types of models may prove to be the next wave of AI development.[394] But the AI litigation involves the current wave in which LLMs and diffusion models dominate.[395] The AI companies' use of large datasets containing copyrighted content reflects a practice that started in university research, where scaling up proved to yield greater advances with bigger datasets.[396] It continued at AI companies by AI researchers, some hired straight from academia. Courts in the AI litigation should not ignore this history: it explains the origin and reason for scaling up with large datasets, which included copyrighted content.

2. Factor 1: The Further Purpose and Different Character of Using Copyrighted Works to Research, Develop, Train, and Improve New AI Models by For-Profit Companies. Putting aside the important difference of commercial use,[397] the purpose of training AI models with copyrighted works is the same: to develop an AI model that works and improves the state of the art. The character of use is also the same: computational analysis of the datasets to facilitate deep learning of the model so that it can respond to human queries, including requests to generate new works. The technical reason AI developers used

391. *See* Sumit Singh, *Top Large Language Models for Writers, Developers, and Marketers: A Comprehensive Comparison*, LABELLER (Oct. 23, 2024), https://www.labellerr.com/blog/comparing-language-models-through-parameters-vs-real-life-experiments/ [https://perma.cc/9UFL-LACM]; Cameron Hashemi-Pour, *What Is the Google Gemini AI Model (Formerly Bard)?*, TECHTARGET (Jan. 8, 2025), https://www.techtarget.com/searchenterpriseai/definition/Google-Gemini [https://perma.cc/6YQY-CYCY]; *Introducing LLaMA: A Foundational, 65-Billion-Parameter Large Language Model*, META (Feb. 24, 2023), https://ai.meta.com/blog/large-language-model-llama-meta-ai/ [https://perma.cc/B3JQ-JC9Z].

392. Alan D. Thompson, *The Memo - Special Edition: Claude 3 Opus*, THE MEMO BY LIFEARCHITECT.AI (Mar. 4, 2024), https://lifearchitect.substack.com/p/the-memo-special-edition-claude-3 [https://perma.cc/NB4P-S93F].

393. *See* Gabriel Snyder, *Yann LeCun, Pioneer of AI, Thinks Today's LLM's Are Nearly Obsolete*, NEWSWEEK (Sep. 5, 2025, at 13:08 ET), https://www.newsweek.com/ai-impact-interview-yann-lecun-artificial-intelligence-2054237 [https://perma.cc/D3CA-ML6U].

394. *See* Dave Bergmann, *What Is Knowledge Distillation?*, IBM, https://www.ibm.com/think/topics/knowledge-distillation [https://perma.cc/95W3-62HK] (last visited July 14, 2025).

395. For example, the lawsuits against OpenAI, Meta, Anthropic, and Google involve LLMs, while the lawsuits against Stability AI, Midjourney, and other image generators involve diffusion models. *See Master List*, *supra* note 30; Hashemi-Pour, *supra* note 391.

396. *See supra* text accompanying 281, 294.

397. *See* discussion *supra* Part III. Granted, the companies' goals include the additional goal of offering an AI generator or app to the public, including potentially in a paid service.

copyrighted works in datasets to train the AI model has not changed from the university to AI companies. In both settings, the purpose of use is to develop a new AI model, and the character of use during the AI training is computational in deep learning.

The history of AI training illuminates the common edifice on which AI training with datasets consisting of copyrighted content lies.[398] If, as the plaintiffs in the copyright lawsuits contend, this purpose and character of AI training cannot be considered transformative or possessing a different purpose or character of use, then the same conclusion should follow for all university researchers as well.[399] That extreme result is certainly a possibility—which would likely slow down, if not stall, AI training at universities and AI companies in the United States.

This Article offers a better way. Courts should conclude that AI training using copyrighted works, whether by university-based researchers or AI companies, can serve a fair use purpose to develop a new technology with public benefits. That finding points in favor of fair use, but the courts must still conduct the rest of the fair use analysis, including the commercial use of the works by AI companies.

This aligns with *Warhol*'s use-by-use analysis. As shown in Table 6 below, we can delineate the two types of uses of copyrighted works potentially involved in the AI litigation—AI training uses and AI outputs.

Table 6. Potential Various Uses by AI Companies and Users of Its AI.

Entity	**Use of Copyrighted Works**
AI company	Use in Research, Developing, or Creating AI Model: Use datasets of many works, including copyrighted works, to train and improve AI models.
User of AI generator	Use in Any Outputs of AI generator: Any specific instance user of AI generator creates outputs that are regurgitations of copyright training materials or otherwise substantially similar.

In the first two district court decisions involving the training of generative AI models, Judges Alsup and Chhabria concluded,

398. *See supra* Sections IV.A.1, B.1 (setting forth the history of AI training).

399. *See, e.g.*, Plaintiffs' Reply to Motion for Partial Summary Judgment and Opposition to Meta's Motion for Partial Summary Judgment at 15–16, Kadrey v. Meta Platforms, Inc., No. 23-CV-03417-VC, 2025 WL 1752484 (N.D. Cal. June 25, 2025) (arguing "Meta's use of copyrighted works to train LLMs is not transformative"); Plaintiffs' Opposition to Anthropic's Motion for Summary Judgment at 14, Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, (N.D. Cal. June 23, 2025) (arguing "Anthropic's use of books in training is commercial and not transformative").

respectively, that Anthropic's and Meta's training of their models with copyrighted books of the plaintiffs was a highly transformative use—and ultimately a fair use.[400] Indeed, in *Bartz v. Anthropic*, Judge Alsup held that "the use of the books at issue to train Claude and its precursors was exceedingly transformative," even "spectacularly so," to develop a new technology that is "among the most transformative many of us will see in our lifetimes."[401] "Anthropic used copies of Authors' copyrighted works to iteratively map statistical relationships between every text-fragment and every sequence of text-fragments so that a completed LLM could receive new text inputs and return new text outputs as if it were a human reading prompts and writing responses."[402] In *Kadrey v. Meta*, Judge Chhabria reached the same conclusion that AI training is highly transformative in developing a new technology:

> There is no serious question that Meta's use of the plaintiffs' books had a "further purpose" and "different character" than the books—that it was highly transformative. The purpose of Meta's copying was to train its LLMs, which are innovative tools that can be used to generate diverse text and perform a wide range of functions. Users can ask Llama to edit an email they have written, translate an excerpt from or into a foreign language, write a skit based on a hypothetical scenario, or do any number of other tasks. The purpose of the plaintiffs' books, by contrast, is to be read for entertainment or education.[403]

The analysis of both Judge Alsup and Judge Chhabria aligns with my recommended approach for analyzing technological fair use in AI training, outlined above.[404] Notably, in both cases, the plaintiffs made no allegation that the respective AI model had produced any infringing outputs of their books.[405] (As discussed later, I disagree with other parts of their opinions—Judge Alsup's apparent

400. *See Bartz*, 2025 WL 1741691, at *6–8; *Kadrey*, 2025 WL 1752484, at *23.

401. *Bartz*, 2025 WL 1741691, at *5, *6–7, *18.

402. *Id.* at *7.

403. *Kadrey*, 2025 WL 1752484, at *9 (citing Google LLC v. Oracle Am., Inc. 593 U.S. 1, 30 (transformative to use copyrighted computer code "to create a new platform that could be readily used by programmers")).

404. *See supra* text accompanying notes 337–50.

405. *Bartz*, 2025 WL 1741691, at *4, *7; *Kadrey*, 2025 WL 1752484, at *7.

categorical view of acquiring pirated books and Judge Chhabria's broad view of market dilution as cognizable harm.[406])

C. How AI's Capability or Functionality to Generate Serves a Fair Use Purpose to Create New, Non-Infringing Works

1. AI's Generative Capability Serves a Fair Use Purpose. Once successfully trained, the generative AI model deployed in a generator will be able to "generate" new content, including text, images, videos, music, and computer programs.[407] That is one of the key innovations of generative AI. There may be disagreement over whether the AI's generative functionality should be analyzed under fair use for AI training, for specific outputs of AI that are allegedly infringing, or both.[408] In *Kadrey v. Meta*, Judge Chhabria asked the parties whether Meta's alleged transformative purpose is based on "what Llama is ultimately capable of producing" or, alternatively, based on training the model "without regard to what the language model ultimately enables people to produce after the copying" used in the training.[409] In its pre-publication report, the Copyright Office took the view that the defendant's use to develop AI is a distinct use that should still "be evaluated in the context of the overall use," such as in providing generators the capability to produce new content.[410]

I am not sure it matters where in the fair use analysis courts analyze AI's generative capability, provided they do so. The question under Factor 1 is the same: whether developing this generative capability of an AI model serves a "further purpose" than the copyright holders' purpose in creating or disseminating the works.

It does. The generative function of AI models serves the *further* purpose of creating a new technology capable of creating

406. *See infra* notes 640–55 and accompanying text (pirated books), 498–525 (market dilution); *see also Comparing Judge Alsup and Judge Chhabria's Fair Use Decisions in* Anthropic, Meta *Cases*, CHAT GPT IS EATING THE WORLD (June 26, 2025), https://chatgptiseatingtheworld.com/2025/06/26/12026/ [https://perma.cc/MH45-PF4F] (comparison of Judge Alsup's and Judge Chhabria's decisions on fair use).

407. *See* Cole Stryker & Mark Scapicchio, *What Is Generative AI?*, IBM, https://www.ibm.com/think/topics/generative-ai [https://perma.cc/QRS7-MQU8] (last visited July 29, 2025).

408. Or perhaps the generative functionality should be analyzed separately for fair use purposes, meaning a use-by-use approach could be applied to three separate uses: (1) training the AI model; (2) developing generative functionality that enables people to create works; and (3) specific outputs of the AI generator.

409. *Expressing Skepticism*, *supra* note 349.

410. PRE-PUBLICATION REPORT, *supra* note 40, at 36–38.

new, non-infringing works.[411] Society benefits from both the new technology and the new, non-infringing works of expression. That is especially so when the technology makes creative production more accessible to people.

2. Fair Use's Line Between Derivative Works and Non-Infringing Works. The Supreme Court has recognized that fair use "permits [and requires] 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."[412] Time and again, the Court has admonished that, in America, "[t]he primary objective of copyright is not to reward the labor of authors, but '[t]o promote the Progress of Science and useful Arts.'"[413] Fair use accommodates people's need to borrow and build on works that have come before, in order to create new works.[414]

To balance these interests, the *Warhol* Court drew an important line between fair uses and derivative works.[415] The Court cited Judge Leval's analysis in *Authors Guild v. Google, Inc.*[416] In that case involving Google's fair use copying of millions of books to create Google Book Search (summarized in Part II above[417]), the Second Circuit drew the same line between fair uses and derivative works:

411. *See* Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *7–8 (N.D. Cal. June 23, 2025) (Anthropic's Claude model served goal of progress because "the purpose and character of using copyrighted works to train LLMs to generate new text was quintessentially transformative").

412. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577 (1994); *accord* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1274 (2023).

413. *See* Feist Publ'ns., Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 349 (1991) (second alteration in original) (quoting U.S. CONST. art. I, § 8, cl. 8).

414. *See Campbell*, 510 U.S. at 575.

415. *Warhol*, 143 S. Ct. at 1275 ("To preserve that right, the degree of transformation required to make 'transformative' use of an original must go beyond that required to qualify as a derivative.").

416. *See id.* at 1276 (citing Authors Guild v. Google, Inc., 804 F.3d 202, 214 (2d Cir. 2015) (Leval, J.)).

417. *See supra* Section III.B.1.

> The more the appropriator is using the copied material for new, transformative purposes, the more it serves copyright's goal of enriching public knowledge and the less likely it is that *the appropriation will serve as a substitute for the original or its plausible derivatives*, shrinking the protected market opportunities of the copyrighted work . . . .
>
> . . . .
>
> . . . [*Campbell*] further explains that the "goal of copyright, to promote science and the arts, is generally furthered by the *creation of transformative works*" and that "[s]uch works thus lie at the heart of the fair use doctrine's guarantee of breathing space within the confines of copyright." In other words, transformative uses tend to favor a fair use finding because a transformative use is *one that communicates something new and different from the original* or expands its utility, thus serving copyright's overall objective of contributing to public knowledge.[418]

Thus, both the *Warhol* Court and the Second Circuit in *Author's Guild* trained their focus in analyzing Factor 1 by distinguishing between derivative works and fair uses to create transformative works. Both explanations of Factor 1 scrutinize whether a defendant's work contains some element(s) copied from the plaintiff's work that might make it a derivative work instead of a fair use, with a transformative purpose, that "communicates something new and different from the original."[419] Moreover, some new works are created without the need for fair use because they do not copy any copyrightable elements from other works—the *Warhol* Court mentioned this category of works that "use existing materials to make valuable new works" based on the many "escape valves" in copyright law.[420]

On a spectrum of potential uses of a copyrighted work, this latter category—i.e., uses to create new works without copying any protected elements from the original work—are likely to have a greater degree of being new and different from the original than works, such as a parody, that do copy protected elements.[421] Table 7

---

418. *Authors Guild*, 804 F.3d at 208, 214 (emphasis added) (second alteration in original) (citations omitted).

419. *See Warhol*, 143 S. Ct. at 1275; *Authors Guild*, 804 F.3d at 214–15.

420. *See Warhol*, 143 S. Ct. at 1287.

421. For example, 2 Live Crew's parody of Roy Orbison's song, "Oh, Pretty Woman," conjures up and copies elements from Orbison's song. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579, 580–82 (1994). By contrast, a non-parodic song that does not copy protected

below sets forth this range, going from the creation of a derivative work based on another work, to a parody fair use of a work, to a completely new, non-infringing work that does not copy any copyrightable or protected element at all. As the bottom row indicates, the last category presents a straightforward case of non-infringement: the new work does not copy any protected element from another work.

Table 7. Potential Uses of Copyrighted Works to Make New Works.

Derivative Work	**Parody Fair Use or Other Fair Use of Copyrighted Work**	**New, Non-Infringing Work with No Copyrightable Elements Copied**
Defendant copies copyrightable elements of work to make derivative work based on the original work.	Defendant copies copyrightable elements of work to make parody of the original work, or a secondary work with a further purpose or different character.	Defendant does not copy any copyrightable element from original work but creates a new and different work that is not substantially similar to original work.
Infringing if unauthorized.	Fair use potentially.	Entirely non-infringing.

The line the *Warhol* Court drew between derivative works and permissible fair uses to create new works is the line the courts should follow in the AI litigation.[422] In Table 7, this line divides the first two columns. Nothing in *Warhol* even remotely suggests that the scope of copyright can extend to restrict non-infringing works, as indicated by the third column in Table 7.[423] To the contrary, the Court's recitation of "escape valves" in copyright, including "the idea–expression

elements of another song, such as Ed Sheeran's "Thinking Out Loud," adds new expressive works not confined to parody an existing work. *See* Structured Asset Sales, LLC v. Sheeran, 120 F.3d 1066, 1079–81 (2d Cir. 2024) (all similarities between Sheeran's song and Marvin Gayle's "Let's Get It On" were unprotected elements); Ben Sisario, *Ed Sheeran Won His Copyright Trial. Here's What to Know*, N.Y. TIMES (May 4, 2023), https://www.nytimes.com/article/ed-sheeran-marvin-gaye-copyright-trial.html [https://perma.cc/8L9W-4UAY].

422. Other lower courts have drawn the same line between derivative works and non-infringing works in analyzing Factor 4. *See, e.g.*, Warner Bros. Ent. Inc. v. RDR Books, 575 F. Supp. 2d 513, 549–50 (S.D.NY. 2008) ("This testimony does not bear on the determination of the fourth factor, however, because a reference guide to the *Harry Potter* works is not a derivative work; competing with Rowling's planned encyclopedia is therefore permissible.").

423. *Warhol*, 143 S. Ct. at 1287.

distinction; the general rule that facts may not receive protection; the requirement of originality; the legal standard for actionable copying," and other doctrines, supports the conclusion that copyright law does not—and should not—restrict the creation of new, non-infringing works.[424] In short, the production of new, non-infringing works in the third column are akin to the "third rail" of copyright law—they should never be touched.

As Judge Alsup concluded in *Bartz v. Anthropic*, the capability of AI generators to produce new, non-infringing works is transformative and should weigh in favor of fair use because it promotes progress through the creation of non-infringing works for the public's benefit.[425] This conclusion comports with the Progress Clause: "the goal of copyright, to promote science and the arts, is generally furthered by the creation of transformative works."[426] Thus, the ability of people to create new, non-infringing works using AI generators should favor fair use. Regurgitations and production of substantially similar derivative works, on the other hand, may militate against fair use, on a use-by-use basis. But, as the *Sony* Court expounded, neither patent law nor copyright law can prohibit technologies capable of substantial non-infringing uses.[427]

3. Generative Capability Is a Method Beyond Copyright's Scope. More fundamentally, what generative AI models learn are *methods* for composing or creating new works—how to write essays and computer programs, how to create images, songs, and videos, and how to generate based on predictions the models make based on their weighting of discrete relationships within the vast data (or works) on which they were trained.[428] When people

424. *Id.*

425. *See* Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *7–8 (N.D. Cal. June 23, 2025).

426. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994); U.S. CONST. art. I, § 8, cl. 8.

427. Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 429 (1984).

428. *See* SUMMERFIELD, *supra* note 383, at 159 ("During training, the function that the LLM learns (encoded in its weights) to predict the next token contains information about the deep structure of grammar, of meaning, and of logic, which can be generalized to entirely new sequences of tokens—allowing it to give fluent and coherent outputs."); *What Is Generative AI?*, TEXAS A&M UNIV.: CTR FOR GEOSPATIAL SCIS., APPLICATIONS & TECH., https://geosat.tamu.edu/genai-overview/ [https://perma.cc/349R-HNB8] (last visited July 18, 2025) ("These AI models are constructed on artificial neural networks trained on extensive datasets, enabling them to learn and replicate data patterns and relationships. The process often begins with a prompt that the AI uses to autonomously generate relevant and coherent content based on its training. This capability to produce novel and creative outputs is seen as an emergent property of the AI's complex structure and its learning process, rather than through explicit

"prompt" AI generators to create something new, the AI generator is capable of doing so only because it has learned a *method* for such generation of content based on predictions.[429] Some generative capability of AI, such as writing computer code, was surprising to researchers—an emergent ability developed by the AI model.[430] And the ability of large language models to learn how to tackle new tasks sought in prompts, without further training, is called *in-context learning*.[431] In their paper on GPT-3, OpenAI researchers found that in-context learning improved with larger models containing more parameters from training.[432]

---

programming."); *What Is an AI Model?*, IBM, https://www.ibm.com/think/topics/ai-model [https://perma.cc/QP3J-W2HE] (last visited July 15, 2025) ("Essentially, when an LLM outputs text, it has computed a high probability of that sequence of words being assembled in response to the prompt it was given."); *How Does Generative AI Work?*, MICROSOFT, https://www.microsoft.com/en-us/ai/ai-101/how-does-generative-ai-work [https://perma.cc/6UR3-GFHK] (last visited July 18, 2025) ("GPT-3.5, with its 175 billion parameters, demonstrates an immense capacity to learn and generate sophisticated text, as each parameter contributes to the model's ability to recognize nuances in language and context, leading to more coherent and contextually relevant outputs."); *Generative Models*, *supra* note 360 ("The trick is that the neural networks we use as generative models have a number of parameters significantly smaller than the amount of data we train them on, so the models are forced to discover and efficiently internalize the essence of the data in order to generate it.").

429. *See* Julian Wallis, *What Is Generative AI? Exploring the Tech Behind This Modern Innovation*, INTUJI, https://intuji.com/what-is-generative-ai-exploring-the-tech/ [https://perma.cc/QJU9-KWCK] (last visited July 19, 2025) ("[T]he case of a text-based generative AI model involves finding a method to represent words as vectors that capture the similarity between frequently co-occurring words or words with similar meanings within sentences."); Ryan O'Connor, *Introduction to Diffusion Models for Machine Learning*, ASSEMBLYAI (May 12, 2022), https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction [https://perma.cc/GY24-EG3E] (describing how diffusion model "learns to reverse this diffusion process in order to generate new data" during training); *see also* Ling Yang et al., *Diffusion Models: A Comprehensive Survey of Methods and Applications*, *in* 56 ACM Computing Surveys 5–6, 8–10, 13–15 (2024), https://arxiv.org/pdf/2209.00796 [https://perma.cc/W4SM-XNJR] (discussing how various diffusion models are trained and operated).

430. *See* ANIL ANANTHASWAMAY, WHY MACHINES LEARN: THE ELEGANT MATH BEHIND MODERN AI 420 (2024); *see also* Stephen Ornes, *The Unpredictable Abilities Emerging from Large AI Models*, QUANTA MAG. (Mar. 16, 2023), https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/ [https://perma.cc/4TMD-P79Z] (discussing emergent abilities of AI models).

431. *See* SUMMERFIELD, *supra* note 383, at 159; Deval Shah, *What Is in-Context Learning, and How Does It Work: The Beginner's Guide*, LAKERA (Aug. 27, 2025), https://www.lakera.ai/blog/what-is-in-context-learning [https://perma.cc/V5H2-E9YW].

432. *See* Brown et al., *supra* note 367, at 24. Granted, AI skeptics deride AI as engaging in mimicry, no better than "stochastic parrots." *See* Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, *in* FACCT '21: PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 610–11, 617 (2021), https://dl.acm.org/doi/pdf/10.1145/3442188.3445922 [https://perma.cc/RG8J-HQB3] ("[A]n LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot."). It goes beyond this Article to wade into this debate, but the slight underestimates the learning capability of not only AI models, but also parrots. *See New Study Provides Evidence that Parrots Can Communicate Needs and Emotions with Humans*, THE

Copyright law provides no basis for copyright holders to attempt to stop AI's generative capability either directly or indirectly through the backdoor of negating fair use.[433] Since *Baker v. Selden*, it is a bedrock principle of copyright law that copyright does not protect methods, a principle now codified in § 102(b) of the Copyright Act.[434] As the Seventh Circuit admonished, "[c]opyright protects actual expression, not methods of expression."[435] Adherence to this principle requires courts to limit infringement claims to *copyrightable* expression.[436] Applying this principle, Judge Alsup correctly rejected the *Bartz* book authors' expansive argument that the AI model's learning of "grammar, composition, and style" from the training datasets was not fair use.[437] Copyright law does not protect these methods.[438] As Molly Shaffer Van Houweling explained, what copyright law leaves unprotected gives people "an affirmative freedom to extract those elements of copyrighted works to which protection does not extend."[439] Based on the "escape valves" that limit the scope of copyright, the *Warhol* Court envisioned "ample space for artists and other creators *to use existing materials to make valuable new works*."[440]

History is replete with examples in which new methods for creation disrupted traditional ones. As Justice Breyer remarked

---

OPEN UNIV. (Dec. 2, 2024), https://www.open.ac.uk/blogs/news/science-mct/new-study-provides-evidence-that-parrots-can-communicate-needs-and-emotions-with-humans/ [https://perma.cc/5J4Z-KG8W]; Schuyler Velasco, *She Taught Her Cockatoo to Read. That Was Just the Beginning*, NGN MAG. (Apr. 28, 2023), https://news.northeastern.edu/2023/04/28/magazine/cockatoo-parrot-reading/ [https://perma.cc/Y9ZC-L9GK].

433. *See generally* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984) (rejecting copyright holders' "unprecedented" copyright claim to stop sale of video tape recorders (VTRs), a technology capable of substantial non-infringing uses, because "there is no basis in the Copyright Act upon which respondents can hold petitioners liable for distributing VTR's to the general public"); Edward Lee, *Freedom of the Press 2.0*, 42 GA. L. REV. 309, 380–83 (2008) (explaining how the *Sony* safe harbor's protection for technologies capable of substantial non-infringing uses serves as a First Amendment safeguard within copyright law consistent the constitutional protection for speech-facilitating technologies under the Free Press Clause).

434. Baker v. Selden, 101 U.S. 99, 103–04 (1879) ("The very object of publishing a book on science or the useful arts is to communicate to the world the useful knowledge which it contains. But this object would be frustrated if the knowledge could not be used without incurring the guilt of piracy of the book."); 17 U.S.C. § 102(b).

435. Peters v. West, 692 F.3d 629, 636 (7th Cir. 2012).

436. *See id.*

437. Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *8 (N.D. Cal. June 23, 2025).

438. *See id.*

439. Molly Shaffer Van Houweling, *The Freedom to Extract in Copyright Law*, 103 N.C. L. REV. 445, 456 (2025).

440. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1287 (2023) (emphasis added).

during oral argument in *Grokster*, "the monks [who were scribes] had a fit when Gutenberg made his press."[441] Likewise, the painter Paul Delaroche lamented the advent of photography: "From today, painting is dead!"[442] Similar concerns were raised by digital audio workstations (DAWs), with the fear that software-based music production would replace human artists.[443] As Donny Osmond explained, people feared DAWs would displace real musicians, but, today, DAWs are used by top musicians, including Dua Lipa, Bruno Mars, Harry Styles, and Justin Bieber.[444]

The advent of these technologies provided general competition between new and existing methods of creation, which even included the potential job displacement of some creators, such as painters and musicians.[445] But copyright law has refrained from interfering with this general competition in the market. Granting copyright holders the ability to stop AI's generative function, even though it creates new, non-infringing works, would expand the limited monopoly of copyright far beyond its permissible bounds. Indeed, it would be unconstitutional because Congress only has the Article I power to grant exclusive rights to authors to "their respective Writings," but not beyond.[446]

4. General Competition v. Specific Substitution of Works. A corollary to this fundamental principle is that copyright is limited to infringement of *specific works*—and does not give any right to copyright holders to monopolize methods for creation or to prevent general competition in the market presented by non-infringing but

441. *See* Peter K. Yu, *Of Monks, Medieval Scribes, and Middlemen*, 2006 MICH. ST. L. REV. 1, 2–3.

442. *See* George Dillard, *"From Today, Painting Is Dead"—What the Invention of Photography Tells Us About AI*, MEDIUM (May 17, 2023), https://worldhistory.medium.com/from-today-painting-is-dead-what-the-invention-of-photography-tells-us-about-ai-1c53900e613 [https://perma.cc/Z44X-PNXK].

443. *See* Ashley Hume & Larry Fink, *Donny Osmond Says AI Is A 'Tool, Not a Substitute'*, FOX NEWS (Sep. 20, 2024, at 02:00 ET), https://www.foxnews.com/entertainment/donny-osmond-says-artificial-intelligence-tool-not-substitute [https://perma.cc/AVF9-Z3WP].

444. *Id.*

445. *See* Kristofer Erickson, *AI and Work in the Creative Industries: Digital Continuity or Discontinuity?*, CREATIVE INDUS. J., Oct. 2024, at 1, 3–5, 7.

446. U.S. CONST. art. I, § 8, cl. 8.

competing works.[447] Therefore, in analyzing fair use, it is important for courts to distinguish between:

> (i) general competition in the market from different methods of production and among non-infringing works that do not contain substantially similar copies of others' works and
>
> (ii) specific substitution between two works, one of which includes copyrightable expression copied from the other.

Fair use is concerned only with specific substitution from a copied work: where the defendant's work copies and incorporates the plaintiff's work into a substitute work, it is typically not fair use.[448] But general competition from works that do not copy any protected elements of the plaintiff's work falls completely outside of fair use's and copyright's concern.[449]

As Judge Alsup explained in rejecting the market dilution theory in *Bartz*:

> Authors contend generically that training LLMs will result in an explosion of works competing with their works — such as by creating alternative summaries of factual events, alternative examples of compelling writing about fictional events, and so on. This order assumes that is so. But Authors' complaint is no different than it would be if they complained that training schoolchildren to write well would result in an explosion of competing works. This is not the kind of competitive or creative displacement that concerns the Copyright Act. The Act seeks to

447. *See* 17 U.S.C. § 106 (exclusive rights of copyright all limited to "the copyrighted work"); *id.* § 102(b) (excluding ideas and methods from copyright protection); *id.* § 107 (fair use doctrine's consideration of market harm to "the copyrighted work," where defendant made use of "the copyrighted work"); *see also* Tori Noble, Mitch Stoltz & Corynne McSherry, *The U.S. Copyright Office's Draft Report on AI Training Errs on Fair Use*, TECHDIRT (May 21, 2025, at 12:46 CT), https://www.techdirt.com/2025/05/21/the-u-s-copyright-offices-draft-report-on-ai-training-errs-on-fair-use/ [https://perma.cc/QB55-EQ4S] ("Traditionally, the fair use analysis requires courts to consider the effects of the use on the market for *the work in question*.").

448. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1263 (2023) ("An independent justification like this is particularly relevant to assessing fair use where an original work and copying use share the same or highly similar purposes, or where wide dissemination of a secondary work would otherwise run the risk of substitution for the original or licensed derivatives of it.").

449. Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *17 (N.D. Cal. June 23, 2025).

> advance original works of authorship, *not to protect authors against competition*.[450]

Similarly, in *Sega*, the Ninth Circuit concluded the defendant's creation of a new, non-infringing game—which the plaintiff called a "competing product"—ultimately served the purpose of copyright:

> [It] led to an increase in the number of independently designed video game programs offered for use with the Genesis console. *It is precisely this growth in creative expression*, based on the dissemination of other creative works and the unprotected ideas contained in those works, that the Copyright Act was intended to promote.[451]

The Ninth Circuit even extolled competition in the marketplace: "[A]n attempt to monopolize the market by making it impossible for others to compete runs counter to the statutory purpose of promoting creative expression and cannot constitute a strong equitable basis for resisting the invocation of the fair use doctrine."[452] The court found the defendant's reverse engineering of copies of the plaintiff's operating system to be a fair use, in part because it "facilitat[ed] the entry of a new competitor" that created non-infringing works.[453]

Similarly, applying the teaching of *Sega*, the Ninth Circuit held that the defendant Connectix's reverse engineering of Sony's copyrighted operating system for its PlayStation console so that Connectix could create an emulator program that was compatible to operate PlayStation games on Apple computers instead of Sony's console was a fair use.[454] During the reverse engineering, Connectix created many intermediate copies of Sony's operating system to figure out how to emulate it, but "none of the Sony copyrighted material was copied into, or appeared in, *Connectix's final product*, the Virtual Game Station."[455] Put simply, Connectix's final program was itself entirely non-infringing.

---

450. *Id.* (emphasis added) (citation omitted); *but see* Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *2, *17–18 (N.D. Cal. June 25, 2025) (disagreeing with Judge Alsup's analysis in *Bartz* and concluding that market dilution based on non-infringing works should be cognizable harm under Factor 4).

451. Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1522–23 (9th Cir. 1992) (emphasis added).

452. *Id.* at 1523–24 (emphasis added).

453. *Id.* at 1514, 1523–24 (emphasis added).

454. Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 601–02 (9th Cir. 2000).

455. *Id.* at 600 (emphasis added).

Sony objected that it would suffer market harm from Connectix's competing program (since consumers didn't need to buy Sony's console to play Sony games), but the Ninth Circuit held that such economic loss was outside the scope of copyright:

> The district court found that "[t]o the extent that such a substitution [of Connectix's Virtual Game Station for Sony PlayStation console] occurs, Sony will lose console sales and profits." *We recognize that this may be so.* But because the Virtual Game Station is transformative, and does not merely supplant the PlayStation console, the Virtual Game Station is a legitimate competitor in the market for platforms on which Sony and Sony-licensed games can be played. *For this reason, some economic loss by Sony as a result of this competition does not compel a finding of no fair use*. Sony understandably seeks control over the market for devices that play games Sony produces or licenses. *The copyright law, however, does not confer such a monopoly.* This factor favors Connectix.[456]

By contrast, specific competition between two works, one of which includes copyrightable expression copied from the other, is potentially substitution of the copyright holder's work under Factor 1 of fair use and market harm to the work under Factor 4. As Judge Chhabria ruled in the book authors' lawsuit against Meta, for the plaintiffs to prove infringement in the outputs of an AI generator, they must show it generates "actual copies of their protected works" or "are similar enough to the plaintiffs' books to be infringing derivative works."[457] If the plaintiffs identify allegedly infringing outputs of the AI generator, the outputs raise concerns of potential substitution and market harm cognizable under copyright law. But this issue should be analyzed separate from training copies, as

456. *Id.* at 607–08 (alterations in original) (emphasis added) (citations omitted) (citing *Sega*, 977 F.2d at 1522–24, "[A]n attempt to monopolize the market by making it impossible for others to compete runs counter to the *statutory purpose of promoting creative expression* and cannot constitute a strong equitable basis for resisting the invocation of the fair use doctrine.").

457. Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2023 WL 8039640, at *1 (N.D. Cal. Nov. 20, 2023); *see also* VMG Salsoul, LLC v. Ciccone, 824 F.3d 871, 878 (9th Cir. 2016) ("A 'use is de minimis only if the average audience would not recognize the appropriation.'" (quoting Newton v. Diamond, 388 F.3d 1189, 1193 (9th Cir. 2004)).

discussed next.[458] The plaintiffs need to prove such a scenario with evidence of specific works copied, not mere speculation.[459]

The line between general competition, including in the methods for creation, and specific substitution by infringing works reflects the *Sony* safe harbor. Copyright holders have no basis to stop substantial non-infringing uses of a technology.[460] AI generators offer the ability for people with disabilities, such as blindness or lack of voice, to create new, non-infringing works.[461] That increase in accessibility and creation of non-infringing works serves the goal of the Progress Clause.[462] The market harm produced by a transformative work is typically not "a harm cognizable under the Copyright Act."[463] That is especially so where the plaintiffs point to no infringing elements in specific works produced by AI generators.[464] The public has an "interest in access to" technologies with substantial non-infringing uses, including ones that increase access to creative production such as to people with disabilities.[465] Indeed, fostering such technologies serves the "ultimate aim [of the Copyright Act], to stimulate artistic

458. Granted, there might be extreme scenarios in which an AI generator was so poorly trained or lacked any guardrails that it routinely produces—at consumers' requests—vast amounts of infringing content, thereby undermining the asserted training purpose. In that egregious scenario, the fair use defense might fail due, not to the lack of a fair use purpose in AI training, but instead due to the shoddy implementation that resulted in a substantial amount of copies under Factor 3 of fair use. *Cf.* Warner Bros. Ent. Inc. v. RDR Books, 575 F. Supp. 513, 548–49 (S.D.N.Y. 2008) (Harry Potter lexicon had a fair use purpose as a reference guide but was not a fair use because too much of underlying works were copied, "more than is reasonably necessary to create a reference guide").

459. *Cf.* Authors Guild v. Google, Inc., 804 F.3d 202, 227 (2d Cir. 2015) (rejecting argument that copies of books stored by Google internally are prone to hacking).

460. *Id.* at 212.

461. *See* Edward Lee, *Prompting Progress: Authorship in the Age of AI*, 76 FLA. L. REV. 1445, 1560–61 (2024).

462. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 575, 579 (1994).

463. *Id.* at 591–92; *see Warner Bros.*, 575 F. Supp. at 550 (stating that the use of parts of *Harry Potter* books for reference in lexicon was transformative and therefore author of *Harry Potter* cannot control who enters that reference market).

464. *See* Tremblay v. OpenAI Inc., 716 F. Supp 3d 772, 778 (N.D. Cal. 2024) (granting dismissal of copyright claims where plaintiffs alleged "every output" was infringing but "fail[ed] to explain what the outputs entail or allege that any particular output is substantially similar – or similar at all – to their books.").

465. Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 440 (1984); *see* Laurie Henneborn, *Designing Generative AI To Work for People with Disabilities*, HARV. BUS. REV. (Aug. 18, 2023), https://hbr.org/2023/08/designing-generative-ai-to-work-for-people-with-disabilities [https://perma.cc/9AYF-6WRQ]; Lee, *supra* note 134, at 201 ("When a new technology expands accessibility and the very notion of who can create, society benefits from the growth of creative works *and* from the expansion of who can create to include people with disabilities and other historically underrepresented groups.").

creativity for the general public good."[466] In our free market system and the marketplace of ideas, people and companies alike must embrace competition, not attempt to stifle it.[467]

5. The Copyright Office's "Uncharted" Theory of Copyright Dilution Is Unconstitutional. The Copyright Office's pre-publication report took a negative view of AI's generative capability, even when used to produce new, non-infringing works.[468] Instead of this creation of new, non-infringing expression weighing in favor of fair use, AI's generative capability should diminish the degree of transformativeness of the AI model.[469] The Copyright Office concluded: "Where a model is trained on *specific types of works* in order to produce content that shares the purpose of appealing to a particular audience, that use is, at best, modestly transformative."[470]

While conceding it was embarking on "uncharted territory," the Copyright Office doubled down on its unbounded view of the scope of copyrights by endorsing a new and untested theory of market dilution for copyright under Factor 4 of fair use.[471] Dilution is a trademark claim for famous marks under the Lanham Act.[472] It is not a concept recognized under the Copyright Act—or even proposed in copyright scholarship prior to the Copyright Office's report. The Copyright Office admitted it was in "uncharted territory" because, in the more than two centuries of copyright law, no fair use decision has ever recognized copyright dilution, much less a cognizable market harm from the creation of purely non-infringing works.[473]

Instead, the fair use cases recognize market harm under Factor 4 only where the defendant's *copy* of the plaintiff's work serves as a

---

466. *Sony*, 464 U.S. at 432 (quoting Twentieth Century Music Corp. v. Aiken, 422 U.S. 151, 156 (1975)).

467. *See* Nat'l Soc'y of Pro. Eng'rs v. United States, 435 U.S. 679, 695 (1978) ("The assumption that competition is the best method of allocating resources in a free market recognizes that all elements of a bargain—quality, service, safety, and durability—and not just the immediate cost, are favorably affected by the free opportunity to select among alternative offers."); Standard Oil Co. v. Fed. Trade Comm'n, 340 U.S. 231, 248 (1951) ("The heart of our national economic policy long has been faith in the value of competition.").

468. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 46 ("The use of a model may share the purpose and character of the underlying copyrighted works without producing substantially similar content. Where a model is trained on specific types of works in order to produce content that shares the purpose of appealing to a particular audience, that use is, at best, modestly transformative.").

469. *See id.*

470. *Id.* (emphasis added).

471. *See id.* at 64–65 nn.367–70, 375.

472. 15 U.S.C. § 1125(c).

473. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 65; Tori Noble et al., *supra* note 447.

potential substitute for the work and thereby poses a *cognizable* market harm.[474] As Judge Leval explained in *Authors Guild*, the concept of market substitution is "the likelihood that potential purchasers may opt to acquire *the copy in preference to the original*."[475] That is shown where the secondary use "results in widespread revelation of *sufficiently significant portions of the original* as to make available a significantly competing substitute."[476] The *Warhol* Court agreed with Judge Leval's analysis of market substitution of "matter protected by the [copyright owner's] interests in the original wor[k] or derivatives of [it]."[477] Under this principle from *Authors Guild*, if an AI-generated work does not infringe copy a plaintiff's work, it does not offer the public any chance of "acquir[ing] the copy in preference to the original."[478] The reason is simple: the AI-generated work is not even a "copy" of or substantially similar to the plaintiff's work.

But, ignoring this line of precedent, the Copyright Office endorsed the new theory of copyright dilution advanced by UMG Recordings, a plaintiff in two copyright lawsuits against AI companies, and other major copyright stakeholders.[479] Under this new theory of copyright dilution, Factor 4 of fair use should recognize market harm from competition putatively created by non-infringing AI-generated content, even though it is not "substantially similar to a specific copyrighted work."[480]

To borrow Justice Scalia's apt phrase in an analogous case, the Copyright Office's new theory of copyright dilution is "a species of mutant copyright law": it misuses a trademark concept to

474. *See, e.g.*, Ringgold v. Black Ent. Television, Inc., 126 F.3d 70, 79 (2d Cir. 1997) ("When all or a substantial portion of text that contains protectable expression is included in another work, solely to convey the original text to the reader without adding any comment or criticism, the second work may be said to have supplanted the original because a reader of the second work has little reason to buy a copy of the original."); Peter Letterese & Assocs., Inc. v. World Inst. of Scientology Enters., Int'l, 533 F.3d 1287, 1317–18 (11th Cir. 2008) ("The unrestricted and widespread dissemination of the [defendant's work] Sales Course—a use that is not transformative of the book and may be regarded as appropriating 'the heart' of its expression—namely, the selection and structure of sales techniques and distinctive descriptions thereof—may well usurp the potential market for [the plaintiff's work] *Big League Sales* and derivative works."); Ty, Inc. v. Publ'ns Int'l Ltd., 292 F.3d 512, 518 (7th Cir. 2002) (explaining why burlesque versions of past characters and stories, such as Frankenstein and Dracula, would not be fair use).

475. Authors Guild v. Google, Inc., 804 F.3d 202, 223 (2d Cir. 2015) (emphasis added); *accord* Fox News Network, LLC v. TVeyes, Inc., 883 F.3d 169, 179 (2d Cir. 2018).

476. *Authors Guild*, 804 F.3d at 223 (emphasis added).

477. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1276 (2023) (alterations in original) (citing *Authors Guild*, 804 F.3d at 207).

478. *Authors Guild*, 804 F.3d at 223.

479. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 65 n.375.

480. *See id.* at 64–65.

protect copyrighted works.[481] The Copyright Office's unbounded view of copyright dilution stretches copyright to entire genres and styles represented in the works contained in the training data. Thus, according to the Copyright Office's illogic, if an AI model was trained on romance novels, the copyright holder gets to stop the AI's capability from generating new, non-infringing romance novels because the new, non-infringing works might "dilut[e]" the market for the original novels and lead to "more competition for sales of an author's works and more difficulty for audiences in finding them."[482] The Copyright Office concluded the same with all AI-generated music—and even relied on the comments from UMG, a plaintiff in two pending copyright lawsuits against Udio and Suno.[483] The same illogic of the Copyright Office applied to imitation of styles from works in the training data, even when the AI-generated content is non-infringing and "not substantially similar to a specific underlying work."[484]

In a matter of two short paragraphs—which cited no legal precedent or even secondary authority in support of copyright dilution—the Copyright Office expanded the scope of copyright to now cover entire genres, uncopyrightable styles, and general competition a copyright holder faces from non-infringing works.[485] The Copyright Office gathered no evidence to substantiate its fears of copyright dilution but simply relied on hearsay statements of interested stakeholders that would not be admissible in a court of law.[486] Had the Copyright Office performed any semblance of evidence-gathering, it would have found that the markets for its two examples of copyright dilution—romance novels and music—are booming today.[487] That

481. Dastar Corp. v. Twentieth Century Fox Film Corp., 539 U.S. 23, 34 (2003).

482. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 65.

483. *Id.*; Shani Rivaux, Macarena Ferreira Fink & Catherine Perez, *Legal Riffs: Music Industry Alleges AI Is Out of Tune*, PILLSBURG (July 10, 2024), https://www.pillsburylaw.com/en/news-and-insights/record-labels-lawsuit-copyright-infringement-generative-ai-music.html [https://perma.cc/M85T-Y5FL].

484. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 66.

485. *Id.* at 65–66.

486. *See id.* at 64–66 (reviewing various comments submitted to Copyright Office); *see also* Smith v. Arizona, 144 S. Ct. 1785, 1792 (2024) (Confrontation Clause "bars only the introduction of hearsay—meaning, out-of-court statements offered 'to prove the truth of the matter asserted.'"); FED. R. EVID. 801–02 (defining and prohibiting hearsay from use as evidence absent an exception).

487. *See* Julia Rittenberg, *Are We Having a Romance Renaissance?*, BOOK RIOT (Mar. 27, 2025), https://bookriot.com/are-we-having-a-romance-renaissance/ [https://perma.cc/3GXQ-KUQD]; *Music Revenues Rise Again in 2024, Boosted by Streaming Subscriptions, Report Shows*, REUTERS (Mar. 19, 2025, at 13:52 CT), https://www.reuters.com/business/media-telecom/music-revenues-rise-again-2024-boosted-by-streaming-subscriptions-report-shows-2025-03-19/ [https://p

includes UMG, which floated the idea of copyright dilution: UMG's revenues were $3.31 billion in the first quarter of 2025, up 9.5%, beating analysts' expectations.[488] And beyond parroting the arguments of interested copyright stakeholders, the only shred of legal support the Copyright Office offered for its newfound theory was that the fair use provision "encompasses *any 'effect'* upon the potential market," citing the Supreme Court's decision in *Campbell*.[489]

But the Copyright Office misread *Campbell* and the fair use provision. Section 107 does not say "any," much less "every" or "all" effects; it simply states "the effect."[490] Interpreting this term, the Supreme Court and a long line of precedent, including *Sega* and *Connectix*, recognize that many effects from transformative uses of copyrighted works are not cognizable market harms under Factor 4.[491] Indeed, *Campbell* itself stated: "when a lethal parody, like a scathing theater review, *kills demand for the original, it does not produce a harm cognizable* under the Copyright Act."[492] Thus, in analyzing the potential market effects from the defendant's use of a copyrighted work, courts examine, as a threshold matter, whether the putative market effect even falls within the scope of copyright's cognizable harms, taking into account how transformative the defendant's purpose was.[493] The Copyright Office ignored this essential analysis in misreading both § 107 and *Campbell*. If dramatically expanding the scope of famous trademarks to encompass trademark dilution required an act of Congress, then, at a minimum, an act of Congress should be required to expand the scope of copyright to encompass copyright dilution as well.[494]

---

erma.cc/DNT5-9L7W]; Elizabeth Dilts Marshall, *Sony Music Posts Record Profits, Fueled by Double-Digit Streaming and Publishing Growth*, BILLBOARD (May 14, 2025), https://www.billboard.com/pro/sony-music-fiscal-earnings-revenue-profits-streaming/ [https://perma.cc/9PYP-DP58].

488. Mauro Orru, *Universal Music's Revenue from Subscriptions, Streaming Beats Forecasts*, WALL ST. J. (Apr. 29, 2025, at 12:36 ET), https://www.wsj.com/business/earnings/universal-musics-subscriptions-streaming-revenue-beats-forecasts-2d7c4d81?st=tQA66p&reflink=desktopwebshare_permalink [https://perma.cc/X9PS-Z8W9].

489. PRE-PUBLICATION REPORT, *supra* note 40, at 65 (emphasis added) (citing 17 U.S.C. § 107; Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 590 (1994)).

490. 17 U.S.C. § 107.

491. *See, e.g.*, Sony Comput. Ent., Inc. v. Connectix Corp., 203 F.3d 596, 607 (9th Cir. 2000); Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1524, 1527–28 (9th Cir. 1992).

492. *Campbell*, 510 U.S. at 591–92 (emphasis added).

493. *Id.* at 591 ("[W]hen, on the contrary, the second use is transformative, market substitution is at least less certain, and market harm may not be so readily inferred.").

494. *See* Erin J. Roth & Robert B. Bennett, Jr., *The Federal Trademark Dilution Act: Potent Weapon or Uphill Battle*, 16 MIDWEST L.J. 1, 7–11 (1999) (history of Congress's enactment of dilution protection for famous marks in Lanham Act in 1995).

In *Kadrey v. Meta*, Judge Chhabria took an approach somewhat similar to the Copyright Office's dilution theory, albeit without discussing it.[495] Unlike the Copyright Office's report, Judge Chhabria found the generative capability of Meta's AI model to be "highly transformative": "The purpose of Meta's copying was to train its LLMs, which are innovative tools that can be used to generate diverse text and perform a wide range of functions."[496] At least under Factor 1 of fair use, Judge Chhabria ruled that the AI model's ability to mimic styles in the training works did not diminish the transformative purpose, given that "style is not copyrightable—only expression is."[497] But, like the Copyright Office's report, Judge Chhabria supported recognizing a new theory of cognizable harm under Factor 4 of fair use based on market dilution that putatively results from non-infringing works of similar genres contained in the training works.[498] Indeed, despite the *Kadrey* plaintiffs failing to sufficiently raise this theory to survive summary judgment, Judge Chhabria expressed his virtually categorical view[499] that, "in most cases," AI training on copyrighted works will be illegal and not fair use due to market dilution, "[n]o matter how transformative LLM training may be."[500]

Judge Chhabria's opinion shows what's at stake. If market dilution is accepted by courts, fair use in AI training will fail "in most cases."[501] As he elaborated: "The upshot is that in many circumstances it will be illegal to copy copyright-protected works to train generative AI models without permission. Which means that the companies, to avoid liability for copyright infringement, will generally need to pay copyright holders for the right to use their materials."[502]

The whole concept of copyright market dilution is deeply flawed. For starters, it stretches Factor 4 of fair use well beyond

---

495. Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *18 (N.D. Cal. June 25, 2025).

496. *Id.* at *9.

497. *Id.* at *10.

498. *Id.* at *18.

499. *Id.* at *1. Judge Chhabria's extensive discussion of market dilution was dicta because he ultimately ruled that the plaintiffs had failed even to raise a genuine issue of material fact to advance this theory of dilution. *See id.* at *19.

500. *Id.* at *1, *18, *23 ("Indeed, it seems likely that market dilution will often cause plaintiffs to decisively win the fourth factor—and thus win the fair use question overall—in cases like this.").

501. *Id.* at *1, *23.

502. *Id.* at *2.

market harm caused by *specific substitution* from works that *copied protected elements* from the plaintiffs' works.[503] In other words, as summarized in Table 7 above, market dilution seeks to extend the scope of copyright to reach the "third rail" of copyright—new, non-infringing works that do not copy protected elements from any prior works.[504] Such a broad approach runs roughshod over the idea–expression dichotomy and uncopyrightability of methods, and destroys the "escape valves" the *Warhol* Court championed as necessary to "provide ample space for artists and other creators to use existing materials to make valuable new works."[505]

Copyright dilution cannot be squared with case law.[506] Take *Sega*. Imagine that other competitors did what Accolade did and created non-infringing video games for Sega's console that obliterated the market for video games, shrinking Sega's revenues.[507] For a court to treat such revenue loss by the copyright holder from general competition posed by non-infringing works would expand the limited monopoly of copyright to a general monopoly against all competition posed by non-infringing works. Or take Picasso, who was roundly criticized for "appropriat[ing]" elements from other artists' works.[508] Given his stature and success, any work of Picasso could easily dilute, if not obliterate, the market for art from which Picasso borrowed some elements. But, unless any borrowing resulted in a substantially similar or infringing work, the creation of new works by Picasso should be extolled, not condemned, even if Picasso diluted the market for the original works from which he borrowed.[509]

503. *See id*. at *16.

504. *See supra* notes 422–25 and accompanying text.

505. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1287 (2023).

506. Both the Copyright Office and Judge Chhabria acknowledged the theory of market dilution is new for copyright law. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 66; *Kadrey*, 2025 WL 1752484, at *18 ("In arguing that this sort of harm doesn't count just because it's never made a difference in a case before, Meta makes the mistake the Supreme Court instructs parties and courts to avoid: robotically applying concepts from previous cases without stepping back to consider context.").

507. *Cf.* Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1527 (9th Cir. 1992) (discussing competition from Accolade's non-infringing video game).

508. *See* Timothy Anglin Burgard, *Picasso and Appropriation*, 73 ART BULL. 479, 487 (1991).

509. Artists borrowing from other artists was not unique to Picasso; Matisse borrowed from Van Gogh, to mention one other example. *See, e.g.*, Martin Bailey, *Matisse Wanted To Buy a Van Gogh Portrait—Instead, His Brother Bought a Bicycle*, ART NEWSPAPER (Mar. 21, 2025), https://www.theartnewspaper.com/2025/03/21/matisse-wanted-to-buy-a-van-gogh-portrait-but-instead-the-money-was-spent-on-a-bicycle [https://perma.cc/FQT6-BEM3].

Next consider James Joyce's *Ulysses*.[510] Joyce wasn't the first author to use the literary style of stream of consciousness, but he is one of its most celebrated authors.[511] Under the Copyright Office's unbounded view of copyright dilution, Joyce would have been an infringer for borrowing this literary style from other authors, including the French writer Édouard Dujardin.[512] Why? To borrow the Copyright Office's own reasoning, "Even when the output is not substantially similar to a specific underlying work, stylistic imitation made possible by its use in training may impact the creator's market."[513] "Joyce picked up a copy of Dujardin's novel . . . in 1903" and "acknowledge[d] a certain borrowing from it."[514] And William Faulkner did the same, borrowing from Joyce's *Ulysses*.[515]

Contrary to the Copyright Office's and Judge Chhabria's unbounded view of copyright dilution, none of this borrowing of literary style weighs against fair use absent proof the defendant's work is a substantially similar copy. As Justice Story, an early articulator the fair use doctrine,[516] admonished in a passage approved by the Supreme Court in *Campbell*: "Every book in literature, science and art, borrows, and must necessarily borrow, and use much which was well known and used before."[517] Indeed,

> Virgil borrowed much from Homer; Bacon drew from earlier as well as contemporary minds; Coke exhausted all the known learning of his profession;

---

510. *See generally* George O'Brien, *Joyce's* Ulysses *Redefines Modern Fiction*, EBSCO (2023), https://www.ebsco.com/research-starters/literature-and-writing/joyces-ulysses-redefines-modern-fiction [https://perma.cc/JKY9-8ZQS] (explaining the history of *Ulysses* and its cultural significance, including its importance in the development of stream of consciousness writing).

511. *See* Melanie M. Steele*, The Pinnacle Development of the Stream-of-Consciousness Technique as Demonstrated in James Joyce's* Ulysses, 3 UNIV. WIS.-SUPERIOR MCNAIR SCHOLARS J. 155, 156 (2002) ("The stream-of-consciousness technique was first used as far back as the eighteenth century, was experimented with by several writers throughout the nineteenth century, was made to embody several emerging characteristics near the turn of the century, and was then brought to its highest point of development by James Joyce in his 1922 novel *Ulysses*, after which it became accepted and widely used in modern literature."); Liz Delf, *What Is Stream of Consciousness?*, OR. STATE UNIV. (Nov. 12, 2019), https://liberalarts.oregonstate.edu/wlf/what-stream-consciousness [https://perma.cc/W6FQ-VHKA].

512. *See* Édouard Dujardin, BRITANNICA, https://www.britannica.com/biography/Edouard-Dujardin [https://perma.cc/6JFJ-9NDK] (last visited Aug. 13, 2025).

513. PRE-PUBLICATION REPORT, *supra* note 40, at 66.

514. RANDELL STEVENSON, MODERNIST FICTION: AN INTRODUCTION 227 n.14 (1992).

515. *See* Edwin Turner, *The Sound and the Fury—William Faulkner*, BIBLIOKLEPT (Mar. 6, 2009), https://biblioklept.org/2009/03/06/the-sound-and-the-fury-william-faulkner/ [https://perma.cc/DH8S-GZ3B].

516. *See* Folsom v. Marsh, 9 F. Cas. 342, 344–45 (C.C.D. Mass. 1841) (No. 4,901); Ned Snow, *The Forgotten Right of Fair Use*, 62 CASE W. RSRV. L. REV. 135, 145 (2011).

517. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 575 (1994) (quoting Emerson v. Davies, 8 F. Cas. 615, 619 (No. 4,436) (C.C.D. Mass. 1845)).

> and even Shakespeare and Milton, so justly and proudly our boast as the brightest originals would be found to have gathered much from the abundant stores of current knowledge and classical studies in their days.[518]

The creation of new works should be extolled if they borrowed from past works without producing an infringing, or substantially similar, copy. As Justice Story further explained, "A copy is one thing, an imitation or resemblance another. There are many imitations of Homer in the Aeneid; but no one would say that the one was a copy from the other."[519]

Digital photography exposes the fallacy of copyright dilution. Digital photographs, spurred by the advent of smartphones, have increased exponentially, with an estimated 1.5 trillion photographs taken in 2022.[520] The widespread availability of digital cameras capable of rendering amazing image quality at the touch of a button has made it more challenging for professional photographers because, as one photographer put it, "[l]et's be honest: everyone's a photographer now."[521] The low barriers to entry in becoming a photographer translates into much greater competition.[522] And as AI is now integrated into cameras and editing tools, the capabilities for ordinary individuals to take high-quality photographs are widely available.[523] If we extended the logic of the Copyright Office's and Judge Chhabria's theory of market dilution to this context, the widespread availability of digital cameras would be bad for society. The technology flooded the market for photographs, making it difficult for professional photographers to compete. The value of some style of photographs likely decreased because everyone with a smartphone can now take an amazing headshot or portrait. The Copyright Office's and Judge Chhabria's theory of market dilution would view this democratization of digital cameras as bad because it "flooded" the market with trillions of photographs.

---

518. *Emerson*, 8 F. Cas. at 619.

519. *Id.* at 622.

520. *See* Susan Enfield, *How Many Photos Will Be Taken in 2022?*, MYLIO, https://news.mylio.com/how-many-photos-taken-in-2022/ (last visited July 20, 2025) [https://perma.cc/EBB6-NVJD].

521. Will Moneymaker, *The Real Struggles Photographers Face Daily*, MONEYMAKER PHOTOGRAPHY, https://www.moneymakerphotography.com/five-problems-youll-face-photographer=/ [https://perma.cc/54XN-SYS3] (last visited July 19, 2025).

522. *See* Christian Perdomo, *Photography in the US - Market Research Report (2015-2030)*, IBIS WORLD (Mar. 2025), https://www.ibisworld.com/united-states/industry/photography/1443/ [https://perma.cc/94BF-P4NM].

523. *See id.*

But, from a First Amendment perspective, the increase in photographs taken by an increasingly greater number of people in the United States is a boon. It is fundamental in this country that "*more speech*, not less, is the governing rule."[524] The proliferation of cameras for both photographs and videos spurred today's Creator Economy in which millions of people have become content creators as a profession and passion, especially on social media.[525] Contrary to the Copyright Office's and Judge Chhabria's market dilution theory, the creation of new, non-infringing works should be valued in this country, not discounted, in the fair use analysis.

The Copyright Office's and Judge Chhabria's newfound theory of copyright dilution is a slippery slope with no stopping point. All output of generative AI is likely to reflect some genre, style, or type of work from the training datasets, including essays, news articles, books, computer programs, images, visual works, songs, videos, and the like. To see just how limitless this theory of dilution is, consider UMG's comment that all AI-generated songs of everyone using AI that was trained on any genre of music constitutes copyright dilution: "All such competition is cognizable as defeating fair use under the fourth factor. *Works in any genre of music* might compete with works in different genres, and music companies all vie for the ears of listeners who favor multiple kinds of music."[526] By the Copyright Office's illogic, every output of generative AI, even if non-infringing, diminishes the degree of transformativeness of the AI to "modestly" transformative "at best," and poses a market harm through copyright dilution from the creation of works in the same genre.[527] Or, as Judge Chhabria concluded, the "upshot" is that AI companies will likely commit copyright infringement if they do not license the training works.[528]

Copyright was never intended as a shield from general competition from non-infringing works. Unless the copyright holder can prove a work in question is an infringing copy or derivative work, the competing work should be not only allowed

524. Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 361 (2010) (emphasis added); United States v. Alvarez, 567 U.S. 709, 719–20 (2012) (Kennedy, J.) (plurality opinion) (scienter requirement for defamation "exists to allow more speech, not less.").

525. *See* Lee, *supra* note 461, at 1561.

526. Universal Music Group, Comment Letter on Artificial Intelligence and Copyright, 56–57 (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-9014 [https://perma.cc/SHN7-XW89] (emphasis added).

527. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 46, 64–65.

528. Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *2 (N.D. Cal. June 25, 2025).

but encouraged.[529] As Judge Alsup explained in rejecting the *Bartz* plaintiffs' market dilution theory, "[t]his is not the kind of competitive or creative displacement that concerns the Copyright Act. The Act seeks to advance original works of authorship, not to protect authors against competition."[530] Protecting copyright holders from competition from non-infringing works in entire genres turns the Copyright Clause on its head. Instead of promoting progress, the goal is to protect copyrights—and perversely to reduce the creation of new, non-infringing works. This approach "threaten[s] to reduce diversity and competition in the marketplace of ideas."[531]

The Copyright Office's unbounded view of the scope of copyright is unconstitutional. First, this interpretation violates the Progress Clause, which limits the grant of exclusive rights to authors to "*their respective [w]ritings*,"[532] and does not allow exclusive rights to extend to the new, non-infringing writings of others.[533] Indeed, as the Supreme Court has explained, "the Framers intended copyright itself to be the engine of free expression."[534] Copyright does so "[b]y establishing a marketable right to the use of *one's expression*."[535] Copyright law has never protected ideas or methods, much less entire literary styles.[536] Unless a style involves copyrightable elements, everyone is free to copy and build on it.[537] As Judge Alsup explained in rejecting the *Bartz* plaintiffs' attempt to expand copyright to proscribe AI training:

> [I]f someone were to read all the modern-day classics because of their exceptional expression, memorize them, and then emulate a blend of their best writing, would that violate the Copyright Act? Of course not. Copyright does not extend to

529. *See* Bartz v. Anthropic, No. C 24-05417 WHA, 2025 WL 1741691, at *17 (N.D. Cal. June 23, 2025).

530. *Id.*

531. Anderson v. Celebrezze, 460 U.S. 780, 794 (1983).

532. U.S. CONST. art. I, § 8, cl. 8 (emphasis added).

533. Congress and the Supreme Court have interpreted "writings" in the Progress Clause broadly to encompass a wide assortment of creations and subject matter. *See* Feist Publ'n, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 346 (1991).

534. Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 558 (1985).

535. *Id.* (emphasis added).

536. *See* Eldred v. Ashcroft, 537 U.S. 186, 190 (2003); Baker v. Selden, 101 U.S. 99, 102 (1879); 17 U.S.C. § 102(b).

537. *See* Steinberg v. Columbia Pictures Indus., Inc., 663 F. Supp. 706, 712 (S.D.N.Y. 1987) ("[S]tyle is one ingredient of 'expression' . . . .").

> "method[s] of operation, concept[s], [or] principle[s]" "illustrated[ ] or embodied in [a] work."[538]

Dolly Parton was a true innovator for the pop country style of music, for example.[539] Her creativity led to many other music icons building on that crossover style, including Taylor Swift, Miley Cyrus, Ariana Grande, and Beyoncé.[540] Granted, some AI generators may copy copyrightable elements when generating a work in response to a person's prompt to create in "the style of" a specific artist.[541] But the proper remedy is a copyright infringement lawsuit, not concocting a mutant species of copyright dilution that penalizes non-infringing works.

Second, recognizing copyright dilution based on new, non-infringing works would likely violate the First Amendment. Consider this analogy. Imagine that a court extended defamation to apply to truthful but unflattering statements about a person in AI-generated articles because of the fear that the "flood" of AI-generated articles would dilute the persona's reputation and make it difficult to find the positive biographical information the person wrote and published. Can a defamation law that imposed liability for dilution resulting from completely *truthful* information withstand First Amendment scrutiny? Of course not. The First Amendment protects truthful information and even personal opinions.[542] Truth is a complete defense.[543]

Similarly, courts must ask the constitutional question that the Copyright Office's report and Judge Chhabria's opinion ignored: Can a copyright law that imposed liability for dilution resulting from completely *non-infringing* works withstand First Amendment scrutiny? As Judge Chhabria's opinion concedes, accepting market dilution from non-infringing AI works as cognizable harm means that, in most cases, AI training is

538. Bartz v. Anthropic, No. C 24-05417 WHA, 2025 WL 1741691, at *7–8 (N.D. Cal. June 23, 2025) (first alternation not in original) (quoting 17 U.S.C. § 102(b)).

539. *See* Emily Lordi, *The Grit and Glory of Dolly Parton*, N.Y. TIMES STYLE MAG. (Dec. 1, 2020, at 17:51 ET), https://www.nytimes.com/2020/11/30/t-magazine/dolly-parton.html [https://perma.cc/SWV9-AK3S].

540. *Id.*

541. *See, e.g.*, *Studio Ghibli v. OpenAI: Is This the Next U.S. Copyright Lawsuit To Drop?*, CHATGPT IS EATING THE WORLD (Mar. 28, 2025), https://chatgptiseatingtheworld.com/2025/03/28/ghibli-studios-v-openai-is-this-next-u-s-copyright-lawsuit-to-drop/ [https://perma.cc/R33P-8ZWQ].

542. *See* Pan Am Sys., Inc. v. Atlantic Ne. Rails & Ports, Inc., 804 F.3d 59, 64–65 (1st Cir. 2015).

543. *Id.* at 64.

infringement,[544] which thereby puts into legal doubt people's ability to use the AI generator developed from the training. For the reasons explained below, I believe adopting a theory of market dilution from non-infringing works violates the First Amendment.

To begin, it's important to recognize that, along with the idea – expression dichotomy, fair use is a First Amendment safeguard that protects non-infringing speech.[545] However, copyright dilution expands the scope of copyright beyond authors' respective writings to encompass the new, non-infringing creative expression of others, here, people who use AI generators to create non-infringing works.[546] As the Supreme Court explained the distinction between creating one's own speech and potentially infringing speech, "[t]he First Amendment securely protects the freedom to make—or decline to make—one's own speech; it bears less heavily when speakers assert the right to make other people's speeches."[547] As long as an AI-generated work is non-infringing, the First Amendment protects the person's own right to create the work.[548]

And copyright law cannot discriminate against the expressive work merely because the person used AI in the process of creating it. To do so would violate not only the Supreme Court's principle of aesthetic non-discrimination, but it would also constitute impermissible content and viewpoint discrimination under the First Amendment.[549] "Under our Constitution, 'esthetic and moral judgments about art and literature . . . are for the individual to make, not for the Government to decree, even with the mandate or approval of a majority.'"[550] As Justice Scalia explained: "[W]hatever the challenges of applying the Constitution to

544. Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *23 (N.D. Cal. June 25, 2025).

545. Eldred v. Ashcroft, 537 U.S. 186, 219–20 (2003).

546. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 66 ("*Even when the output is not substantially similar to a specific underlying work*, stylistic imitation made possible by its use in training may impact the creator's market." (emphasis added)).

547. *Eldred*, 537 U.S. at 221.

548. *See* Eugene Volokh, Mark Lemley & Peter Henderson, *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 651, 657–59 (2023); Glik v. Cunniffe, 655 F.3d 78, 82 (1st Cir. 2011) (holding that making a recording of a police officer carrying out their duties in public is protected speech).

549. *See* Bleistein v. Donaldson Lithographing Co., 188 U.S. 239, 251 (1903) ("It would be a dangerous undertaking for persons trained only to the law to constitute themselves final judges of the worth of pictorial illustrations, outside of the narrowest and most obvious limits."); Brown v. Ent. Merchs. Ass'n., 564 U.S. 786, 790 (2011); Edward Lee & Andrew Moshirnia, *The AI Penalty: Is There a Bias Against AI-Generated Works?*, 2024 MICH. ST. L. REV. 641, 720–22 (explaining how disfavoring AI-generated works created by humans is viewpoint discrimination); *Eldred*, 537 U.S. at 219–20.

550. *Brown*, 564 U.S. at 790 (quoting United States v. Playboy Ent. Grp., Inc., 529 U.S. 803, 818 (2000)).

ever-advancing technology, 'the basic principles of freedom of speech and the press, like the First Amendment's command, do not vary' when a new and different medium for communication appears."[551] A person using generative AI today to create books in stream of consciousness or any other style is entitled to as much First Amendment protection as Faulkner or Joyce.[552]

The Copyright Office's approach to dilution is content-based.[553] The content-based discrimination is evidenced in the differential treatment to AI-generated content based on their similarity to the expressive style, genre, or type of works in the training datasets on which the AI model was trained.[554] The Copyright Office's approach to copyright dilution singles out for unfavorable treatment under fair use based on the presence of certain communicative content—expressing certain styles, genres, and types of works. For example, non-infringing content in the style of romance novels or in the genre of music created by people using AI are treated as copyright dilution because of their communicative content (i.e., romance novels and music) if the same types of works were in the AI training datasets.[555] This discrimination based on the communicative content of the AI-generated work is content-based, requiring strict scrutiny.[556]

Moreover, the Copyright Office's endorsement of copyright dilution is also likely viewpoint-based discrimination.[557] In adopting dilution,

---

551. *Id.* (quoting Joseph Burstyn, Inc. v. Wilson, 343 U.S. 495, 503 (1952)).

552. *Cf. id.* ("Like the protected books, plays, and movies that preceded them, video games communicate ideas—and even social messages—through many familiar literary devices (such as characters, dialogue, plot, and music) and through features distinctive to the medium (such as the player's interaction with the virtual world). That suffices to confer First Amendment protection.").

553. *See* Reed v. Town of Gilbert, Ariz., 576 U.S. 155, 169 (2015) (finding a content-based sign regulation because it "singles out specific subject matter for differential treatment"); PRE-PUBLICATION REPORT, *supra* note 40, at 65.

554. PRE-PUBLICATION REPORT, *supra* note 40, at 64–65 (noting AI-generated content "pose[s] a serious risk of diluting markets for works of the same kind as in their training data," such as romance novels and music).

555. *See id.*

556. *See Reed*, 576 U.S. at 169–70.

557. *See* R.A.V. v. City of St. Paul, 505 U.S. 377, 380, 393–94 (1992) (holding that a law prohibiting "fighting words" only related to race, color, creed, religion, or gender was viewpoint discrimination); Iancu v. Brunetti, 588 U.S. 388, 393–94 (2019) (holding that a trademark registration bar against "immoral" and "scandalous" marks was viewpoint discrimination); Matal v. Tam, 582 U.S. 218, 235–36 (2017) (plurality) (holding that a trademark registration bar against marks that may disparage persons was viewpoint discrimination); Rosenberger v. Rector & Visitors of Univ. of Va., 515 U.S. 819, 831–32 (1995) (holding that a state university's singling out religious student organizations and excluding them from benefits available to all other (nonreligious) student organizations was viewpoint discrimination).

the Copyright Office favorably cited commenters that discriminate against non-infringing AI-generated works as inferior creations.[558] For example, the Science Fiction and Fantasy Writers Association called AI-generated works "*trash*."[559] Even more disturbingly, the Copyright Office favorably cited and even quoted this "trash" comment in support of its endorsement of the theory of copyright dilution.[560] Indeed, the record before the Copyright Office was rife with such viewpoint-based discrimination with many calling AI-generated content "trash," "garbage," "junk," "inferior," "low quality," "slop," "disgusting," "cheap," "crap," an "incestuous cesspool," and even worse pejoratives.[561] This type of judgment about the dilutive and tarnishing quality of non-infringing AI-generated works is similar to the determinations of "scandalous" trademarks under the Lanham Act the Supreme Court found to be unconstitutional viewpoint discrimination by the

558. PRE-PUBLICATION REPORT, *supra* note 40, at 65.

559. *Id.* at 65 & n.374 (emphasis added) ("The harm creators and audiences are already experiencing is a flood of trash, directly enabled by generative AI with no restrictions on output. . . . AI-generated material . . . literally crowds human writers out.").

560. *Id.*

561. *See, e.g.*, Science Fiction and Fantasy Writers Association, Comment Letter on Artificial Intelligence and Copyright (Dec. 6, 2023), https://www.regulations.gov/comment/COLC-2023-0006-10349 [https://perma.cc/L6EY-B9CG]; Anonymous, Comment on Artificial Intelligence and Copyright (Nov. 12, 2023), https://www.regulations.gov/comment/COLC-2023-0006-10060 [https://perma.cc/26BQ-MTHY] ("It's hot garbage, not actually art, doesn't take skill to make, and is made via theft."); Grace Ginaven, Comment on Artificial Intelligence and Copyright (Sep. 1, 2023), https://www.regulations.gov/comment/COLC-2023-0006-4683 [https://perma.cc/ZJV9-EUAL] ("These companies who control AI use it to undercut and scab hardworking Americans out of livelihoods while producing junk."); John Snell, Comment on Artificial Intelligence and Copyright (Sep. 1, 2023), https://www.regulations.gov/comment/COLC-2023-0006-4817 [https://perma.cc/Z9SV-5QL4] ("[A]ll they can do is steal, reconstitute, and spit out images and text which are vastly inferior to anything created by human hand or eye."); The Authors Guild, *supra* note 197, at 3 ("In the book market, there is a serious risk of market dilution from machine-generated works. Generative AI systems are already being used to produce low-quality ebooks that attempt to compete with authors' works."); Copyright Alliance, Comment Letter on Artificial Intelligence and Copyright (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-8935 [https://perma.cc/VB6W-CJRB] ("Flooding the market with AI-generated works creates competition against human-created works and makes it difficult for consumers to find the higher-quality, human-created works they prefer in the sea of low-quality outputs."); Eli Franklin, Comment on Artificial Intelligence and Copyright (Oct. 5, 2023), https://www.regulations.gov/comment/COLC-2023-0006-7243 [https://perma.cc/YUK5-TJCL] ("AI is a force of nothing but evil and sho[u]ld be heavily regulated as soon as possible, lest this nation or the world plunge into an age of meaningless, unethically made, mindless slop."); Anonymous, Comment on Artificial Intelligence and Copyright (Sep. 3, 2023), https://www.regulations.gov/comment/COLC-2023-0006-6553 [https://perma.cc/8C5T-NVDH] ("AI 'art' is disgusting."); Anonymous, Comment on Artificial Intelligence and Copyright (Sep. 1, 2023), https://www.regulations.gov/comment/COLC-2023-0006-5460 [https://perma.cc/F3AP-H4UA] ("Every level of our society needs to summarily reject handing over IP as a trade-in for mediocre, cheap, stolen AI-produced works."); Eric C., Comment on Artificial Intelligence and Copyright (Sep. 28, 2023), https://www.regulations.gov/comment/COLC-2023-0006-7071 [https://perma.cc/4A98-AWBF] (AI produces "cheap copies of art" and "a cheap piece of crap"); Allora White, Comment on Artificial Intelligence and Copyright (Sep. 7, 2023), https://www.regulations.gov/comment/COLC-2023-0006-1753 [https://perma.cc/6GGZ-63TU] (AI will produce "an incestuous cesspool of stolen art").

Trademark Office.[562] Here, the only infringing "trash" penalized as dilution are the works people created with AI tools.

It's fine for non-governmental actors to express such viewpoints. But, under the First Amendment, the Copyright Office has no authority to discriminate against a non-infringing work as "trash," no matter what technology its creator used. As the Supreme Court has long recognized, "one man's vulgarity is another's lyric. Indeed, we think it is largely because governmental officials cannot make principled distinctions in this area that the Constitution leaves matters of taste and style so largely to the individual."[563] The wisdom of this principle has been validated throughout history, including at the start of the twentieth century when people in the United States, including leading physicians, attacked the paintings of Picasso, Matisse, and other modern artists, especially the Cubists, as "degenerate"—and a direct threat to society.[564] Shortly before this regrettable period during the inception of modern art, Justice Holmes astutely admonished that the government must refrain from evaluating artistic worth: the "very novelty" of new forms of art "would make them repulsive until the public had learned the new language in which their author spoke."[565]

But the Copyright Office's copyright dilution is premised on the view that expression created by persons using AI are inferior—or "trash"—and therefore dilutive even though the expression is completely non-infringing. Of all the non-infringing works of the same style or genre, it singles out only one class for unfavorable treatment under fair use: the content created by people using AI.[566] Why? Because of the view that non-infringing AI-generated content—aka "trash"—*dilutes* the content of the copyright holders.[567] And the only class of creators whose works or expression is treated as dilutive—and harmful under fair use— under the Copyright Office's approach are people who use generative AI to create romance novels, music, or other genre of work contained in the AI training datasets.[568] In other words, the Copyright Office's approach disfavors or penalizes only the non-infringing works of people who believe that AI offers a legitimate way to create new works—such use manifests their "motivating ideology or . . . opinion or perspective" on the usefulness

562. *Iancu*, 588 U.S. at 394.
563. Cohen v. California, 403 U.S. 15, 25 (1971).
564. *See* Lee & Moshirnia, *supra* note 549, at 722–23.
565. Bleistein v. Donaldson Lithographing Co., 188 U.S. 239, 251 (1903).
566. PRE-PUBLICATION REPORT, *supra* note 40, at 64–65.
567. *Id.*
568. *Id.*

of generative AI in artistic and literary endeavors.[569] Of course, many commenters to the Copyright Office strongly disagreed with—and even condemned—that viewpoint and the use of AI,[570] reflecting a larger backlash, if not moral panic, against the use of AI in creative endeavors.[571] But the First Amendment does not permit the Copyright Office to disfavor the viewpoint of one class of creators over another.[572] The law can no more single out and penalize the non-infringing works of AI creators as it could discriminate against Cubist artists for adopting a new approach to create that many other people attacked.[573]

And the constitutional infirmity of the Copyright Office's discriminatory treatment of AI creators' works is not cured by its effectuation through the denial of fair use. As the Supreme Court recognized in the context of tax exemptions, "[t]o deny an exemption to claimants who engage in certain forms of speech is in effect to penalize them for such speech."[574] Under the First Amendment, the "government cannot suppress ideas indirectly any more than it can do so directly."[575] Copyright dilution is a species of mutant copyright law that has no place in our copyright system.[576] As Judge Posner warned in a (non-AI) case involving collectors' guide books, "the fair-use doctrine is not intended to set up the courts as judges of the quality of expressive works."[577] To do so "rais[es] a First Amendment question."[578]

At a minimum, courts must apply strict scrutiny to such an overbroad application of a mutant copyright law because it alters the traditional contours of fair use and the idea-expression dichotomy by

569. Rosenberger v. Rector & Visitors of Univ. of Va., 515 U.S. 819, 829 (1995).

570. *See* comments cited *supra* note 561.

571. *See* Lee & Moshirnia, *supra* note 549, at 674–80.

572. *See Rosenberger*, 515 U.S. at 831 (holding that a state university engaged in viewpoint discrimination by denying funding only to student groups with religious perspectives); *see also* Nat'l Endowment for the Arts v. Finley, 524 U.S. 569, 587 (1998) (viewpoint discrimination may arise "[i]f the NEA were to leverage its power to award subsidies on the basis of subjective criteria into a penalty on disfavored viewpoints").

573. *See* Lee & Moshirnia, *supra* note 549, at 723–24; *cf.* R.A.V. v. City of St. Paul, 505 U.S. 377, 384 & n.4 (1992) (discussing how government singling out one subset of speech based on the content of the speech constitutes viewpoint discrimination); Police Dept. of City of Chi. v. Mosley, 408 U.S. 92, 95 (1972) (holding that a city ordinance violated the First Amendment violation by allowing "[p]eaceful picketing on the subject of a school's labor-management dispute is permitted, but all other peaceful picketing is prohibited.").

574. Speiser v. Randall, 357 U.S. 513, 518 (1958).

575. Brooklyn Inst. of Arts & Scis. v. City of N.Y., 64 F. Supp. 2d 184, 199 (E.D.N.Y 1999).

576. *Cf.* Dastar Corp. v. Twentieth Century Fox Film Corp., 539 U.S. 23, 34 (2003) (declining to allow a cause of action that would create a mutant copyright law that limits the public's federal right to copy and use expired copyrights).

577. Ty, Inc. v. Publ'ns Int'l Ltd., 292 F.3d 512, 515, 523 (7th Cir. 2002).

578. *Id.* at 523.

stretching copyright to penalize other people's non-infringing works.[579] The fears of rampant market dilution or the cannibalization of the market with AI-generated works are highly speculative—in *Bartz v. Anthropic*, Judge Alsup even called the assertion "baloney."[580] The First Amendment does not permit the recognition of copyright dilution to limit the creation of this non-infringing expression. Less restrictive alternatives exist under U.S. law to handle the worst fraudulent practices some copyright holders decry.[581]

D. Technology Usage: Alleged Infringement in Outputs of AI Generators

Now let's turn to alleged uses of copyrighted works that may arise when people use the AI model. These are potentially different uses of the copyrighted works that should be analyzed separately from the fair use analysis of the training and development of the AI model. Given the myriad amounts and types of outputs AI generators can produce, I am writing a separate article to provide a more in-depth analysis of some of the most salient problems.[582] I briefly summarize here a few general principles when analyzing infringing outputs on a use-by-use basis.

---

579. *See* Eldred v. Ashcroft, 537 U.S. 186, 221 (2003) ("[W]hen, as in this case, Congress has not altered the traditional contours of copyright protection, further First Amendment scrutiny is unnecessary.").

580. Derek Slater, *AI Training, the Licensing Mirage, and Effective Alternatives to Support Creative Workers*, TECH POL'Y .PRESS (June 2, 2025), https://www.techpolicy.press/ai-training-the-licensing-mirage-and-effective-alternatives-to-support-creative-workers/ [https://perma.cc/S2PS-4N9V].

581. *See, e.g.*, *North Carolina Musician Charged with Music Streaming Fraud Aided by Artificial Intelligence*, U.S. ATTY'S OFFICE S. DIST. OF N.Y. (Sep. 4, 2024), https://www.justice.gov/usao-sdny/pr/north-carolina-musician-charged-music-streaming-fraud-aided-artificial-intelligence [https://perma.cc/3VUZ-WTJL]; Andersen v. Stability AI Ltd., 744 F. Supp. 3d 956, 977–78, 81 (N.D. Cal. 2024) (holding that plaintiffs stated claims for false endorsement and vicarious trade dress infringement under the Lanham Act for alleged use of their names and trade dress to mimic their styles in Midjourney's AI generator). For example, the Authors Guild points to the example of the author Jane Friedman, who saw "on Amazon 'a cache of garbage books' written under her name on subjects she is known for." The Authors Guild, *supra* note 197, at 3. That allegation can be addressed by a false endorsement or false advertising claim under the Lanham Act. *See Andersen*, 744 F. Supp. at 977, 981; 15 U.S.C. § 1125(a). And a copyright claim arises for unauthorized abridgements or extensive summaries of books. *See* Twin Peaks Prods., Inc. v. Publ'ns Int'l, Ltd., 996 F.2d 1366, 1372–73 (2d Cir. 1993); 17 U.S.C. § 101 (definition of "derivative work" includes "abridgment" of a work).

582. *See* Edward Lee, AI Output: Copyright Infringement v. Fair Use (2025) (unpublished manuscript) (on file with the author).

1. The Need to Identify Substantially Similar Outputs. If the AI model produces and publicly distributes works substantially similar to the works on which it was trained, such "regurgitated" copies would typically not serve a fair use purpose because they serve the same purpose as the copyright holders' publication of the works, absent some other justification such as parody, comment, or research. By contrast, for non-infringing works that the AI model produces, there is simply no "use" of the copyright holders' works in the outputs of the AI model. The general competition between the copyright holders' works and non-infringing AI-generated works is not cognizable as a market harm under fair use.[583]

2. Internal "Memorization" Should Not Be Infringing Absent a Corresponding Infringing Output. Courts have correctly rejected the most expansive theories of infringement in some of the early complaints filed that alleged that the AI model itself constituted one giant infringing derivative work.[584] As Judge Chhabria concluded in *Kadrey v. Meta*, such a theory "would have to mean that if you put the Llama language model next to Sarah Silverman's book, you would say they're similar. That makes my head explode when I try to understand that."[585] Underlying Judge Chhabria's conclusion is the fundamental requirement of a copyright claim that the plaintiff must prove substantial similarity between the defendant's alleged copy and the plaintiff's own work as viewed by the ordinary lay audience.[586] Under the *de*

583. *See* Bartz v. Anthropic, No. C 24-05417 WHA, 2025 WL 1741691, *6–7 (N.D. Cal. June 23, 2015).

584. *See Judge Chhabria Poised to Dismiss Some of Sarah Silverman's Copyright Claims v. Meta*, CHAT GPT IS EATING THE WORLD (Nov. 11, 2023), https://chatgptiseatingtheworld.com/2023/11/11/judge-chhabria-poised-to-dismiss-some-of-sarah-silvermans-copyright-claims-v-meta/ [https://perma.cc/A6FP-H584].

585. *Id.* (emphasis omitted).

586. *See* Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2023 WL 8039640, at *1 (N.D. Cal. Nov. 20, 2023) ("The plaintiffs allege that the 'LLaMA language models are themselves infringing derivative works' because the 'models cannot function without the expressive information extracted' from the plaintiffs' books. This is nonsensical. . . . There is no way to understand the LLaMA models themselves as a recasting or adaptation of any of the plaintiffs' books."). *But see* Andersen v. Stability AI Ltd., 744 F. Supp. 3d 956, 973, 975 n.16 (N.D. Cal. 2024) (distinguishing *Kadrey* from the claims against image generators, which included a theory that the models store "compressed copies" of the training images, but still requiring proof of substantial similarity).

minimis doctrine, it is not infringement when a lay audience would not even notice the alleged appropriation.[587]

Therefore, the potential "memorization" of training materials by the AI model during the process of training should not constitute infringement absent a substantially similar output reflecting a memorized copy.[588] A. Feder Cooper and James Grimmelmann define memorization as follows: "[A] model has 'memorized' a piece of training data when (1) it is possible to reconstruct from the model (2) a (near-)exact copy of (3) a substantial portion of (4) that specific piece of training data."[589] AI researchers have tried to avoid AI models memorizing content from the training datasets, but a complete solution apparently has not been found.[590] One study (in pre-print) using adversarial extraction (not typically used by the ordinary AI user) found that memorization often occurs in larger models but "the extent of memorization varies widely from model to model and, within a model, even from work to work."[591]

Memorization raises a potential copyright issue because the model may have stored a copy of a copyrighted work from the training datasets. As Cooper and Grimmelmann contend: "If a generative-AI model memorizes its training data, these training data are *in the model*."[592] Memorization of training materials explains how so-called regurgitations of near-exact copies of training materials occur with AI generators.[593] If an AI generator produces a regurgitated copy from the training materials, the "memorized content must be encoded in the model's parameters."[594] The scholars stop short, however, of taking any copyright position on a regurgitated or memorized copy.[595]

587. *See* VMG Salsoul, LLC v. Ciccone, 824 F.3d 871, 878 (9th Cir. 2016) ("A 'use is de minimis only if the average audience would not recognize the appropriation.'") (quoting Newton v. Diamond, 388 F.3d 1189, 1193 (9th Cir. 2004)).

588. *See Bartz*, 2025 WL 1741691, at *7–8.

589. A. Feder Cooper & James Grimmelmann, *The Files Are in the Computer: On Copyright, Memorization, and Generative AI*, CHI.-KENT. L. REV. (forthcoming 2025).

590. *See* Brown et al., *supra* note 367, at 29–33 (discussing attempts to prevent memorization in GPT-3 model).

591. *See* A. Feder Cooper et al., *Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models*, ARXIV (July 10, 2025), https://arxiv.org/abs/2505.12546 [https://perma.cc/XH7S-URY6].

592. Cooper & Grimmelmann, *supra* note 589, at 5, 29, 31–33.

593. *Id.* at 2, 15.

594. *Id.* at 25.

595. *Id.* at 23–24.

In my view, the correct position from a copyright law perspective is that a purely internal memorized copy of a work is *not* infringement absent an infringing output of the work. First, as explained at length above, the use of copies of works to train an AI model serves a different, transformative purpose—namely, to develop a new AI program or technology with public benefits. Second, memorization that is purely internal to the model is not substitutional without any outputs that infringe. Third, the fair use cases involving large databases containing exact copies of billions of works scraped from the internet and millions of books digitized from libraries demonstrate that the mere storage of internal copies for a legitimate fair use purpose to create a new technology does not undermine its claim to fair use.[596] Fourth, a copyright infringement claim must be proven with evidence of a copy of the plaintiff's work: a general allegation that models typically memorize some training materials does not prove it has memorized the plaintiff's works. Finally, an analogy to human memorization further supports the lack of infringement in purely internal AI memorization. When humans memorize poems, songs, or other copyrighted works (whether intentionally, subconsciously,[597] or due to a photographic memory[598]), they are not committing infringement based on that memorization alone. Instead, they might infringe the copyright for a work they memorized if they produced a copy of it and distributed it to the public.[599]

596. *See supra* notes 159–60.

597. *See* Sheldon v. Metro-Goldwyn Pictures Corp., 81 F.2d 49, 54 (2d Cir. 1936) ("With so many sources before them they might quite honestly forget what they took; nobody knows the origin of his inventions; memory and fancy merge even in adults. Yet unconscious plagiarism is actionable quite as much as deliberate.").

598. *See Photographic Memory*, NEWSCIENTIST, https://www.newscientist.com/definition/photographic-memory/ [https://perma.cc/C5W9-AS7D] (last visited July 20, 2025); *see also* Timothy J. McFarlin, *Fixation*, ELGAR ENCYCLOPEDIA INTELL. PROP. L. (forthcoming 2025) (discussing how brain scanning technology may eventually show how parts of copyrighted works are memorized or "fixed" in the brain).

599. Based on a study of memorization by AI models, some researchers suggest that the distribution of the model containing some memorized copies of works, such as Meta's open-source LlaMa 3.1, "could be seen as . . . potentially infringing distributions of reproductions of" the memorized works. *See* Cooper et al., *supra* note 591, at 12. Even if so, presumably the entities and individuals downloading the open-source model would themselves have possible fair use defenses in developing and improving the model. Indeed, if, in the above example, Meta is considered the secondary user of the copyrighted work, the entities building on LlaMA can be considered tertiary users, who have potentially even stronger claims of fair use because presumably the amount of memorized works is far less than the universe of works Meta copied.

My approach comports with how Judge Alsup ruled in *Bartz v. Anthropic*.[600] Judge Alsup assumed for the purposes of summary judgment that Anthropic's model memorized all copies of the plaintiffs' books on which it was trained, but he nonetheless recognized that neither human nor AI memorization of expression in books is infringing when used merely to create new, non-infringing works that emulate unprotected elements of the books.[601] Indeed, it is "quintessentially transformative" to use copyrighted works "to turn a hard corner and create something different" by "generat[ing] new text" that itself is not infringing.[602]

3. Guardrails to Avoid AI Generation of Infringing Content Is Important to Fair Use. AI generators are not perfect in avoiding the generation of potentially infringing content. Examples of regurgitated and allegedly substantially similar content abound.[603] Some are the basis of the pending copyright lawsuits against AI companies.[604] Yet, AI generators often include guardrails or filters to minimize infringing outputs, although they are not foolproof.[605] Indeed, we can reasonably conjecture that AI generators are producing far more non-infringing than infringing content—the latter requires an output that is substantially similar to specific copyrighted works.[606] An average user of

600. Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *8 (N.D. Cal. June 23, 2025).

601. *Id.* at *7–8. The factual assumption of complete memorization of training data makes sense for the purposes of summary judgment, but it does not appear to be accurate. *See* Cooper et al., *supra* note 591, at 2 ("There is no evidence that most training data is memorized—especially not in high-quality, contemporary LLMs."); *id.* at 12 ("Our experiments show that the extent of memorization varies widely from model to model and, within a model, even from work to work in the Books3 dataset.").

602. *Bartz*, 2025 WL 1741691, at *8.

603. *See, e.g.*, Stuart A. Thompson, *We Asked A.I. to Create the Joker. It Generated a Copyrighted Image.*, N.Y. TIMES (Jan. 25, 2024), https://www.nytimes.com/interactive/2024/01/25/business/ai-image-generators-openai-microsoft-midjourney-copyright.html [https://perma.cc/8WGC-7BVX]; Chloe Xiang, *AI Spits Out Exact Copies of Training Images, Real People, Logos, Researchers Find*, VICE (Feb. 1, 2023, at 15:47 CT), https://www.vice.com/en/article/ai-spits-out-exact-copies-of-training-images-real-people-logos-researchers-find/ [https://perma.cc/4EK3-V2QK].

604. *See, e.g.*, New York Times' Complaint, *supra* note 32, ¶¶ 98–107.

605. *See* PRE-PUBLICATION REPORT, *supra* note 40, at 23–24; Mackenzie Ferguson, *Anthropic Sets the Stage for AI Innovation with New Copyright Guardrails*, OPENTOOLS (Dec. 31, 2024), https://opentools.ai/news/anthropic-sets-the-stage-for-ai-innovation-with-new-copyright-guardrails [https://perma.cc/W6QN-3QPK]; Yi Dong et al., *Building Guardrails for Large Language Models*, ARXIV, 1–2, (May 29, 2024), https://arxiv.org/abs/2402.01822 [https://perma.cc/U5CC-99X5] (discussing guardrails for LLMs including those to avoid copyright infringement).

606. *See* Samuelson, *supra* note 129, at 1551–52 ("In fact, outputs of the generative AI systems in litigation are overwhelmingly not substantially similar in expression to the

ChatGPT drafting a letter, conducting research, creating a spreadsheet, or just brainstorming is unlikely to be generating anything close to infringing content, for example. And even the U.S. Copyright Office has allowed people to register the copyrights for works, parts of which were AI-generated.[607] The Copyright Office would not be registering works it viewed as infringing.

Nonetheless, the generation of infringing content by AI generators is a valid concern. Does the potential or actual generation of infringing content undermine the fair use defense of AI generators?

Potentially, it does. In the extreme case, it could show the complete failure of an AI company's training if the AI model ended up regurgitating copies of all works it was trained on. In this scenario, the AI company had a legitimate fair use purpose to develop a new AI model, but its execution failed with the character of the outputs routinely involving substantially similar copies of works in the training datasets. An analogous example from the analog world occurred in the Harry Potter Lexicon case in which a publisher had a legitimate fair use purpose in creating a reference guide to keep track of the complex universe of characters, spells, and settings for the Harry Potter, but created the Lexicon in such a sloppy way, copying too much without quotations or attribution (relevant to Factor 3 of fair use), that the court ruled against fair use.[608] After the decision, the publisher cleaned up the Lexicon and published it.[609]

works on which the models were trained, although the extent to which models are likely to produce similar outputs depends on the diversity (or lack thereof) and volume of data used to train the models and the extent to which developers have used technical safety measures to prevent infringing outputs."(footnote omitted)); *compare Bartz*, 2025 WL 1741691, at *7–8 ("Authors do not allege that any LLM output provided to users infringed upon Authors' works. Our record shows the opposite. Users interacted only with the Claude service, which placed additional software between the user and the underlying LLM to ensure that no infringing output ever reached the users."), *with* Kadrey v. Meta Platforms, Inc., No. 23-CV-03417-VC, 2025 WL 1752484, at *14 (N.D. Cal. June 25, 2025) ("Meta's LLMs won't output any meaningful amount of the plaintiffs' books . . . .").

607. *See Has the Copyright Office Become More Receptive to AI-Generated Works? Yes, if They Embody Selection, Coordination, Arrangement of Human Creators*, CHAT GPT IS EATING THE WORLD (Mar. 14, 2025), https://chatgptiseatingtheworld.com/2025/03/14/has-the-copyright-office-become-more-receptive-to-ai-generated-works-yes-if-they-embody-selection-coordination-arrangement-of-humans/ [https://perma.cc/4JLJ-BPVH].

608. Warner Bros. Ent. v. RDR Books, 575 F. Supp. 2d 513, 523, 547–48 (S.D.N.Y. 2008).

609. *See The Lexion (The Harry Potter Lexicon Reader's Guide Series Book 2)*, AMAZON.COM, https://www.amazon.com/Lexicon-Harry-Potter-Readers-Guide-ebook/dp/B008FKBN2Q/ [https://perma.cc/558A-GANN] (last visited Aug. 16, 2025).

Courts in past technology fair use cases weighed favorably a company's implementation of guardrails to reduce the potential for substitutional copies being publicly disseminated. For example, in *Authors Guild v. Google*, Judge Leval recited at length the measures Google implemented to avoid users gaming the snippet view function of Google Book Search: these measures "substantially protect[] against its serving as an effectively competing substitute for Plaintiffs' books."[610] Judge Leval concluded: "The result of these restrictions is . . . that a searcher cannot succeed, even after long extended effort to multiply what can be revealed, in revealing through a snippet search what could usefully serve as a competing substitute for the original."[611] Judge Leval weighed favorably Google's security protections to avoid hackers getting access to its database of digital copies of the books, as well as Google's contractual obligations imposed on participating libraries to avoid the same with respect to their library digital copy.[612] Other technology fair use cases adopt this approach, weighing favorably the implementation of guardrails to minimize potential substitution of the copyright holders' works.[613] This approach to guardrails is consistent with the courts' rejection of speculation as the basis for finding market harm to the copyright holders.[614]

E. Coda on Nonexpressive Use

Considerable discussion—and disagreement—over the fair use analysis of AI training has involved the concept of nonexpressive use. In earlier scholarship, Matthew Sag, James Grimmelmann, and other scholars characterized the use of copyrighted works by machines or technologies in the fair uses cases (summarized above) as nonexpressive uses of the copyrighted works.[615] Sag, who coined the term "nonexpressive use" and elaborated its theory, explained the concept in the

610. Authors Guild v. Google, Inc., 804 F.3d 202, 222–23 (2d Cir. 2015).

611. *Id.* at 222.

612. *See id.* at 228–29.

613. *See, e.g.*, Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 100–01 (2d Cir. 2014).

614. *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 453–54 (1984); Samuelson, *supra* note 129, at 1490 ("When courts decide that plaintiffs' claims of market harm are speculative, as the Court concluded in *Sony Corporation of America v. Universal City Studios, Inc.*, fair use defenses tend to succeed." (footnote omitted)).

615. *See* James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 661–64, 666–67 (2016); Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607, 1625–44 (2009); Murray, *supra* note 65, at 279–80.

following way: "[N]onexpressive uses of copyrighted works—i.e., acts of copying that do not communicate the author's original expression to the public—should not generally be regarded as infringing."[616] With the recent burst of AI models, Sag has made a strong case that AI training involves nonexpressive uses of copyrighted works, which have a different or transformative purpose than the purpose of copyright holder's rights in expressive uses: the AI model dissects and identifies statistical relationships among bits of the content in entire datasets.[617] Yet, Sag conceded that the problem of "memoriz[ation]" by AI models "is a big deal": "If ordinary and foreseeable uses of generative AI result in model outputs that would infringe on the inputs no matter what intervening technological steps were involved, then the nonexpressive use rationale would no longer apply."[618] Other legal scholars vigorously disagree that nonexpressive use is a fair use purpose in any situation.[619]

It goes beyond the scope of this Article to wade into this debate, which has been well argued. But I would be remiss in not commenting on how it fits with my theory. In my view, it is not critical for the courts to decide whether to classify the alleged use as nonexpressive or expressive when analyzing the purpose and character of the use under Factor 1.

The *Google* decision itself did not do so.[620] Had it, the Court likely would have classified the Java declaring code as expressive to programmers.[621] The Court described the declaring code as intended to provide "names that would prove intuitively easy to remember."[622] The declaring code was meant to be easily understood by computer programmers, so "[i]t must be designed and organized in a way that is intuitive and understandable to developers."[623] The code was expressive to computer programmers, but that didn't disqualify Google's fair use argument. Both *Warhol* and *Google* eschewed an "all or nothing" approach to fair use and

616. *See* Sag, *supra* note 615, at 1625; Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 EMORY L.J. 1164, 1167 n.12 (2025).

617. *See* Sag, *supra* note 244, at 307–09.

618. *Id.* at 312 (footnote omitted); *see also* Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887, 1907, 1909, 1911–15 (2024).

619. *See, e.g.*, Brauneis, *supra* note 347, at 12–13; David W. Opderbeck, *Copyright in AI Training Data: A Human-Centered Approach*, 76 OKLA. L. REV. 951, 975–92 (2024); Sobel, *supra* note 244, at 68–69; Charlesworth, *supra* note 48, at 328.

620. *See* Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1202–04 (2021).

621. See *id.*

622. *Id.* at 1202.

623. *Id.*

stressed the analysis of factors is a "matter of degree."[624] These teachings caution against reducing the Factor 1 analysis into an "all or nothing" contest whose outcome hinges on whether a use is labeled "nonexpressive" or "expressive."

Placing too much weight on the distinction between "nonexpressive" and "expressive" use is a mistake—similar to the "artificial" and "semantic distinction between 'studying' and 'use'" of a computer program that the Ninth Circuit rejected in determining fair use in *Sony Computer Entertainment, Inc. v. Connectix Corp*.[625] Courts should reject the mistaken view that expressive uses must be unfair. Fair use has *never* required a user to avoid using expression. Indeed, such a misconception of fair use flies in the face of the whole notion of having a *further* purpose to justify a fair use of a copyrighted work.[626]

V. RESPONDING TO CRITICISMS

This final Part addresses criticisms of my positions above. These criticisms are not exhaustive, but they encompass the more salient counterarguments to my analysis.

A. University Research Is Saved Only by Noncommercial Use?

Critics may argue that AI training with copyrighted works is categorically not transformative. Indeed, one critic makes the bold, categorical assertion: "AI [c]opying [i]s [n]ot [t]ransformative," and arguments to the contrary are all "sleight[s] of hand that distract[] from the reality that protected expressive content is being stored in the model."[627] Yet, even advancing this line, some critics may be willing to concede that university-based AI training, given its noncommerciality, might have a legitimate fair use purpose due to its lack of commercial exploitation, perhaps citing personal time-shift recordings in *Sony*.[628] Courts sometimes view verbatim copying for noncommercial, educational, or personal use as serving a fair use purpose, even though the purpose is not transformative or different from the copyright

624. *Id.* at 1197–98; Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1273–74 (2023).

625. Sony Comput. Ent., Inc. v. Connectix Corp, 203 F.3d 596, 604 (9th Cir. 2000).

626. For example, a parody copies expression in part for the expressive purpose to conjure up the original work. *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 588 (1994).

627. Charlesworth, *supra* note 48, at 347, 358.

628. *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 442, 449–51, 454–55 (1984).

holder's purpose.[629] Recording broadcast TV shows for later viewing and circulating copied excerpts of works at universities for classes provide two examples.[630] The Second Circuit suggested that an individual making a copy of a scientific article for research might be a fair use, while a corporation's systematic process of copying journal articles was not.[631]

However, if the only saving grace for university-based AI training is merely its noncommercial nature, that is a tenuous basis for it to rest.[632] As Robert Brauneis, who rejects the fair use arguments in AI training, points out, the case law for university-based copying for "nonprofit educational purposes" has never recognized a "blanket exemption from copyright," especially not for copying of entire works.[633] In *Campbell v. Acuff Rose Music*, the Supreme Court admonished that "the mere fact that a use is educational and not for profit does not insulate it from a finding of infringement, any more than the commercial character of a use bars a finding of fairness."[634] These cautionary words may be particularly salient in the context of AI researchers at universities, given not only the hybrid roles that some prominent researchers have at both universities and AI companies, but also how university research positions can be stepping stones to founding a startup or finding employment at tech companies with highly lucrative salaries.[635]

The Second Circuit's rejection of fair use asserted by the Internet Archive, despite its noncommercial use in providing people access to copyrighted books in an online library typically under the restriction that only one person can loan out a digital copy scanned from the published book owned by Internet Archive's two partners, a library and a bookstore, should be a warning sign to all universities conducting AI

629. *See id.* at 451; Cambridge Univ. Press v. Becker, 446 F. Supp. 3d 1145, 1270–71 (N.D. Ga. 2020) (finding some digital excerpt copies of copyrighted works were fair use, but others not).

630. *See Sony Corp. of Am.,* 464 U.S. at 420, 456; *Cambridge Univ. Press*, 446 F. Supp. 3d at 1161.

631. *See* Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 916 (2d Cir. 1994).

632. *See* Desai & Reidl, *supra* note 246, at 83 ("If the research is published and used in limited academic settings, the work should be protected under fair use. Things change if the researcher envisions starting a small company that might offer commercial products, or seeks to sell the company outright, like when Google bought Deep Mind.").

633. *See* Brauneis, *supra* note 347, at 12–14.

634. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 584 (1994).

635. *See* Cade Metz, *A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit*, N.Y. TIMES (Apr. 19, 2018), https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html [https://perma.cc/QU73-D96G].

research in the United States.[636] If such noncommercial use did not save the nonprofit Internet Archive from copyright infringement, it might not save the nonprofit universities where AI training occurs.[637] That is, not unless AI training has a different or further purpose under Factor 1.[638]

It does. As explained above, the use of copyrighted materials to train AI models at universities has a fair use purpose that is distinct and far more compelling than its noncommercial character. AI training—at both universities and AI companies, including ones that are hiring AI researchers from universities who sometimes retain their university affiliations—is advancing AI innovation in the United States, a goal consistent not only with Progress Clause but also the national policy and priority of the United States.

B. How to Weigh the Use of Copies of Books from "Shadow" Libraries Consisting of "Pirated" Books

Another major controversy is the use of so-called shadow libraries, or online repositories of "pirated" books, copied without authorization of their authors. These shadow libraries have drawn public condemnation and considerable media attention in the AI litigation.[639] Thus far, two district courts in the AI lawsuits have ruled on the use of pirated books by AI companies—with conflicting approaches.

In *Bartz v. Anthropic*, Judge William Alsup treated Anthropic's building of a permanent library of pirated books it downloaded as a use of the books that was *separate from* Anthropic's later use of the books to train its AI model, under the use-by-use approach of *Warhol*.[640] Anthropic's "[p]irated [l]ibrary [c]opies" was not fair use.[641] Indeed, Judge Alsup referred to

636. *See* Hachette Book Grp., Inc. v. Internet Archive, 115 F.4th 163, 181–82, 185, 196 (2d Cir. 2024). For three months during the pandemic, the Internet Archive increased the number of persons who can access the same book to 10,000 people. *Id.* at 176 ("The NEL ran from March 24, 2020, to June 16, 2020, when IA reinstated its lending controls after this lawsuit was filed.").

637. *See id.* at 186 ("IA does not profit directly from its exploitation of the Works in Suit. For that reason, its use of the Works is non-commercial in nature.").

638. *See id.* ("[B]ecause IA's [noncommercial] use of the Works is not transformative, the first fair use factor favors Publishers.").

639. *See* Creamer, *supra* note 332.

640. Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *11–12 (N.D. Cal. June 23, 2025) (citing Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 533 (2023)).

641. *Id.* at *18.

Anthropic's conduct as "stealing."[642] And he even suggested, without deciding, that the result would be the same even if Anthropic had downloaded the pirated books without storing them in a permanent library: "Such piracy of otherwise available copies is inherently, irredeemably infringing even if the pirated copies are immediately used for the transformative use and immediately discarded."[643] In short, in denying fair use as to this library-building, Judge Alsup took a categorical approach, treating the acquisition of pirated books as virtually *per se* infringing.[644]

By contrast, Judge Chhabria took a flexible approach.[645] Instead of a separate use, Judge Chhabria viewed Meta's downloading of pirated books as a part of the same use of the books to train its AI model: "There is no serious question that Meta's use of the plaintiffs' books had a 'further purpose' and 'different character' than the books—that it was highly transformative."[646] Disagreeing with Judge Alsup, Judge Chhabria concluded that the "downloading must still be considered in light of its ultimate, highly transformative purpose: training Llama," and that same purpose applied to Meta's use of the copies as a first step to determine whether the books would be "good training data" and simply "cross-referencing" copies in different datasets.[647] However, the training purpose did not completely exonerate Meta's use of pirated books. Under Factor 4, Judge Chhabria suggested that Meta's use of the pirated books datasets would present a possible market harm if "Meta's copying helped others acquire copyrighted works, potentially including the plaintiffs' works, without paying for them (and without any indication that those other people were acquiring the works for fair use purposes)."[648] But the plaintiffs failed to present any evidence on this issue.[649]

The use of shadow libraries is sometimes framed by parties as an example of "bad faith" that should weigh against fair use.[650]

642. *Id.* at *2, *18.

643. *Id.* at *11.

644. Judge Alsup did leave open for further argument other uses of the pirated books copies that Anthropic may have had. *See id.* at *19 ("Nothing is foreclosed as to any other copies flowing from library copies for uses other than for training LLMs.").

645. *See* Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *12–*13 (N.D. Cal. June 25, 2025).

646. *Id.* at *12, *9.

647. *Id.* at *12.

648. *Id.* at *21.

649. *Id.*

650. *See* Plaintiffs' Opposition to Anthropic's Motion for Summary Judgment at 10–11, Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691 (N.D. Cal. Apr. 25, 2025).

However, both Judges Alsup and Chhabria found the case law unclear (and perhaps doubtful as to the relevance of bad faith) to rely on it as a basis for their fair use analysis.[651]

One could write an entire article analyzing this controversy.[652] Because I do not have that luxury, let me offer a few remarks. First, being an author myself, I completely understand the concerns and even outrage voiced by the book authors in the lawsuits. The authors will have their day in court to air their concerns—and all signs indicate that courts will entertain them seriously.

Second, fair use is, in many respects, a doctrine in which the ends justify the means. After all, the defendant's purpose of use is the central consideration of Factor 1, and, if the defendant's purpose serves a "further," fair use purpose (the ends), then the defendant's act of copying (the means) can be deemed lawful and not infringing. This approach of fair use is somewhat analogous to the public necessity doctrine in tort law in which the ends justify the means—i.e., destroying someone else's property or engaging in activity that would otherwise constitute a trespass to property may be justified if it serves a countervailing, larger public interest.[653] Judge Alsup's suggestion that pirated books are "inherently, irredeemably infringing," no matter the further fair use purpose offered warrants, at the very least, further justification—and scrutiny.[654] If criminal counterfeit goods can, once confiscated, lawfully be repurposed for charitable

651. *Kadrey*, 2025 WL 1752484, at *11; Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *12 & n.5 (June 23, 2025) ("But [Anthropic's] bad faith is not the basis for this decision.").

652. A more thorough discussion is provided in Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 U.C. DAVIS L. REV. 893, 955 (2019). *See also* Mark A. Lemley, *The Fruit of the Poisonous Tree in IP Law*, 103 IOWA L. REV. 245, 268 (2017) ("Application of the fruit of the poisonous tree doctrine is unwise when the defendant has made substantial contributions that do not merely take or adapt the plaintiff's work."); Pamela Samuelson, *Legally Speaking: How to Think About Remedies in the Generative AI Copyright Cases*, COMMC'NS. OF ACM, July 2024, at 27, 29 ("Thus far, courts have been unwilling to adopt a 'fruit of the poisonous tree' theory of copyright liability when a preparatory use of protected works infringes, but a subsequent product derived in part from the earlier use does not." (footnote omitted)).

653. *See* Monica E. Eppinger, *The Challenge of the Commons: Beyond Trespass and Necessity*, 66 AM. J. COMP. L. 1, 18 (2018); Surocco v. Geary, 3 Cal. 69, 69 (1853); Field v. City of Des Moines, 39 Iowa 575, 577–78 (1874).

654. *Bartz*, 2025 WL 1741691, at *11.

distribution in some cases,[655] why should pirated books be treated as forever, "irredeemably" infringing?

The Copyright Act's text supports taking a flexible approach. Unlike the provision for the first-sale doctrine and several other exceptions to copyright, the text of the fair use provision, § 107, contains no requirement of a "lawfully made" copy to fall within the exception.[656] Accordingly, courts should eschew creating any bright-line rules for how to weigh the defendant's use of shadow libraries, but should evaluate it on a case-by-case basis. There should be no presumption that it automatically weighs against fair use, contrary to the suggestion of the Copyright Office's pre-publication report.[657]

Nor should courts adopt Judge Alsup's categorical approach, especially not his suggestion that such copies are "inherently, irredeemably infringing even if the pirated copies are immediately used for the transformative use and immediately discarded."[658] Such a *per se* approach is too rigid and at odds with the Supreme Court's teachings that fair use is a flexible doctrine and "is not to be simplified with bright-line rules."[659]

The Supreme Court's approach is consistent with Congress's codification of fair use in the Copyright Act as a multi-factor balancing test.[660] On the very question of building a library of TV show recordings without permission as discussed in *Sony*, the Supreme Court in *Grokster* took an equivocal stance, saying it was

---

655. *See* Kristina Rae Montanaro, *"Shelter Chic": Can the U.S. Government Make It Work?*, 42 VAND. J. TRANSNAT'L L. 1663, 1679 (2009) (discussing Customs and Border Protection's distribution of seized counterfeit clothing to victims displaced from Hurricane Katrina); *NCDA & Partners Mark End of Project Donating Nearly 100K Seized Counterfeit Jackets to Charity*, AHRC NASSAU (Apr. 28, 2022) https://www.ahrc.org/ncda-partners-mark-end-of-project-donating-nearly-100k-seized-counterfeit-jackets-to-ny-charities [https://perma.cc/H8KU-WMTF] (Nassau County distribution of 100,000 counterfeit jackets to people in need through more than 160 nonprofits and charities).

656. *Compare* 17 U.S.C. § 109(a) ("[T]he owner of a particular copy or phonorecord lawfully made under this title . . . ."), *with id.* § 107 ("[F]air use of a copyrighted work . . . ."); Kirtsaeng v. John Wiley & Sons, Inc., 568 U.S. 519, 537 (2013) (discussing "lawfully made" copy requirement in §§ 109(c), 109(e), and 110(1)); *and* 17 U.S.C. § 108(c)(2) ("lawful possession of such copy" by library or archives).

657. PRE-PUBLICATION REPORT, *supra* note 40, at 52 ("In the Office's view, the knowing use of a dataset that consists of pirated or illegally accessed works should weigh against fair use without being determinative.").

658. *Bartz*, 2025 WL 1741691, at *11.

659. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577 (1994).

660. *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 448 n.31 (1984) (noting the committee reports to Copyright Act "eschewed a rigid, bright-line approach to fair use").

not "necessarily infringing."[661] And, in *American Geophysical Union v. Texaco, Inc.*, the Second Circuit left open the possibility that, had Texaco's scientist made use of the copies of the journal articles in the lab, in a paper format more suitable for use and possible destruction from chemicals in the lab, "these purposes might suffice to tilt the first fair use factor in favor of Texaco if these purposes were dominant."[662] The Second Circuit left open possible fair uses in archival or library copies in other cases.[663] Similarly, Judge Alsup's analysis of library-building in Anthropic should not be understood to set forth a bright-line rule that the acquisition and use of any "pirated" book is never fair use, regardless of the circumstances related to the use such as by researchers at universities or for AI training without library-building.[664]

A more flexible approach is especially warranted when the fair use question involves a practice of far greater significance beyond the parties in the case. Judge Alsup's opinion does not define what makes a copy "pirated."[665] But if "pirated" simply means unauthorized copy, that expansive approach would not only cast a shadow over the legality of *every* dataset of copyrighted content compiled without permission, but would also threaten to swallow the fair use inquiry—which always involves an unauthorized use.[666] How else would AI researchers at universities and companies get large datasets of millions of works without

661. Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 931 (2005) (discussing *Sony*, 464 U.S. at 424, 454–55). *But see In re* Aimster Copyright Litig., 334 F.3d 643, 647 (7th Cir. 2003) (concluding in dicta that the library building in *Sony* should be considered infringement).

662. Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 915, 919 (2d Cir. 1994).

663. *Id.* at 920 ("We do not mean to suggest that no instance of archival copying would be fair use, but the first factor tilts against Texaco in this case because the making of copies to be placed on the shelf in Chickering's office is part of a systematic process of encouraging employee researchers to copy articles so as to multiply available copies while avoiding payment.").

664. *See* Bartz v. Anthropic PBC, No. C 24-05417 WHA, 2025 WL 1741691, at *12–13 (N.D. Cal. June 23, 2025).

665. *Id.* at *2 ("[I]n January or February 2021, another Anthropic cofounder, Ben Mann, downloaded Books3, an online library of 196,640 books that he knew had been assembled from unauthorized copies of copyrighted books — that is, pirated.").

666. *See Judge Alsup's Solomonic Judgment on Fair Use in AI Training & Acquiring Pirated Books: Is It the Blueprint for the Future of AI Training? Part I: Pirated Copies*, CHAT GPT IS EATING THE WORLD (June 25, 2025), https://chatgptiseatingtheworld.com/2025/06/25/judge-alsups-solomonic-judgment-on-fair-use-in-ai-training-acquiring-pirated-books-is-it-the-blueprint-for-the-future-of-ai-training-part-i-pirated-copies/ [https://perma.cc/G5NC-946R]; Kyle Wiggers, *AI Training Data Has a Price Tag That Only Big Tech Can Afford*, TECHCRUNCH (June 1, 2024, 06:00 PT), https://techcrunch.com/2024/06/01/ai-training-data-has-a-price-tag-that-only-big-tech-can-afford/ [https://perma.cc/HND9-AAZX].

some unauthorized copies, either scraped from the internet or downloaded from online repositories? If no unauthorized acquisition of data is legal or allowed, then it might dampen AI research in the United States. Judge Alsup's opinion is equivocal on this issue, at times stressing the importance of the "initial copy" the defendant obtained, while other times distinguishing fair use decisions that allowed the unauthorized copying of online content for search engines, including pirated copies of images.[667] In any event, 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."[668]

Given the justification for AI researchers at universities and AI companies to engage in scaling techniques with large datasets used to train AI models, taking a more flexible, case-by-case approach to the use of pirated books is prudent.[669] A court could find that a defendant's use of shadow libraries was limited to serve the further purpose of developing better AI models for the public's benefit and was undertaken with sufficient guardrails to avoid widespread dissemination of any copies. In such case a court would be justified in finding the defendant's use of shadow libraries does not weigh against fair use. Conversely, if a defendant's use of shadow libraries factors was undertaken without sufficient guardrails or if it did not reasonably serve the purpose of developing the AI model, a court could find that the defendant's use of shadow libraries weighs against fair use. As Judge Chhabria suggested in *Kadrey v. Meta*, the use of pirated books datasets from shadow libraries in a way that bolstered their further use by others may weigh against fair use as a market harm if substantiated by evidence, including when the purpose of use was transformative.[670]

667. *Compare Bartz*, 2025 WL 1741691, at *12–14 (initial copies), *with id.* at *13–14 (distinguishing Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1164 n.8 (9th Cir. 2007) as a case involving "copies of images that had been pirated by third-party websites were used to index those same websites while indexing the entire web," whereas Anthropic's point was to build a library of pirated books).

668. *See id.* at *2, *18 ("But the test requires that we contemplate the likely result were the conduct to be condoned as a fair use — namely to steal a work you could otherwise buy (a book, millions of books) so long as you at least loosely intend to make further copies for a purportedly transformative use (writing a book review with excerpts, training LLMs, etc.), without any accountability.").

669. *See, e.g.*, *supra* notes 266–68, 352–56.

670. Kadrey v. Meta Platforms, Inc., No. 23-cv-03417-VC, 2025 WL 1752484, at *12, *14 (N.D. Cal. June 25, 2025).

C. All Copyrighted Works for AI Training Must Be Licensed?

The plaintiffs in the copyright lawsuits will argue against fair use due to their lost licensing of their works, which should allegedly count as both substitution and market harm. But this argument risks circularity and can always be asserted for every unauthorized use in which fair use is raised. The same argument was raised in *Google*, yet the Court nonetheless ruled that the jury had ample evidence to conclude Google's use was different in purpose, that is, to develop a new technology for smartphones, and the market effects also weighed in favor of fair use.[671] As *Google* instructs, "we must take into account the public benefits the copying will likely produce."[672]

Some people may disagree with—or even detest—AI. It raises a host of serious, if not vexing, concerns for governments and societies, with some fearing massive job losses, if not human extinction. And the prospect that tech companies, big and small, can profit off the labor of individuals is probably distasteful, if not disgusting, to many.[673] I get that. But the plaintiffs in the lawsuits will have their chance to present evidence of cognizable market harm to their works—and courts will weigh it in the balance of fair use. As the Supreme Court has oft recognized, "The primary objective of copyright is not to reward the labor of authors, but '[t]o promote the Progress of Science and useful Arts.'"[674]

Progress in the United States is served by both the production of new works and the development of new technologies.[675] The progress in developing AI in the United States occurs amidst fierce competition with China to develop the world's leading AI models.[676] This Article has explained why the use of copyrighted works to train AI models serves a highly transformative purpose in creating innovative new technology central to this nation's top priority. Fair use is a flexible doctrine, affording courts the ability to balance competing interests, including the

671. Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1203–04 (2021).

672. *Id.* at 1206.

673. *See, e.g.*, *Elton John Warns of AI Stealing Artists' Profits*, GEO NEWS (Jan. 26, 2025), https://www.geo.tv/latest/587474-elton-john-warns-of-ai-stealing-artists-profits [https://perma.cc/UAN4-SL9Z].

674. Feist Publ'ns, Inc. v. Rural Tele. Serv. Co., 499 U.S. 340, 349 (1991) (second alteration in original).

675. *See supra* notes 231–39 and accompanying text.

676. *See supra* notes 11–29 and accompanying text.

larger public interest.[677] Heeding the Supreme Court's fair use precedents counsels caution—and avoiding any rigid, categorical approach that might render all AI training, both at companies and universities, unlawful. To borrow the *Google* Court's sage words, "Given the rapidly changing technological, economic, and business-related circumstances, we believe we should not answer more than is necessary to resolve the parties' dispute."[678] And, as the White House AI Czar David Sacks advised, "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."[679]

VI. CONCLUSION

This Article explains how courts in the litigation against AI companies should analyze Factor 1 of fair use, drawing on the Supreme Court's recent fair use decisions in *Warhol* and *Google*, along with the Court's other technology fair use decision in *Sony*. *Warhol* instructs that courts must analyze fair use, *use-by-use*. Use to train and develop an AI model is one use. Use to generate an allegedly infringing copy is another use. Each must be analyzed separately. In analyzing the use in AI training, courts should consider the history of using datasets with copyrighted works by universities researchers, who originated the practice. This history illuminates that the practice had an important technological reason for such use of works: using larger and more diverse datasets—or the technique of *scaling up*—led to better results in developing AI models that actually worked and advanced the state of the art. This history shows that an AI developer's use of copyrighted works to train AI models to research, develop, train, and improve an AI model can serve a legitimate fair use purpose: namely, to create a new technology with public benefits. But the outputs of AI generators, once deployed to the public, must be analyzed on a use-by-use basis. Some outputs, such as regurgitated or memorized copies, may lack a fair use purpose and constitute copyright infringement. Finally, courts should reject the Copyright

677. *See Google*, 141 S. Ct. at 1206; *see also* Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 928 (2005) ("The more artistic protection is favored, the more technological innovation may be discouraged; the administration of copyright law is an exercise in managing the tradeoff."); Goldstein v. California, 412 U.S. 546, 559 (1973) ("Where the need for free and unrestricted distribution of a writing is thought to be required by the national interest, the Copyright Clause and the Commerce Clause would allow Congress to eschew all protection.").

678. *Google*, 141 S. Ct. at 1197.

679. David Sacks (@DavidSacks), X (June 24, 2025, at 12:10 CT), https://x.com/davidsacks/status/1937558998166954092 [https://perma.cc/F8NL-BL3D].

Office's pre-publication report's endorsement of a new, expansive theory of "copyright dilution" that penalizes, under fair use, people's creation of new, non-infringing works using AI. The Constitution does not permit authors to assert copyright beyond "their respective writings" to penalize the non-infringing writings of others. Copyright is intended to be the engine of free expression, not its enemy.