AI's Data Dilemma: Training Generative AI with Copyrighted Data

By Kaitlyn Chambers*

Introduction

Generative artificial intelligence is changing the world. ChatGPT, self-driving cars, medical imaging, autonomous delivery bots, content creation—the list goes on and on. Society interacts with generative artificial intelligence ("gen AI") daily, and its uses—and controversies—are growing exponentially. AI, however, has grown faster than the law, and the American legal system has not yet caught up.[1] As AI models become more common and sophisticated, it is important to establish guidelines and regulations that allow both developers and creators to protect their work while also promoting the rights and safety of consumers. To ensure the fair, safe, and innovative use of generative AI, the United States needs to pass fair training legislation permitting the use of copyrighted material in AI training sets, provided certain conditions are met.

AI copyright issues can be divided into two distinct categories: (1) copyrighted data used to train AI models, and (2) copyright protection for works created using generative AI.[2] This Paper focuses on the first category, addressing copyright issues related to the training of generative AI models and ultimately showing that AI training is fair use and should be protected. Congress should pass fair training legislation to formalize and ensure this training right. This Paper will lay out why such a law is necessary and what the law should include.

---

* J.D. Candidate, University of San Francisco School of Law, 2026; B.A., Global Studies, University of Minnesota, 2020. Thank you to *U.S.F. Law Review*, especially Jordan Mitchell, for insightful edits and supporting this Comment and to my family for their encouragement and support throughout my three years of law school. This Comment is dedicated to the memory of my grandfather, Richard Breiter, my first editor.

1. Prac. L. Intell. Prop. & Tech., Thompson Reuters, Expert Q&A on Artificial Intelligence (AI) Licensing 2 (2019).

2. Julia Yoon & Chris Borges, *Informing the Innovation Policy Debate: Key Concepts in Copyright Laws for Generative AI*, Ctr. for Strategic & Int'l Stud. (Apr. 12, 2024), https://www.csis.org/blogs/perspectives-innovation/informing-innovation-policy-debate-key-concepts-copyright-laws [https://perma.cc/C9NE-GVLR].

Part I begins by discussing generative AI and copyright basics. Part II analyzes common law copyright and fair use issues in training generative AI models, giving particular focus to the recent decision in *Thomson Reuters v. Ross Intelligence*, and ultimately concluding that gen AI training is fair use. Part III considers international legislation regulating AI training and what approaches should be adopted in the United States. Part IV lays out proposed legislation for gen AI training that protects developers, copyright owners, and end-users. Finally, Part V concludes that fair training legislation is necessary to promote safe, fair, accurate, and accessible generative AI models.

I. Background

Artificial intelligence and copyright law are complex fields. Section A provides a high-level overview of generative AI. Section B discusses copyright basics in the United States.

A. Generative Artificial Intelligence

Artificial intelligence is a general term for computer systems and machines that can simulate human learning, reasoning, decision-making, problem-solving, and creativity.[3] Generative AI refers to technology that can create original content.[4] Gen AI tools are built using machine learning ("ML") and deep learning.[5]

ML, developed in the 1980s, involves creating models by "training an algorithm to make predictions . . . based on data."[6] This enables computers to learn patterns and make decisions without being explicitly programmed.[7] ML algorithms analyze data, identify patterns, and improve their learning over time.[8] Traditional ML models make predictions but do not create new data.[9] Deep learning developed in the

3. *What Is Artificial Intelligence?*, NASA (May 13, 2024), https://www.nasa.gov/what-is-artificial-intelligence/ [https://perma.cc/D68M-C9TT]; Cole Stryker & Eda Kavlakoglu, *What Is Artificial Intelligence (AI)?*, IBM, https://www.ibm.com/think/topics/artificial-intelligence [https://perma.cc/S4YE-822Z] (last visited Oct. 4, 2025).

4. Stryker & Kavlakoglu, *supra* note 3.

5. *Id.*

6. *Id.*

7. *Id.*

8. *Id.*

9. *See* Adam Zewe, *Explained: Generative AI*, MIT News (Nov. 9, 2023), https://news.mit.edu/2023/explained-generative-ai-1109 [https://perma.cc/6BWB-U2ZC] ("Generative AI can be thought of as a machine-learning model that is trained to create new data, rather than making a prediction about a specific dataset."); Arun Nanda, *Generative vs Discriminative*

2010s as a subset of ML, employing "deep neural networks[] that more closely simulate the complex decision-making power of the human brain."[10] Deep neural networks extract information and relationships from large volumes of unstructured data and make predictions about what that data represents without requiring human intervention.[11]

Generative AI, the focus of this Paper,[12] is built using deep learning techniques and, unlike earlier models, can create original content.[13] Using neural network-based models, gen AI produces outputs that mimic human creativity and originality.[14] Models are trained on large datasets, and they learn to recognize patterns in the data.[15] Based on their learning, the systems build predictive models that can generate new content from prompts or inputs.[16] "At a high level, generative models encode a simplified representation of their training data, and then draw from that representation to create new work that's similar, but not identical, to the original data."[17]

Generative AI is transforming industries like art, design, entertainment, education, and software development by enhancing creativity, automating tasks, and personalizing user experiences.[18] However, it also poses ethical and intellectual property challenges and can facilitate the spread of misinformation or deceptive content.[19]

Models: Differences & Use Cases, DataCamp (Sep. 2, 2024), https://www.datacamp.com/blog/generative-vs-discriminative-models [https://perma.cc/Z3NV-BLUY] (comparing traditional discriminative ML models to generative models and noting that discriminative models "cannot generate new data points").

10. Stryker & Kavlakoglu, *supra* note 3.

11. *Id.*

12. This Comment discusses generative AI only (AI that creates new content itself). This will be an important distinction when it comes to *Thomson Reuters v. Ross Intelligence*. *See* discussion *infra* Section II.C.

13. *See, e.g.*, Kim Martineau, *What Is Generative AI?*, IBM (Apr. 20, 2023), https://research.ibm.com/blog/what-is-generative-AI [https://perma.cc/2XBT-VVAZ]; *Generative Artificial Intelligence (AI)*, Harv. Univ.: Info. Tech., https://www.huit.harvard.edu/ai [https://perma.cc/2LHW-WNYA] (last visited Oct. 8, 2025).

14. *See* Abid Ali Awan, *What Are Neural Networks?*, DataCamp (Aug. 30, 2023), https://www.datacamp.com/blog/what-are-neural-networks [https://perma.cc/79GE-JPRV].

15. Zewe, *supra* note 9.

16. *Id.*

17. Stryker & Kavlakoglu, *supra* note 3.

18. *Generative AI Uses: Transforming Industries and Enhancing Creativity*, Tekrowe (Mar. 6, 2025), https://www.tekrowe.com/blogs/post/generative-ai-uses-transforming-industries-enhancing-creativity [https://perma.cc/S9ZY-RSUE].

19. Neelam Koshiya, *Ethics in Generative AI: Addressing Bias, Misinformation, and Intellectual Property Challenges*, Sci. Times (Feb. 23, 2025, at 20:27 ET), https://www.sciencetimes.com/articles/60309/20250223/ethics-generative-ai-addressing-bias-misinformation-intellectual-property-challenges.htm [https://perma.cc/LXX2-NU83]; Araz Taeihagh, *Governance of Generative AI*, 44 Pol'y & Soc'y 1, 2 (2025); Jamey Keaten, *AI Developers Should Counter Misinformation and Protect Fact-Based News*, AP News (May 5, 2025, at 14:56 CT),

B. Copyright Basics

In the United States, copyright protection is grounded in the Constitution.[20] The Patent and Copyright Clause grants Congress the express power "[t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."[21] Copyright law protects "[o]riginal works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."[22] Original works of authorship may include "literary, dramatic, musical, and artistic works, such as poetry, novels, movies, songs, computer software, and architecture."[23] The Copyright Office has made it clear that to be copyrighted, a work requires *human* authorship.[24] Accordingly, gen AI output is unprotected if it lacks significant human involvement.[25]

Copyright protects the *expression* of a work, and there is no copyright protection for ideas or facts.[26] Registration is not required for protection to exist—copyright exists from the moment a work is created and fixed in a tangible form.[27] Copyright provides a creator with exclusive rights of reproduction, adaptation or preparation of derivative works, distribution, public performance, public display, and public performance by digital-audio transmission.[28] When a party copies a work without permission for any of the purposes listed above, copyright has been infringed.

---

https://apnews.com/article/ebu-wan-ifra-ai-misinformation-news-tech-61fb43f20d945753a8c86881aa631d65 [https://perma.cc/Z7V9-5TXR].

20. U.S. Const. art. I, § 8, cl. 8.

21. *Id.*

22. 17 U.S.C. § 102(a).

23. *Copyright in General*, U.S. Copyright Off., https://www.copyright.gov/help/faq/faq-general.html [https://perma.cc/R275-XNZH] (last visited Oct. 4, 2025).

24. Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190 (Mar. 16, 2023) (to be codified at 37 C.F.R. pt. 202).

25. *See* Thaler v. Perlmutter, 687 F. Supp. 3d 140 (D.C. Cir. 2023) (analyzing the Copyright Officer's denial of a copyright application on the ground that copyright law extends only to works created by human beings); Letter from Robert J. Kasunic, Assoc. Reg. of Copyrights & Dir. of Registration Pol'y & Prac., U.S. Copyright Off., to Van Lindberg, Att'y for Kristina Kashtanova (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf [https://perma.cc/MYG7-NP6A] (finding that AI program users lack sufficient control over generated images and AI output could not be predicted, so AI output does not have enough human authorship to be copyrightable).

26. 17 U.S.C. § 102(b). This provision embodies the idea-expression dichotomy.

27. *Copyright in General*, *supra* note 23.

28. 17 U.S.C. § 106(1)–(6).

II. Generative AI Training Is Not Copyright Infringement

To respond effectively to user prompts, sophisticated generative AI models must be trained on large amounts of data.[29] ChatGPT-3.5, a popular large language model ("LLM"), was trained using more than 570 gigabytes of filtered text data, which "is equivalent to roughly 300 billion words or 1.3 million books."[30] The majority of data used in training sets is copyrighted.[31]

This Part analyzes the copyright issues that arise when training gen AI models. Specifically, this Part will demonstrate that (A) training gen AI on a wide variety of data benefits society and is aligned with the purposes of copyright; (B) gen AI training is fair use, and a fair training law is needed to protect this right; (C) the first AI fair use decision, *Thomson Reuters v. Ross Intelligence*, supports the argument that gen AI training is fair use, despite ruling against the AI developer; and (D) in June 2025 the Northern District of California provided early favorable fair use rulings for gen AI developers.

A. The Purpose of Copyright and AI's Societal Benefits

The purpose of copyright law is to promote the progress of science and useful arts by providing authors and inventors with exclusive rights to their creations for a limited time.[32] This incentivizes the creation of new works by offering a financial reward and ultimately benefits the public by expanding access to knowledge and creative works once the period of exclusive control expires.[33]

According to some experts, the issue is that "[o]ur basic copyright doctrines don't fit generative AI."[34] However, the use of copyrighted material in AI training sets *is* aligned with the purposes and ideals of copyright law, and there are strong policy reasons to allow such training.

29. Yoon & Borges, *supra* note 2.

30. *Id.*

31. Adam Buick, *Copyright and AI Training Data—Transparency to the Rescue?*, 20 J. INTELL. PROP. L. & PRAC. 182, 186 (2025) ("[A] large majority of the content used to train most generative AI models will be protected by the law of copyright. This is an inevitable result of the fact that copyright arises automatically for any work that meets a minimal set of requirements . . . . [W]hile there is a large body of public domain works for which copyright has expired, almost all of these works will date from before the 1950s—any AI trained exclusively on such works would therefore be hopelessly outdated.").

32. U.S. CONST. art. I, § 8, cl. 8.

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

34. Mark A. Lemley, *How Generative AI Turns Copyright Upside Down*, 25 COLUM. SCI. & TECH. L. REV. 190, 212 (2024).

First, society benefits when AI is trained on a wide variety of sources, as limiting access to data could lead to biased models and outputs. The quality and quantity of training data directly affect the accuracy of the AI model.[35] Overly strict copyright protection negatively affects society because developers, who might make valuable uses of copyrighted works, may be reluctant to do so for fear of infringement and penalties.[36] Second, the AI models' use of the data is transformative and used for different purposes and in different markets than the underlying work. Third, training sets copy the non-protected aspects of a work—the ideas, facts, methods, and functions—not the protectable expressions, qualifying as fair use.[37] Fourth, a licensing regime is overburdensome and implausible because of the millions of works and authors in training sets, and licensing requirements often lead to biased models that harm society.[38]

The use of copyrighted works to train gen AI is fair use, but copyright owners' concerns should not be taken lightly. Inevitably, training sets involve copying of entire works without alteration. Often, gen AI models are used for commercial purposes, and some copyright owners argue that for-profit developers should pay for the copyrighted works they use to create profitable outputs.[39] Critics have also expressed concerns that gen AI models with unlimited ability to consume copyrighted materials could spread propaganda, misinformation, and deep fakes;[40] perpetuate and amplify biases;[41] contribute to worker exploitation[42] or displacement;[43] invade privacy;[44] exacerbate global warming and other environmental problems;[45] or even threaten humanity and take over the

35. Andrew W. Torrance & Bill Tomlinson, *Training Is Everything: Artificial Intelligence, Copyright, and "Fair Training,"* 128 DICK. L. REV. 233, 242 (2023).

36. *See* Mark S. Nadel, *How Current Copyright Law Discourages Creative Output: The Overlooked Impact of Marketing*, 19 BERKELEY TECH. L.J. 785, 803–06 (2004).

37. *See* Torrance & Tomlinson, *supra* note 35, at 245.

38. *See infra* Section IV.A.

39. Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 765 (2021).

40. *Strengths and Weaknesses of Gen AI*, UNIV. OF LEEDS, https://generative-ai.leeds.ac.uk/intro-gen-ai/strengths-and-weaknesses/ [https://perma.cc/DK94-MSF8] (last visited Oct. 4, 2025); Carlo Salvato, *Challenges and Negative Aspects of Generative AI: A Comprehensive Analysis*, MEDIUM (May 18, 2024), https://medium.com/al-game-code/challenges-and-negative-aspects-of-generative-ai-a-comprehensive-analysis-09226994b43e [https://perma.cc/BH65-KQM9].

41. Salvato, *supra* note 40.

42. *Strengths and Weaknesses of Gen AI*, *supra* note 40.

43. George Lawton, *Generative AI Ethics: 11 Biggest Concerns and Risks*, TECHTARGET (Mar. 3, 2025), https://www.techtarget.com/searchenterpriseai/tip/Generative-AI-ethics-8-biggest-concerns [https://perma.cc/E4VF-G2SB].

44. *Id.*

45. Kate Crawford, *Generative AI's Environmental Costs Are Soaring—And Mostly Secret*, NATURE (Feb. 20, 2024), https://www.nature.com/articles/d41586-024-00478-x [https://perma.cc/EVF5-QPQ3].

world.[46] While all (or at least most) of these concerns may be legitimate, copyright law is not the solution.

Although copyright serves many important purposes, in terms of the concerns mentioned above, "copyright is not the right tool to regulate those abuses."[47] In other words, copyright litigation is being used as a weapon to target AI developers for concerns that are protected by the fair use defense or fall outside the scope of copyright altogether. These concerns are better addressed through alternative means, such as rules surrounding gen AI *output* (rather than input), the extent to which AI may be used, employment protection laws, privacy laws, and environmental regulation.[48] Targeting training is not the solution, especially since greater access to a variety of sources increases accuracy, safety, and fairness.

B. Gen AI Training Satisfies the Fair Use Test

Fair use is an affirmative defense to copyright infringement. This defense protects gen AI training sets from copyright infringement claims. The fair use exception provides a judicial solution to gen AI training; however, to better protect developers, ensure unbiased models for the benefit of society, and to preempt litigation and free up the courts, Congress should create a legislative solution that codifies the judicial fair use doctrine and secures a "fair training" safe harbor for gen AI training.

In considering whether the copying of a protected work is fair use, courts consider four factors: (1) the purpose and character of the use,

46. James Grimmelmann, *Copyright for Literate Robots*, 101 Iowa L. Rev. 657, 676–78 (2016).

47. Lemley & Casey, *supra* note 39, at 749.

48. *See, e.g.*, Taeihagh, *supra* note 19, at 6 (emphasizing that addressing the risks of gen AI requires broader governance rather than solely technical aspects of training inputs because "[t]he traditional regulatory frameworks for data protection, intellectual property, and the like were often defined based on human agency and are unable to adequately address the complexities of generative AI"); Xiongbiao Ye et al., *Privacy and Personal Data Risk Governance for Generative Artificial Intelligence: A Chinese Perspective*, 48 Telecomm. Pol'y, no. 10, 2024, at 8–12 (underscoring that existing legal models are insufficient for generative AI and that new or strengthened privacy laws are necessary to protect individuals from risks arising from AI outputs and data use, not just data collection); Tsai-Fang Chen, *AI, Job Displacement, and the WTO: Identifying Legal Gaps and Charting a Worker-Centered Path*, Yale J. Int'l L. (Nov. 10, 2025), https://yjil.yale.edu/posts/2025-11-10-ai-job-displacement-and-the-wto-identifying-legal-gaps-and-charting-a-worker [https://perma.cc/T3AN-3SGP] (proposing a labor-sensitive framework, including prohibitions, import restrictions, tariffs, safeguard measures, and subsidies, to protect workers from job displacement in light of AI); Mona Sloane & Elena Wüllhorst, *A Systematic Review of Regulatory Strategies and Transparency Mandates in AI Regulation in Europe, the United States, and Canada*, 7 Data & Policy e11, e11-9 (2025).

including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.[49] All factors must be considered, though the weight given to each factor varies on a case-by-case basis.[50] In AI training cases, the bulk of the analysis focuses on the first and fourth factors,[51] though each is discussed in turn below.

1. Factor One: Purpose and Character of the Use

Factor one analyzes the purpose and character of the use.[52] This factor, according to the Supreme Court, turns on "whether the new use merely 'supersedes the objects' of the original creation . . . or instead adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message."[53] In other words, the first factor largely examines the extent to which the new work is "transformative," though a use "need not be transformative to be fair."[54] Arguably, if a use is non-expressive, it does not supersede or substitute for the original work, and the use is equated to a highly transformative use.[55]

Copyrighted works are used in AI training sets for learning purposes. AI models have no interest in particular art or artistic choices.[56] As Mark Lemley and Bryan Casey put it, a training set needs

> photos of stop signs so it can learn to recognize stop signs, not because of the artistic choices you made in lighting or composing your photo. Similarly, a[n] [AI model] wants to see what you wrote to learn how words are sequenced in ordinary conversation, not because it finds your prose particularly expressive or because it wants to use your turn of phrase.[57]

49. 17 U.S.C. § 107.

50. Kadrey v. Meta Platforms, 788 F. Supp. 3d 1026, 1059 (2025).

51. Emily Poler, *AI & Copyright: What's 'Fair' Game?*, *in* Month-In-Brief: Bus. Regul. & Regulated Indus. 23 (Feb. 2024); Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc., 765 F. Supp. 3d 382, 397 (D. Del. 2025) (citing Authors Guild v. Google, Inc., 804 F.3d 202, 220 (2d Cir. 2015)).

52. 17 U.S.C. § 107(1).

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

54. Lemley & Casey, *supra* note 39, at 783.

55. Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 Nw. U. L. Rev. 1607, 1647 (2009).

56. Lemley & Casey, *supra* note 39, at 773.

57. *Id.*

Copyright law does not protect ideas, facts, functions, or methods of a work.[58] Training sets use the *unprotected* portions of copyrighted works—the idea behind the work, not the expression itself.[59] Such use is unprotected by copyright law.[60] Furthermore, data mining serves functional, non-expressive purposes, as copyrighted works in training sets are used for technical, noncommunicative purposes.[61] Courts have consistently held that non-expressive use by search engines is not copyright infringement.[62] As with search engines, AI training sets use copyrighted work internally, and the original works are not shared with the public.[63] Copyright owners may argue that training sets are not fair use because AI models are "consumer[s] of the copyrighted work, outputting a profitable technology rather than new creativity."[64] However, this argument is beyond the scope of copyright law. In these cases, courts are focused on the *use* of the copyrighted work, and an AI model's use of the work is ultimately for learning and training purposes (whether or not the models are one day of a commercial nature).[65] Therefore, factor one ultimately supports the argument that using copyrighted works to train gen AI models constitutes fair use.

2. Factor Two: Nature of the Original Work

Factor two, the nature of the copyrighted work, "has rarely played a significant role in the determination of a fair use dispute."[66] The more factual a work, the lower the degree of protection. However, AI training sets consist of so many works that fall under some degree of copyright protection that this factor carries little weight.[67]

58. 17 U.S.C. § 102(b).

59. Jenny Quang, *Does Training AI Violate Copyright Law?*, 36 Berkeley Tech. L.J. 1407, 1419–20 (2021).

60. Google LLC v. Oracle Am., Inc., 143 S. Ct. 1183, 1196 (2021).

61. Quang, *supra* note 59, at 1409, 1420.

62. *See* Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1523, 1527–28 (9th Cir. 1992), *as amended* (Jan. 6, 1993) (drawing a "distinction between the copying of works in order to make independent creative expression possible and the simple exploitation of another's creative efforts"). Where copying is the "only way to gain access to the ideas and functional elements embodied in a copyrighted computer program and where there is a legitimate reason for seeking such access, [copying] is a fair use of the copyrighted work, as a matter of law." *Id.*; *see also* Authors Guild v. Google, Inc., 804 F.3d 202, 225 (2d Cir. 2015) ("Google's making of a complete digital copy of Plaintiffs' works for the purpose of providing the public with its search and snippet view functions (at least as snippet view is presently designed) is a fair use and does not infringe Plaintiffs' copyrights in their books.").

63. Lemley & Casey, *supra* note 39.

64. *Id.*

65. *Id.* at 765–66.

66. *Authors Guild*, 804 F.3d at 220.

67. *Id.*; Kadrey v. Meta Platforms, 788 F. Supp. 3d 1026, 1049 (N.D. Cal. 2025).

3. Factor Three: Amount and Substantiality of the Portion Used

Training sets copy entire works without alteration, directly implicating fair use factor three, the amount and substantiality of the portion used in relation to the copyrighted work as a whole. However, as the Second Circuit has noted, what matters here is not "the amount and substantiality of the portion used *in making a copy*, but rather the amount and substantiality of *what is thereby made accessible* to a public for which it may serve as a competing substitute."[68] Generally, the copyrighted data used in training sets is never made accessible to the public.[69] Output from the final AI model will not serve as a competing substitute for the original work because a consumer cannot access the original work from the model.[70] This factor, therefore, weighs in favor of the AI developers since the amount and substantiality of the copyrighted work made accessible to the public is nonexistent or minimal.

4. Factor Four: Effect of the Use on the Potential Market for or Value of the Original Work

Factor four considers the effect of the use upon the potential market for or value of the copyrighted work. Training sets do not impact the potential market for or value of the underlying copyrighted work because the data is not shared with the public and does not compete with the work.[71] For example, a gen AI model trains on a book to learn the rules of syntax and the flow of the English language, not to copy that book and provide it in whole or in part to the public. A consumer would still have to buy the book from the copyright owner or authorized seller. AI models are creating new markets and targeting different consumers, and there is little to no direct competition with copyright owners.[72] This factor may be more of an issue when AI models "memorize" training data[73] and regurgitate copyrighted works in their outputs, but copyright infringement in AI output is a separate

68. *Authors Guild*, 804 F.3d at 222 (internal quotation marks omitted).

69. Engine Advocacy, Comment Letter in Response to Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, Docket No. PTO-C-2019-0038 (Jan. 10, 2020), https://www.uspto.gov/sites/default/files/documents/Engine%20Advocacy_RFC-84-FR-58141.pdf [https://perma.cc/G88H-SC2A].

70. *Id.*

71. Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc., 765 F. Supp. 3d 382, 400 (D. Del. 2025).

72. *Id.*

73. *See* Matthew Sag, *Copyright Safety for Generative AI*, 61 Hous. L. Rev. 295, 326 (2023).

issue from infringement in training sets and is outside the scope of this Paper.[74] Infringement in outputs should not preclude fair use of copyrighted works in training sets, as input and output are two entirely separate acts.

Furthermore, courts have found that scanning and indexing content for search purposes can be fair use and does not affect the market for such content.[75] Similarly, AI training is an intermediary process that enables innovation. For all these reasons, factor four favors the fair use argument for gen AI training.

Ultimately, the use of copyrighted works to train generative AI models is fair use (or, more aptly, fair training) under U.S. copyright law. Yet copyright owners continue to sue AI developers, arguing that training is not fair use, as seen in the following case.

C. *Thomson Reuters v. Ross Intelligence*: The First AI Fair Use Decision

On February 11, 2025, the District Court of Delaware shook the tech world by declaring that Ross Intelligence's copying of Westlaw headnotes to train its machine-learning platform was not fair use.[76] While many AI cases remain pending,[77] the *Thomson Reuters* case is the first decision addressing whether the use of copyrighted works in AI training qualifies as fair use. While copyright owners heralded the decision as a major win and the public was stirred into a frenzy over the perception that AI training is not fair use,[78] a close reading of the case

74. For discussion of copyright infringement in AI outputs, see, for example, Mark A. Lemley, *How Generative AI Turns Copyright Upside Down*, 25 COLUM. SCI. & TECH. L. REV. 190, 190 (2024); Vicenç Feliú, *Our Brains Beguil'd: Copyright Protection for AI Created Works*, 25 U.S.F. INTELL. PROP. & TECH. L.J. 105, 113–20 (2021); Mackenzie Caldwell, Note, *What Is an "Author"?—Copyright Authorship of AI Art Through a Philosophical Lens*, 61 HOUS. L. REV. 411, 437–41 (2023); Dylan Jignesh Patel, *Authored by Artificial Intelligence: An Analysis of AI Use in Copyright*, 47 AM. J. TRIAL ADVOC. 419, 419 (2024).

75. Authors Guild v. Google, Inc., 804 F.3d 202, 224 (2d Cir. 2015); Authors Guild v. HathiTrust, 755 F.3d 87, 100 (2d Cir. 2014).

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

77. *See Master List Copyright Lawsuits v. AI Companies in U.S.*, CHAT GPT IS EATING THE WORLD (Aug. 27, 2024), https://chatgptiseatingtheworld.com/2024/08/27/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/ [https://perma.cc/M8CU-HCGD]; Julia Barnett, *The Current State of GenAI Copyright Lawsuits*, MEDIUM (Sep. 18, 2024), https://generative-ai-newsroom.com/the-current-state-of-genai-copyright-lawsuits-203a1bd0f616 [https://perma.cc/448A-NEAM].

78. *See* Kevin Madigan, *First of Its Kind Decision Finds AI Training Is Not Fair Use*, COPYRIGHT ALL. (Feb. 12, 2025), https://copyrightalliance.org/ai-training-not-fair-use/ [https://perma.cc/QY6H-ZVXL] ("AI Training Is Not Fair Use . . . Judge Bibas's opinion is the first decision in a case that addresses whether AI training is fair use, and it unequivocally holds that it is not."); Joseph A. Meckes & Joseph Grasser, *Court: Training AI Model Based on*

reveals that the holding does not implicate *generative* AI training. In fact, the decision may actually strengthen the argument that generative AI training is fair use. Copyright owners will use this case in their prosecution of AI developers, but AI developers can also use it to distinguish their gen AI models.

Thomson Reuters, a legal-research platform, sued Ross Intelligence, a startup developing an AI legal-research tool, alleging copyright infringement.[79] To train its AI search tool, Ross sought to license Thomson Reuters' Westlaw content, but after Thomson Reuters refused, Ross acquired thousands of legal question-and-answer compilations from a third party, which included material derived from Westlaw headnotes.[80] When Thomson Reuters found out, it sued Ross.[81]

Most importantly, Ross's AI model was *not* generative AI. Judge Bibas stressed this at multiple points throughout the opinion.[82] The court's rationale for infringement in this context of machine-learning training does not carry over to gen AI. To start, Ross's AI is distinguishable from the gen AI models at issue in this Paper because Ross did not use the same techniques employed by gen AI. Ross did not directly ingest (or "data mine") the Westlaw headnotes. Rather, when Thomson Reuters refused to license the Westlaw content to Ross, Ross hired a third party to use the Westlaw data to create Bulk Memos, which were then used to train Ross's model.[83]

Analyzing the first fair use factor, Judge Bibas ruled that "Ross's use is not transformative because it does not have a 'further purpose or different character' from Thomson Reuters's."[84] Rather, "Ross was using [the] headnotes as AI data to create a legal research tool to compete with Westlaw," and its (non-generative) AI model "spits back relevant judicial opinions that have already been written" in a process

Copyrighted Data Is Not Fair Use as a Matter of Law, NAT'L L. REV. (Feb. 11, 2025), https://natlawreview.com/article/court-training-ai-model-based-copyrighted-data-not-fair-use-matter-law [https://perma.cc/EN54-8RJH] ("Training AI model[s] based on copyrighted data is not fair use as a matter of law."). *But see* Tori Noble, *Artificial Intelligence, Copyright, and the Fight for User Rights: 2025 in Review*, ELEC. FRONTIER FOUND. (Dec. 25, 2025), https://www.eff.org/deeplinks/2025/12/artificial-intelligence-copyright-and-fight-user-rights-2025-review [https://perma.cc/666W-CPU3] (explaining that as these cases finished their trials and moved to appeals, copyright owners increased their efforts to control fair uses through various means to influence how the internet operates).

79. *Thomson Reuters*, 765 F. Supp. 3d at 391.

80. *Id.* at 390–91.

81. *Id.* at 391.

82. *Id.* at 398–99 ("Ross's AI is not generative AI . . . . Because the AI landscape is changing rapidly, I note for readers that only non-generative AI is before me today.").

83. *Id.* at 390–91.

84. *Id.* at 397–98.

that "resembles how Westlaw uses headnotes and key numbers to return a list of cases with fitting headnotes."[85] Ross copied the headnotes to "develop a competing legal research tool. So, Ross's use is not transformative."[86] Unlike Ross's AI, gen AI serves a further purpose and different character from the copyrighted material it trains on. In other words, gen AI's use is transformative, so this case is distinguishable from those involving gen AI. Unlike Ross's use, copying in gen AI training is necessary for innovation.

Factor two is less relevant here, as the judge favored Ross under this factor because the headnotes had creative elements but were far from the most creative works.[87] Factor three also went to Ross, and Judge Bibas's reasoning shows that factor three will tend to favor AI developers of any kind since the final product does not make copyrighted material available to the public.[88]

Ross lost on factor four because it used the Westlaw data to create "a legal-research search engine that uses artificial intelligence," and Ross and its product are direct competitors to Thomson Reuters' Westlaw.[89] Here too, gen AI cases are distinguishable because, unlike Ross's directly competing output, gen AI output is generally not a market substitute for copyrighted work.[90] Judge Bibas further concluded that "[t]here is nothing that Thomson Reuters created that Ross could not have created for itself or hired LegalEase to create for it without infringing Thomson Reuters's copyrights."[91] Whereas Ross could have easily created its own version of Westlaw headnotes to train on, gen AI developers cannot create the millions of books, photos, and videos used in their training models. Furthermore, Ross used the expressive elements of Westlaw headnotes to train its AI to learn legal principles, while gen AI uses the non-expressive, unprotected elements of copyrighted works to learn syntax, structure, and ideas.

85. *Id.* at 398.

86. *Id.* at 399.

87. *Id.*

88. *Id.* at 400 ("Ross's output to an end user does not include a West headnote. What matters is not 'the amount and substantiality of the portion used in making a copy, but rather the amount and substantiality of what is thereby made accessible to a public for which it may serve as a competing substitute.' *Authors Guild*, 804 F.3d at 222 (internal quotations omitted). Because Ross did not make West headnotes available to the public, Ross benefits from factor three.").

89. *Id.* at 390.

90. Engine Advocacy, *supra* note 69; Bartz v. Anthropic, 797 F. Supp. 3d 1007, 1032 (N.D. Cal. 2025).

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

Ultimately, *Thomson Reuters* shows that, unlike older machine-learning models, generative AI training is fair use. AI developers can use this case to their advantage, distinguishing their training sets from Ross's and strengthening the argument that gen AI training is transformative, non-expressive, and non-competing.

D. Generative AI Training in the Courts

While the Delaware District Court was the first court to address fair use in AI training, in June 2025, the Northern District of California became the first court to address fair use in the context of *generative* AI training. *Bartz v. Anthropic PBC* and *Kadrey v. Meta Platforms, Inc.* provided early, favorable fair use decisions for AI developers, with two different judges ruling in favor of gen AI developers. While these cases were a massive win for gen AI training, they did not create binding precedent, and both rulings were narrow and fact-specific.[92] The judge underscored in *Kadrey*, for example, that the ruling stood for "the proposition that these plaintiffs made the wrong arguments" and affected the rights of the authors bringing the case, not a wider group of authors whose works had been used to train gen AI models.[93] However, these cases lay the groundwork for future litigation and show why fair training legislation is warranted.

On June 23, 2025, the Northern District of California granted summary judgment for Anthropic, the developer of the gen AI model Claude, ruling that the use of copyrighted books to train Claude was fair use.[94] Notably, the court distinguished books acquired legally from pirated books, ruling that pirated books used to build a central library were not justified by fair use.[95] The court found that all factors except the nature of the copyrighted work favored fair use, and that gen AI technology is "among the most transformative many of us will see in our lifetimes."[96]

As predicted, following *Thomson Reuters*, the author-plaintiffs in *Bartz* tried to use *Thomson Reuters* to argue that Anthropic's training was not fair use.[97] The court distinguished the case at issue from *Thomson Reuters*, emphasizing that Ross's AI was not generative AI and

92. Kadrey v. Meta Platforms, 788 F. Supp. 3d 1026, 1036–37 (N.D. Cal. 2025) ("[I]n the grand scheme of things, the consequences of this ruling are limited.").

93. *Id.*

94. *Bartz*, 787 F. Supp. 3d at 1034.

95. *Id.* at 1033.

96. *Id.*

97. *Id.* at 1022.

its training use was thus not transformative.[98] Because Anthropic was using copyrighted data to train *generative* AI, *Thomson Reuters* did nothing to strengthen the *Bartz* authors' arguments, and the court held that "the purpose and character of using copyrighted works to train LLMs to generate new text was quintessentially transformative."[99] The court went even further, asserting that using copyrighted books to train gen AI is no different from using textbooks to train students to write well.[100]

Another fact highlighted repeatedly by the judge was that training Anthropic's LLM "did not result in any exact copies nor even infringing knockoffs of [Authors'] works being provided to the public," and the author-plaintiffs did not allege that "any infringing copy of their works was or would ever be provided to users by" the LLM.[101] For this reason, factor four highly favored Anthropic since "[t]he copies used to train specific LLMs did not and will not displace demand for copies of Authors' works," and the authors had no compelling counter-arguments.[102]

Just two days after the *Bartz* decision, the Northern District issued its decision in *Kadrey v. Meta*. In *Kadrey*, Meta's use of copyrighted books to train its gen AI model was highly transformative, and the court noted that factor one favored Meta because there was "no serious question that Meta's use of the plaintiffs' books had a 'further purpose' and 'different character' than the books."[103] Factor two favored the plaintiffs but played no significant role in the analysis.[104] Factor three also favored Meta: "even though it copied the plaintiffs' books in their entirety[,] [t]he amount that Meta copied was reasonable given its relationship to Meta's transformative purpose" since "[e]veryone agrees that LLMs work better if they are trained on more high-quality data" and "feeding a whole book to an LLM does more to train it than would feeding it only half of that book."[105]

98. *Id.*
99. *Id.*
100. *Id.* ("Like any reader aspiring to be a writer, Anthropic's LLMs trained upon works not to race ahead and replicate or supplant them—but to turn a hard corner and create something different."); *id.* at 1032 ("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 advance original works of authorship, not to protect authors against competition.").
101. *Id.* at 1018, 1031.
102. *Id.* at 1031.
103. Kadrey v. Meta Platforms, 788 F. Supp. 3d 1026, 1044 (N.D. Cal. 2025).
104. *Id.* at 1049–50.
105. *Id.* at 1050.

After analysis of the first three fair use factors, the balance weighed in favor of Meta, and "the plaintiffs needed to win decisively on the fourth factor" to defeat Meta's fair use argument.[106] While *Bartz* suggested the fourth factor weighs heavily in favor of gen AI developers, *Kadrey* suggests that plaintiffs have three main ways of arguing under the fourth factor: (1) "the model will regurgitate their works . . . [and allow] users to access those works or substitutes for them for free via the model;" (2) "unauthorized copying for training harms [the] market" for licensing their works; or (3) "even if the model can't regurgitate their own works or generate substantially similar ones, it can generate works that are similar enough (in subject matter or genre) that they will compete with the originals and thereby indirectly substitute for them."[107]

Meta was able to easily defeat the first theory of harm because throughout the training process, it implemented "mitigations" that prevented its gen AI models from "'memorizing' and outputting certain text from their training data, including copyrighted material."[108] This post-training was so successful that neither Meta's nor the plaintiffs' expert could get the model to regurgitate more than fifty words and punctuation marks from the plaintiffs' books.[109] When used by gen AI developers, this training technique protects the original works from copying and ensures that the model cannot be used to read, generate, or meaningfully access the copyright owner's works.[110]

The plaintiffs focused most of their argument on the theory that Meta's use of the copyrighted books harms the market for licensing books for gen AI training. The court ruled that "whether such a market exists or is likely to develop is irrelevant, because this market is not one that the plaintiffs are legally entitled to monopolize," and "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."[111]

The court noted that while the first two arguments clearly failed in this case, the third was more promising.[112] However, the plaintiffs'

106. *Id.* at 1060.
107. *Id.* at 1051.
108. *Id.* at 1041.
109. *Id.* at 1041–42.
110. *Id.* at 1042, 1051.
111. *Id.* at 1052; *see also* Bartz v. Anthropic, 797 F. Supp. 3d 1007, 1032 (N.D. Cal. 2025) (explaining that a market for licensing works for LLM training "is not one the Copyright Act entitles Authors to exploit").
112. *Kadrey*, 788 F. Supp. 3d at 1051.

argument was far too weak to counter Meta's.[113] Because the plaintiffs did not provide sufficient evidence, market dilution, while addressed by the court, could not be ruled on, and this leaves open issues for future litigation.[114] *Kadrey* did, however, add a new consideration to the fourth fair use factor: "when considering market dilution, the proper consideration isn't to a world with no LLMs, but to a world where LLMs weren't training on copyrighted books."[115] This adds new sub-elements to factor four: whether the model is capable of generating works that compete with copyrighted works through indirect substitution; what the AI-generated work is; the impact the competition actually has on sales of the works with which it competes; and the threat to the market for the rightsholders' works "in a world where [gen AI] developers can copy those [works] compare[d] to the threat to the market for the [rightsholders' works] in a world where the developers can't copy them."[116]

While a major win for Meta and other gen AI developers, *Kadrey* is quite narrow. The court reaffirmed that fair use is a "fact-specific doctrine that requires case-by-case analysis."[117] Meta's use of the copyrighted works to train its gen AI model was highly transformative, so to show that the training wasn't fair use, the plaintiffs needed a strong market harm argument under the fourth factor.[118] The plaintiffs failed to present any meaningful evidence on market dilution, so Meta's highly transformative use of the copyrighted data to train its gen AI models was determined to be fair use on summary judgment.[119]

Though narrow, *Bartz* and *Kadrey* will be useful for gen AI developers sued in court for copyright infringement. The developers and judges in both cases provided strong arguments that developers will be able to use in future litigation to show that training gen AI models on copyrighted works is fair use.

As this Part has shown, the use of copyrighted data in gen AI training sets benefits society and is fair use. Yet litigation and ambiguity persist. To prevent unnecessary litigation and free up the courts, the United States must take steps to definitively protect gen AI training and set up legislative guardrails for fair use of copyrighted works.

113. *Id.*
114. *Id.* at 1055–57.
115. *Id.* at 1053.
116. *Id.* at 1056.
117. *Id.* at 1059.
118. *Id.* at 1060.
119. *Id.*

To determine what Congress should include in a training law, it is helpful to look at regulations tested abroad.

III. International Regulation of AI Training

The issue of copyright infringement in AI training is not unique to the United States. This issue has been litigated, debated, and regulated all over the world with varying degrees of success.[120] This Part considers gen AI training regulation abroad, particularly in the European Union, the United Kingdom, Japan, Singapore, and Brazil. Considering the strengths and weaknesses of international legislation sheds light on how best to protect both AI developers and copyright owners and provides insights into how the United States should structure its AI training policies.

A. The European Union

The European Union Copyright Directive addresses copyright and related rights in the Digital Single Market. Article 3(1) permits "reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access."[121] Article 4(1) requires Member States to provide an exception for "reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining."[122] Article 2(2) defines text and data mining ("TDM") as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."[123] This includes extracting data for commercial TDM uses, but copyright owners may opt out of this exemption.[124] Thus, in the EU, unless a rightsholder has explicitly opted out, publicly available online material may be reproduced for

120. Sara Wright & Ben Hitchens, *Artificial Intelligence and Copyright Case Tracker*, CMS LEGAL (Oct. 31, 2025), https://cms.law/en/mco/publication/artificial-intelligence-and-copyright-case-tracker [https://perma.cc/B9VF-MJ4V]; *AI Regulations Around the World - 2026*, MIND FOUNDRY (Jan. 13, 2026), https://www.mindfoundry.ai/blog/ai-regulations-around-the-world [https://perma.cc/43FL-APNG].

121. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, 2019 O.J. (L 130) 92, 113.

122. *Id.* art. 4.

123. *Id.* art. 2(2).

124. *Id.* arts. 3–4 (stating rightsholders can only opt out of Article 4(1) and have no opt-out right for Article 3(1)).

TDM purposes. The EU has also retroactively extended its TDM exception to cover training activities occurring in non-EU countries once the completed model is placed on the EU market.[125]

While the Copyright Directive provides more clarity than American law, the limitations imposed by Article 4 (primarily the opt-out provision) disadvantage AI developers. The ability of copyright owners to opt out of data mining and AI training could propagate bias. Furthermore, critics argue that "the Directive effectively creates and legitimizes a derivative market for text and data mining, which right holders may wish to control, license or even entirely prohibit."[126] On the other hand, proponents argue that the licensing market for training data could be circumvented if AI developers rely on Article 3(1), which provides no possibility for rightsholders to reserve their rights or opt out if the data is used for scientific purposes.[127] Instead of being restricted by copyright owner opt-out rights under 4(1), AI developers could access training data from an open data repository that was created by a nonprofit organization for scientific research purposes.[128] This potential loophole in the Copyright Directive could lead to exploitation and litigation.[129]

A U.S. law must anticipate and safeguard against similar loopholes by allowing equally for both non-commercial and commercial uses of non-expressive elements of copyright data. Additionally, the U.S. law should limit licensing to prevent the creation of a derivative TDM market.

The EU AI Act requires that AI developers "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model."[130] Such a disclosure

125. Mattias Rattzén, *Location Is All You Need: Copyright Extraterritoriality and Where to Train Your AI*, 26 COLUM. SCI. & TECH. L. REV. 175, 254 (2024).

126. *See* P. Bernt Hugenholtz, *The New Copyright Directive: Text and Data Mining (Articles 3 and 4)*, KLUWER COPYRIGHT BLOG (July 24, 2019), https://legalblogs.wolterskluwer.com/copyright-blog/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ [https://perma.cc/CLK8-Z9CJ].

127. Rattzén, *supra* note 125, at 213.

128. *Id.*

129. *See* Jennifer Rankin, *EU Accused of Leaving 'Devastating' Copyright Loophole in AI Act*, GUARDIAN (Feb. 19, 2025, at 07:13 ET), https://www.theguardian.com/technology/2025/feb/19/eu-accused-of-leaving-devastating-copyright-loophole-in-ai-act?utm [https://perma.cc/K9H8-SWAW].

130. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), art. 53, 2024 O.J. (L) 1, 84 ("[T]his summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties

increases transparency and helps both rightsholders and the public to identify which works have been included in training sets.[131] It allows rightsholders to better protect and police their works, and it holds developers accountable, giving the public the opportunity to ensure that a model isn't relying on biased data or misrepresenting sources. Because it increases transparency, improves AI safety, and allows for monitoring of AI developers, the United States should include a similar public summary provision in its AI policy.

B. The United Kingdom

The United Kingdom has a similar TDM exception, though its exception is narrower than the EU's. Under section 29A of the Copyright, Designs and Patents Act 1988, copies made for TDM purposes do "not infringe copyright in the work provided that . . . the work [is] for the sole purpose of research for a non-commercial use" and the beneficiary had lawful access to the work.[132] Unlike the loophole created by Article 3(1) of the EU Copyright Directive, section 29A(2)(a) of the U.K. law provides that copyright is infringed if "the copy is transferred to any other person, except where the transfer is authorised by the copyright owner." As a result, AI developers cannot rely on third parties to collect data, and TDM by AI developers is only permitted for non-commercial research uses.[133]

In 2021, the U.K. government proposed a new TDM exception that benefited AI training sets and did not provide opt-out provisions or compensation for copyright owners.[134] The proposal received significant backlash and was withdrawn in 2023.[135] In December 2024,

with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used.").

131. *Artificial Intelligence—Questions and Answers*, EUR. COMM'N (Aug. 1, 2024), https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683 [https://perma.cc/96DZ-LTJP].

132. Copyright, Designs and Patents Act 1988, c. 48, § 29A (U.K.).

133. *Id.*

134. *Open Consultation: Artificial Intelligence and Intellectual Property: Copyright and Patents*, GOV. U.K.: INTELL. PROP. OFF. (Oct. 29, 2021), https://web.archive.org/web/20211101024732/ https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents [https://perma.cc/N95G-8U2L].

135. *Consultation Outcome: Artificial Intelligence and Intellectual Property: Copyright and Patents: Government Response to Consultation*, GOV. U.K.: INTELL. PROP. OFF. (June 28, 2022), https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation [https://perma.cc/FY79-KG5L].

the government proposed another TDM exception that would include transparency provisions and would allow for rights reservations and opt-outs, similar to the EU's Article 4.[136] This proposal has also faced significant backlash from creative industries since it would permit copying unless rightsholders explicitly opted out.[137] AI developers, on the other hand, are concerned that including the opt-out provision in the exemption could lead to biased training models and could have a "chilling effect" on AI research.[138]

A U.S. training law may face similar backlash to that in the U.K., and Congress should be prepared to address such criticism.

C. Japan

Japan has been a leader in AI law since 2009, when it passed Article 47(7), authorizing TDM by all users and for all purposes, whether commercial or non-commercial.[139] An amendment to the Copyright Act in 2018 sought to remove perceived copyright barriers to AI by (1) permitting all users to "analyse and understand copyrighted works for machine learning;" (2) permitting "electronic incidental copies of works, recognizing that this process is necessary to carry out machine learning activities but does not harm copyright owners;" and (3) allowing "the use of copyrighted works for data verification when conducting research."[140] These exceptions to copyright law are based on the idea that "the purpose of the exploitation is not to personally

136. *Copyright and Artificial Intelligence*, Gov. U.K.: Intell. Prop. Off. (Dec. 17, 2024), https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence [https://perma.cc/Z9R5-SCRA].

137. Letter from Keith Kupferschmid, Chief Exec. Off., Copyright Alliance to U.K. Intell. Prop. Off. (Feb. 25, 2025), https://copyrightalliance.org/wp-content/uploads/2025/02/AS-SUBMITTED-Copyright-Alliance_UKIPO-AI-Copyright-Consultation-Comments.pdf [https://perma.cc/W9FW-ZAPG]; *UKAI Publishes the 'Creative Industries & AI Report' and Rejects Proposals to Move to an Opt-Out Model for Copyright*, UKAI, https://ukai.co/creative-industries-report/ [https://perma.cc/LZ7X-BR83] (last visited Oct. 8, 2025); Paul Glynn, *Artists Release Silent Album in Protest Against AI Using Their Work*, BBC (Feb. 25, 2025), https://www.bbc.com/news/articles/cwyd3r62kp5o [https://perma.cc/H7RX-TXXY]; Robert Booth, *Paul McCartney Warns AI 'Could Take Over' as UK Debates Copyright Laws*, Guardian (Dec. 10, 2024, at 10:29 ET), https:// www.theguardian.com/technology/2024/dec/10/paul-mccartney-ai-warning-uk-debates-copyright-laws [https://perma.cc/K3NC-XC3P].

138. *See* Jennifer Cass & David Wagget, *The UK's New AI-Copyright Consultation: A Path to Clarity or Controversy?*, Dentons (Jan. 2, 2025), https://www.dentons.com/en/insights/articles/2025/january/2/the-uks-new-ai-copyright-consultation-a-path-to-clarity-or-controversy [https://perma.cc/G2V9-B4TE].

139. *Japan Amends Its Copyright Legislation to Meet Future Demands in AI and Big Data*, Eur. All. for Rsch. Excellence (Sep. 3, 2018), https://eare.eu/japan-amends-its-copyright-legislation-to-meet-future-demands-in-ai-and-big-data/ [https://perma.cc/ZN5X-286H].

140. *Id.*

'enjoy' . . . the thoughts or sentiments expressed in the work" and that the use does not "unreasonably prejudice the interests of the copyright owner."[141] Congress can rely on these justifications in passing an American AI law and responding to criticism like that in the U.K.

Notably, the analysis in Japan is primarily output-based. "Whether data reproduction for training generative AI models will infringe copyright will depend on the circumstances, in particular what is the intended output of the model and how similar outputs can become to the original source material through prompting."[142]

D. Singapore

Like the EU, Singapore provides a TDM exception. The Singapore exception extends to both non-commercial and commercial organizations, does not limit the purposes to which TDM is carried out, and cannot be overridden by contracts or opt-outs.[143] The Singapore Copyright Act 2021 permits copying for computational data analysis purposes, including:

> using a computer program to identify, extract, and analyse information or data from a work or recording, and using the work or recording as an example of a type of information or data to improve the function of a computer program in relation to that type of information or data.[144]

Gen AI falls into this category, so AI developers can reproduce copyrighted works for training purposes.[145] In Singapore, this permissive TDM exception has greatly benefited businesses, and it significantly lowers the cost of AI development, making AI models more accessible.[146] The exception provides legal clarity for developers and copyright owners alike, encourages growth of AI and technology, and supports "data-driven innovation."[147]

---

141. Rattzén, *supra* note 125, at 217.

142. *Id.*

143. Pin-Ping Oh, *Coming Up in Singapore: New Copyright Exception for Text and Data Mining*, Bird & Bird (Sep. 19, 2021), https://www.twobirds.com/en/insights/2021/singapore/coming-up-in-singapore-new-copyright-exception-for-text-and-data-mining [https://perma.cc/9WM5-DBA6].

144. Copyright Act 2021, §§ 190–91, 243–44 (Sing.).

145. Bryan Tan, *Text and Data Mining in Singapore*, ReedSmith (Feb. 5, 2024), https://www.reedsmith.com/articles/entertainment-and-media-guide-to-ai/text-and-data-mining-in-singapore/ [https://perma.cc/7PN8-U7PE].

146. Oh, *supra* note 143.

147. *A Balancing Act: Singapore's Computational Data Analysis Exception*, Infinity Legal LLC (Jan. 12, 2024, at 16:20 PT), https://infinitylegal.com.sg/copyright-law-singapore/ [https://perma.cc/3A3N-CH7E].

As seen in Singapore, TDM exceptions benefit society and support technological growth and innovation. Because of the issues with opt-out provisions, the United States would benefit from an approach similar to Singapore's—a TDM exception with no opt-outs for both commercial and non-commercial uses.

E. Brazil

In December 2024, Brazil's Senate voted on and approved an AI Bill which includes (1) disclosure obligations for AI companies to detail which copyrighted works were used for AI training; (2) establishment of a regulatory body to oversee licensing negotiations for use of copyrighted works; (3) obligations for AI developers to inform users of the use of content to develop AI systems, except where a copyright owner opts-out; and (4) an exception for the use of protected works to train AI by scientific and research organizations, museums, public archives, and libraries with a lawful access requirement and a requirement that the use not prejudice the normal exploitation of the original works.[148] The United States should look to Brazil's analysis, as it would benefit from disclosure obligations, regulatory oversight, and requirements that training uses not prejudice the original works.

F. International Harmonization (or Lack Thereof)

Perhaps the biggest issue facing AI developers is the lack of international harmonization when it comes to copyright law. As Mattias Rattzén observed, "for AI developers, location is indeed all you need" because "[i]n the absence of international harmonization, there is . . . a high likelihood that the same type of training activity would be considered copyright infringement in some countries but not in others."[149] Ultimately, even if the United States clarifies its AI training laws, developers could still be liable for infringement in other countries.[150]

148. Daniel Gullino, *Understanding the Rules for Artificial Intelligence in Brazil; the Bill Has Been Approved by the Senate*, O GLOBO (Dec. 11, 2024, at 01:15 PT), https://oglobo.globo.com/english/especial/understand-the-rules-for-artificial-intelligence-in-brazil-the-bill-has-been-approved-by-the-senate.ghtml [https://perma.cc/H8FQ-QURE].

149. Rattzén, *supra* note 125, at 175–76.

150. *See* Barry Scannel, *GEMA v. Open AI - Europe's Direction on AI Infringement?*, WILLIAM FRY (Nov. 12, 2025), https://www.williamfry.com/knowledge/gema-v-openai-europes-direction-on-ai-infringement/ [https://perma.cc/XA67-SD4C]; *see also International Intellectual Property Disputes: Cross-Border Challenges and Enforcement Strategies*, MICHAEL EDWARDS, https://michaeledwards.uk/international-intellectual-property-disputes-cross-border-challenges-and-enforcement-strategies [https://perma.cc/5C42-F6Q4] (last visited Feb. 7, 2026) ("Intellectual property rights are generally territorial, meaning that a court's ruling in one jurisdiction does not automatically apply in another. This means that businesses often have

This lack of international harmonization also creates potential for abuse by developers and bad actors. Until the countries of the world come together to harmonize AI training and copyright, copyright owners and the public may be harmed by AI developers finding loopholes and circumventing national copyright laws. On the other hand, AI developers face harm if they must conduct separate training activities for different jurisdictions, leading to greater expense, risk, and complexity. With no international standard for AI training, the focus may need to shift from an input analysis to an output analysis, [151] though copyright guidelines for AI output differ drastically by country.[152]

IV. The U.S. Solution

AI training is fair use, but how can the United States best protect both AI developers and copyright owners? Regulatory regimes abroad provide insight into what does and does not work. As this Comment has shown, TDM exceptions, licensing regimes, opt-out policies, AI bills, and copyright amendments have all been utilized to address issues posed by gen AI training. Based on analysis of these use cases, the best approach for the United States is to adopt a fair training legislative safe harbor. This Part will first address why licensing doesn't work before turning to a legislative solution.

A. Why Licensing Doesn't Work

As discussed above, many countries apply licensing regimes, and some rightsholders argue that AI developers should be required to obtain a license to use copyrighted data for gen AI training, even if such training is fair use.[153] Licenses better protect copyright owners and lay out what developers can and can't do with the copyrighted works, promoting responsible, transparent, and ethical usage of the

to litigate the same dispute in multiple jurisdictions, leading to increased costs and uncertainty. . . . For example, a trademark ruling in the United States may not be enforceable in China if Chinese courts determine that the US court's decision conflicts with domestic trademark law.").

151. *See* Rattzén, *supra* note 125, at 288.

152. *See Copyright Ownership of Generative AI Outputs Varies Around the World*, Cooley (Jan. 29, 2024), https://www.cooley.com/news/insight/2024/2024-01-29-copyright-ownership-of-generative-ai-outputs-varies-around-the-world [https://perma.cc/46CP-6LA5].

153. *See generally* Ass'n of Am. Publishers, Artificial Intelligence and Copyright (Docket No. 2023-6) Notice of Inquiry and Request for Comments (2023) (explaining that AI developers must obtain licenses from copyright holders before using their works for training, compensating creators fairly and ensuring continued investment in high-quality content).

licensed data.[154] Licenses ensure copyright holders are compensated for the use of their creative works and set parameters that limit abuse of their intellectual property.[155] Licensing regimes benefit developers, too, by securing developer access to data, limiting uncertainty, preventing litigation, offering a competitive edge over rivals, leveraging partnerships, and allowing developers to specify requirements for the format of the data.[156]

In reality, a licensing regime is not practical. ChatGPT was trained on the equivalent of 1.3 million books.[157] Even a nominal licensing fee would have cost OpenAI millions of dollars. Not only is that an extremely burdensome cost for start-up companies and small developers trying to break into the gen AI space, but it is also unimaginable that an AI developer—especially small actors—would be able to track down the copyright owners of each of those 1.3 million books to pay their licensing fee.[158] Even if the company could track down millions of copyright owners, the time it takes to negotiate licenses with each of them would hamper development and slow technological growth.

And what if a copyright owner decides *not* to license their work? This could lead to biased AI models. Limiting training data to open-source datasets and works that have been explicitly licensed can lead to smaller training sets that are prone to bias.[159] The more sources in a training set, the higher quality the data, and better data decreases the chances that the AI model will produce biased outcomes.[160] When copyright law is uncertain, it "privileges access to certain works over others, encouraging AI creators to use easily available, legally low-risk sources of data for teaching AI, even when those data are demonstrably

154. Nico Klingler, *AI Licenses: What You Should Know for Your Applications*, VISO.AI (Oct. 9, 2024), https://viso.ai/deep-learning/ai-licenses/ [https://perma.cc/G7EJ-NUWJ].

155. *Licensing Content to Train AI—An Emerging Frontier?*, BRISTOWS (Oct. 8, 2024), https://inquisitiveminds.bristows.com/post/102jl8j/licensing-content-to-train-ai-an-emerging-frontier [https://perma.cc/WZ3X-PV46].

156. *Id.*

157. Yoon & Borges, *supra* note 2.

158. *Id.*; Stefano Maffulli, *Why Datasets Built on Public Domain Might Not Be Enough for AI*, OPEN SOURCE INITIATIVE (May 7, 2024), https://opensource.org/blog/why-datasets-built-on-public-domain-might-not-be-enough-for-ai [https://perma.cc/69K4-R49Z]; Matthew Sag, *The False Hope of Content Licensing at Internet Scale*, PROMARKET (Nov. 19, 2025), https://www.promarket.org/2025/11/19/the-false-hope-of-content-licensing-at-internet-scale/ [https://perma.cc/6KSP-HBSL].

159. *See* Matthew G. Hanna et al., *Ethical and Bias Considerations in Artificial Intelligence/Machine Learning*, 38 MOD. PATHOLOGY 1, 8 (2025) (explaining bias mitigation is best done by actively seeking out and including diverse data sets that represent different backgrounds).

160. *See* Thomas Dethmann & Jannis Spiekermann, *Ethical Use of Training Data: Ensuring Fairness and Data Protection in AI*, LAMARR INST. FOR MACH. LEARNING & A.I. (July 3, 2024), https://lamarr-institute.org/blog/ai-training-data-bias/ [https://perma.cc/4TC9-USXQ].

biased."[161] Biased gen AI models risk reinforcing harmful stereotypes, unfairly disadvantaging certain groups, amplifying and perpetuating negative perceptions and treatment of minority groups, and producing flawed decision-making.[162] For example, in healthcare, if data of women and minority groups is inadequate or under-representative, an AI diagnosis system trained on such data may return lower accuracy results for a Latina woman versus a Caucasian man.[163] Similarly, if an AI-powered predictive policing tool is trained solely on historical arrest data or data from predominantly black neighborhoods, the model can reinforce patterns of racial discrimination, profiling, and disproportionate targeting of minorities.[164] Or if an autonomous vehicle is trained on inadequate photos of street signs, the car may misidentify a stop sign and cause an accident.

One suggested solution is compulsory licensing,[165] which would alleviate bias by requiring all copyright owners to license their works for AI training and prohibiting opt-outs. Though all data would be accessible, mandatory licensing would significantly increase development costs, which may lead to higher prices for AI models, reducing their affordability and limiting public access. It would also disproportionately disadvantage smaller actors and limit new developers' ability to compete with established tech giants.[166] Furthermore, mandatory licensing could still lead to biased AI models, as developers may attempt to avoid licensing fees by training their models on older, less accurate, or unverified data.[167]

A licensing requirement, while benefiting copyright owners, would ultimately slow innovation and would limit gen AI development to a small number of rich companies. AI developers are going to continue producing AI technology, and licensing requirements create an economic incentive to avoid licenses and train on biased data, ultimately

161. Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash. L. Rev. 579, 579 (2018).

162. *What Is AI Bias?*, SAP (Oct. 29, 2024), https://www.sap.com/resources/what-is-ai-bias [https://perma.cc/9YVF-SHS4].

163. James Holdsworth, *What Is AI Bias?*, IBM, https://www.ibm.com/think/topics/ai-bias [https://perma.cc/H95Z-WH3Z] (last visited Oct. 8, 2025).

164. *Id.*

165. World Intell. Prop. Org., Understanding Copyright and Related Rights 1, 15 (2016).

166. Quang, *supra* note 59, at 1408.

167. Thomas Margoni & Martin Kretschmer, *A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology*, 71 GRUR Int'l 685, 700 (2022).

hurting end-users and society. Broader access to training data will make AI algorithms fairer, safer, and more accurate.

B. The Solution—A Fair Training Safe Harbor Law

Gen AI training is fair use under U.S. law, but without a clear mandate from Congress, uncertainty and litigation will persist. A legislative safe harbor would provide both AI developers and copyright owners with legal certainty. A safe harbor is a regulatory framework that ensures certain conduct is lawful so long as specific criteria are met.[168]

Congress should combine the idea of a safe harbor, a TDM exception, and the fair use defense to pass a law codifying a fair training right for gen AI, legislatively permitting the use of copyrighted works to train gen AI models. A safe harbor for AI training would protect AI developers from legal liability, provided they abide by certain standards. It would democratize AI by making it easier for smaller creators to enter the market and by removing the deterrent of fear of litigation.[169] Clear legislation permitting training on copyrighted data would mitigate bias in algorithms and make safer, fairer models more accessible to the public, promoting innovation and technological advancement.

The goal of this safe harbor is to remove ambiguity, develop common standards in AI training, and help developers maintain a competitive edge.[170] To be effective, the law should clearly and broadly define data mining and training, taking guidance from the EU's Article 2 and Singapore's Copyright Act. It should explicitly allow reproductions, derivations, and sharing for training purposes.[171] Following Singapore's approach, the safe harbor should allow for both non-commercial and commercial uses equally, but to protect copyright owners, it should be limited to functional, non-expressive uses of copyrighted data.[172] Importantly, this law would apply only to training *input* and would not preclude lawsuits for infringement in AI *output*. Like Singapore's TDM

168. *Safe Harbor*, Legal Info. Inst. (June 2024), https://www.law.cornell.edu/wex/safe_harbor [https://perma.cc/N5RN-ZZGL]. A well-known example of a copyright-related safe harbor is the 1998 Digital Millennium Copyright Act, which protects online service providers from liability for transitory communication, system caching, storage of information by end users, and information location tools so long as several conditions are met. 17 U.S.C. § 512.

169. Quang, *supra* note 59, at 1408.

170. Hunter Walk, *Instead of Asking AI Companies to 'SLOW DOWN' We Should Encourage Them to Move Even Faster*, LinkedIn (Apr. 6, 2023), https://www.linkedin.com/pulse/instead-asking-ai-companies-slow-down-we-should-encourage-hunter-walk/ [https://perma.cc/945M-5JFL].

171. Quang, *supra* note 59, at 1433.

172. *Id.*

exception and the U.K.'s first proposal, there should be no provision for opt-out or rights reservation, as such provisions contribute to bias and hinder innovation. An effective safe harbor would also establish a regulatory body and should provide for independent evaluation of gen AI models to ensure compliance.[173] Finally, like both the EU's and Brazil's AI Acts, the U.S. AI law should include transparency, disclosure, and reporting requirements.

In sum, to remove ambiguity in training and promote innovation and societal benefits, Congress should pass fair training legislation codifying the right to use copyrighted data to train gen AI.

Conclusion

The United States needs to update and clarify its approach to generative AI training. Gen AI training meets the fair use test, but until Congress passes a law to explicitly address fair training, uncertainty and legal battles will persist. Taking lessons from other countries, Congress should pass fair training legislation that codifies gen AI fair use and ensures that AI developers will be able to effectively and ethically train their AI models without fear of legal action by copyright owners. The law should not include an opt-out provision, but it should include transparency and disclosure requirements. This will protect AI developers in the United States and promote safe, fair, accurate, and accessible gen AI models for the use and betterment of American society. However, until there is international cooperation and agreement on fair training, developers will have to be wary of where they train and publish their models, as even though training is fair use in the United States, it may be considered infringement in other countries.

The continued growth of gen AI is inevitable. To ensure that society is offered the best, safest, and most accurate AI products, the international community must come together and agree that using copyrighted data to train AI is legal and should, in fact, be encouraged.

173. Shayne Longpre et al., *A Safe Harbor for AI Evaluation and Red Teaming*, KNIGHT FIRST AMEND. INST. AT COLUM. UNIV.: BLOG (Mar. 5, 2024), https://knightcolumbia.org/blog/a-safe-harbor-for-ai-evaluation-and-red-teaming [https://perma.cc/5WGY-WTV9].