PennState Dickinson Law

DICKINSON LAW REVIEW
PUBLISHED SINCE 1897

Volume 128 | Issue 1

Fall 2023

Training is Everything: Artificial Intelligence, Copyright, and "Fair Training"

Andrew W. Torrance

Bill Tomlinson

Essays

Training is Everything: Artificial Intelligence, Copyright, and "Fair Training"

Andrew W. Torrance & Bill Tomlinson*

Abstract

In this Essay, we analyze the arguments in favor of, and against, viewing the use of copyrighted works in training sets for AI as fair use. We call this form of fair use "fair training." We identify both strong and spurious arguments on both sides of this debate. In addition, we attempt to take a broader perspective, weighing the societal costs (e.g., replacement of certain forms of human employment) and benefits (e.g., the possibility of novel AI-based approaches to global issues such as environmental disruption) of

* Dr. Andrew W. Torrance, Ph.D. is a Paul E. Wilson Distinguished Professor of Law at the University of Kansas School of Law and a Visiting Scientist at the Massachusetts Institute of Technology Sloan School of Management. Dr. Bill Tomlinson, Ph.D. is a Professor of Informatics at the University of California, Irvine's Donald Bren School of Information and Computer Sciences and an Adjunct Professor at Te Herenga Waka - Victoria University of Wellington. The authors would like to thank Amanda McElfresh and Lauren Stahl for their expert research and editing. We would also like to acknowledge the useful comments and suggestions on early drafts of this Essay generously provided by Mark A. Lemley and Matthew Sag. This material is based upon work supported by the National Science Foundation under Grant No. DUE-2121572.

allowing AI to make easy use of copyrighted works as training sets to facilitate the development, improvement, adoption, and diffusion of AI. Finally, we suggest that the debate over AI and copyrighted works may be a tempest in a teapot when placed in the wider context of massive societal challenges such as poverty, inequality, climate change, and loss of biodiversity, to which AI may be part of the solution.

TABLE OF CONTENTS

INTRODUCTION: AI AND ITS LEAP INTO THE PUBLIC CONSCIOUSNESS

There once was a time when the idea of talking to a machine seemed like something straight out of science fiction. Yet, in just a matter of a few short years, AI has become almost commonplace.[1] One of the pioneers of AI, British mathematician Alan Turing, once said, "we can only see a short distance ahead, but we can see plenty

1. *See* George Siemens, *Not Everything We Call an AI is Actually Artificial Intelligence. Here's What to Know*, THE CONVERSATION (Dec. 25, 2022), https://tinyurl.com/bddurzjz [https://perma.cc/8Z8G-GKWX] ("Late last month, AI, in the form of ChatGPT, broke free from the sci-fi speculations and research labs and onto the desktops and phones of the general public.").

there that needs to be done."[2] In the last year, AI has finally achieved its first true milestone of democratization, and is in the midst of changing and disrupting the way we live and work.[3]

AI has been a rapidly growing field in recent years, and finally gained significant usage by the general public in 2022.[4] It did so not because of a popular Hollywood movie, like *The Terminator*, or the extravagant claim of a company or pundit. Rather, it earned this newfound attention from the public due to its sudden usefulness and practicality.[5] The democratization of AI technology is largely credited to companies such as OpenAI, Stability AI, and Discord, who have made it easier for individuals without formal computer science training to use and benefit from AI applications.[6] For instance, in quick succession, OpenAI, a software company based in San Francisco, released a graphics generator (DALL-E2), a text generator (GPT3.5), and then chatbots (i.e., ChatGPT-3.5, followed by ChatGPT-4) capable of carrying on compelling conversations with humans.[7] OpenAI, in particular, has released a range of AI tools including a graphics generator, a text generator, and a chatbot that have captured the public imagination.[8] These democratized forms of AI have paved the way for AI to be adopted in a practical mode by a wider range of people, marking a key milestone in the development of AI technology.[9]

However, there is a *sine qua non* lurking behind these democratized sources of AI that has triggered a substantial legal response.[10]

---

2. Alan M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433, 460 (1950).

3. *See Generative AI Poised to Change the Way We Live According to Experts*, VA. TECH NEWS (Jan. 31, 2023), https://tinyurl.com/576x2w93 [https://perma.cc/642Y-CC5P].

4. *See* Siemens, *supra* note 1.

5. *See id.*

6. *See* Harry Guinness, *A Guide to the Internet's Favorite Generative AIs*, POPULAR SCI. (Jan. 11 2023, 6:00 PM), https://tinyurl.com/587cfe68 [https://perma.cc/A5U9-N44Q] (discussing the various AIs available to the public including Stability and Discord).

7. Johan Moreno, *OpenAI Positioned Itself as the AI Leader in 2022. But Could Google Supersede It In '23?*, FORBES (Dec. 29, 2022, 4:53 PM), https://tinyurl.com/5ckdb4w7 [https://perma.cc/QLN7-RDAJ].

8. *See* Ryan Browne, *All You Need to Know About ChatGPT, the A.I. Chatbot That's Got the World Talking and Tech Giants Clashing*, CNBC (Apr. 17, 2023, 2:29 AM), https://tinyurl.com/mvksxdhy [https://perma.cc/BM96-PVJ6] (discussing OpenAI's different AI tools).

9. *See* Siemens, *supra* note 1.

10. *See* Chloe Xiang, *Artists Are Suing Over Stable Diffusion Stealing Their Work for AI Art*, VICE (Jan. 17, 2023, 11:31 AM), https://tinyurl.com/4tvxym57 [https://perma.cc/K3K6-PH4L] (discussing the recently filed "class action lawsuit against Stability AI, DeviantArt, and Midjourney, alleging that the text-to-image AI tools have infringed the rights of thousands of artists and other creatives 'under the guise of artificial intelligence'"); *see* Blake Brittain, *Getty Images Lawsuit Says Stability AI*

To learn how to behave, the current revolutionary generation of AIs must be trained on vast quantities of published images, written works, sounds, or other forms of data, many of which fall within the core subject matter of copyright law.[11] To some, the use of copyrighted works as training sets for AI is merely a transitory and non-consumptive use that does not materially interfere with owners' content or copyrights protecting it.[12] Companies that use such content to train their AI engines often believe such usage should be considered "fair use" under U.S. law (sometimes known as "fair dealing" in other countries).[13] By contrast, many copyright owners, as well as their supporters, consider the incorporation of copyrighted works into training sets for AI to constitute misappropriation of owners' intellectual property, and, thus, decidedly not fair use under the law.[14] The future trajectory of AI and its applications hinges on these issues, and, given the transformative nature of AI and its potential to impact society in myriad ways, these issues have become increasingly important, relevant, and needful of resolution.

The purpose of this Essay is to analyze the arguments for and against considering the unlicensed use of copyrighted works in training sets for AI as fair use, fair dealing, or "fair training."[15] The Essay will explore the implications of these arguments for copyright law and the future of AI technology. By examining this issue in detail, the Essay aims to contribute to a greater understanding of the complex relationship between AI and copyright law.

The recent rapid advance of AI marks a notable inflection point in human history. Marco Iansati and Karim Lakhani describe the singularity of this time, "[j]ust as in the Industrial Revolution, the age of AI is transforming the economy. However, the speed and breadth of the impact appear to be many times as great. It will not take a

Misused Photos to Train AI, REUTERS (Feb. 6, 2023, 11:32 AM), https://tinyurl.com/mwd52hby [https://perma.cc/X7LR-J4WW].

11. *See* 17 U.S.C. § 102 (detailing works protected under copyright, including "literary works," "musical works," "dramatic works," "pantomimes and choreographic works," "pictorial, graphic, and sculptural works," "motion pictures," "sound recordings," and "architectural works").

12. *See* James Vincent, *The Scary Truth About AI Copyright is Nobody Knows What Will Happen Next*, VERGE (Nov. 15, 2022, 10:00 AM) https://tinyurl.com/56xtcusr [https://perma.cc/34PE-2MQB] (discussing the arguments in favor of the fair use defense for AI).

13. *See* Taysir Awad, *Universalizing Copyright Fair Use: To Copy, or Not to Copy?*, 30 J. INTELL. PROP. L. 1, 3–6 (2022) (discussing the concepts of fair use and fair dealing and the countries that use each of these concepts).

14. *See generally* Vincent, *supra* note 12 (discussing the potential copyright implications of AI).

15. We independently conceived of the phrase "fair training" ourselves. However, we do not claim we are the first to use this phrase. In fact, we would be surprised if others had not employed it previously.

hundred years for digital transformation to pervade every sector of the global economy."[16]

As further evidence of the growing role of AI in society, we also note that we wrote this Essay in collaboration with ChatGPT (Jan. 9th, 2023 version). We did so, in part, to investigate how scholars and AI could collaborate to produce scholarship.[17]

Writing this Essay was, in part, an experiment in a new form of scholarly production. As a consequence, some published work by our colleagues may have been inadvertently missed by the process we describe above. We beg their indulgence for any omissions resulting from our experimental writing method. Nevertheless, we have tried to incorporate relevant references wherever we could within the parameters of our experiment. One strategy to accomplish this has been to post an early draft of our Essay on SSRN for anyone to review. Several colleagues, having read the draft, did kindly suggest references to add; we have, indeed, added these.

As AI advances and becomes a more integral part of our daily lives, the need for a comprehensive examination of its legal and ethical implications becomes increasingly pressing. Given the pivotal role played by training sets, it is imperative for individuals, organizations, and policymakers to closely consider the relationship between AI and copyright law and collaborate towards a solution that benefits society at large. As Mark Twain once wrote: "Training is everything."[18] This Essay is intended to serve as a catalyst for this much-needed discourse and calls for a proactive approach to balancing the advancements of AI, especially in the arena of training, with the protections of copyright law.

---

16. *See* MARCO IANSATI & KARIM R. LAKHANI, COMPETING IN THE AGE OF AI: STRATEGY AND LEADERSHIP WHEN ALGORITHMS AND NETWORKS RUN THE WORLD 206 (2020).

17. While that system contributed substantially to the text, we are omitting it from the author list in line with the recommendation of Springer Nature, a major scientific publisher. *See Tools Such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use*, NATURE (Jan. 24, 2023), https://tinyurl.com/3944sty7 [https://perma.cc/TTN5-CFGJ]. In line with best practices in producing scholarship with AI, we have run this Essay through the TurnItIn plagiarism detection software to ensure that ChatGPT did not inadvertently commit plagiarism or violate copyright. *See* Bill Tomlinson, Andrew W. Torrance & Rebecca W. Black, *ChatGPT and Works Scholarly: Best Practices and Legal Pitfalls in Writing with AI*, 76 SMU L. REV. F. 108, 124 (2023). As of September 7, 2023, a draft of this Essay had no plagiarism through TurnItIn.

18. MARK TWAIN, THE TRAGEDY OF PUDD'NHEAD WILSON 67 (1894) ("Training is everything. The peach was once a bitter almond; cauliflower is nothing but cabbage with a college education.—*Pudd'nhead Wilson's Calendar.*").

I. PRIOR SCHOLARLY VIEWS ON TRAINING DATA AND COPYRIGHT

A wealth of previous scholarship has been published about training data for AI and the law. We survey some of it in this Part.

Long before the recent flurry of interest in AI and training data, Matthew Sag identified and analyzed the risks involved in the use of copyrighted works of authorship by "copy-reliant technologies."[19] He observed that:

> Copy-reliant technologies, such as Internet search engines and plagiarism detection software, do not read, understand, or enjoy copyrighted works, nor do they deliver these works directly to the public. They do, however, necessarily copy them in order to process them as grist for the mill, raw materials that feed various algorithms and indices.[20]

He argued that, in general, such use of copyrighted work should be viewed as "nonexpressive," and, consequently, should not qualify as copyright infringement.[21] Sag extended his analysis to show that a similar analysis applied in the cases of text data-mining and library digitization.[22] In addition, Sag demonstrated how the same principle of non-expressive use for copy-reliant technologies applies in the case of text-mining for training of machine-learning AI systems.[23]

Mark Lemley and Bryan Casey have made a strong case that the use of copyrighted works of authorship should generally be allowed to be used in training sets for AI.[24] As they have argued:

> [Machine learning] systems should generally be able to use databases for training, whether or not the contents of that database are copyrighted. There are good policy reasons to do so. First, we need to encourage people to compile new databases and to open them up for public scrutiny or innovation. Broad access to training sets will further these objectives. . . . And because training sets are likely to contain millions of different works with thousands of different owners, there is no plausible option simply to license

19. Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. UNIV. L. REV. 1607, 1608 (2009).

20. *Id.*

21. *Id.*

22. *See generally* Matthew Sag, *Orphan Works as Grist for the Data Mill*, 27 BERKLEY TECH. L.J. 1503 (2012); Matthew Jockers, Matthew Sag & Jason Schultz, *Don't Let Copyright Block Data Mining*, 490 NATURE 29 (2012).

23. *See generally* Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, J. COPYRIGHT SOC'Y. U.S.A. 66 (2019) (expressly tying the concept of non-expressive use to machine learning and AI).

24. *See generally* Mark A. Lemley & Bran Casey, *Fair Learning*, 99 TEX. L REV. 743 (2021).

> all of the underlying photographs, videos, audio files, or texts for the new use. So allowing a copyright claim is tantamount to saying, not that copyright owners will get paid, but that the use won't be permitted at all, at least without legislative intervention. While we share some of the concerns about the uses to which [machine learning] systems may be put, copyright is not the right tool to regulate those abuses.[25]

Lemley and Casey also use the phrase "fair learning" to refer to how fair use may allow the incorporation of copyrighted works of authorship into training datasets used to train AI systems.[26] Henderson and others suggest that existing foundation models usually use training data that contains copyrighted works of authorship.[27] They warn that such use can trigger legal liability:

> In the United States and several other countries, copyrighted content may be used to build foundation models without incurring liability due to the fair use doctrine. However, there is a caveat: If the model produces output that is similar to copyrighted data, particularly in scenarios that affect the market of that data, fair use may no longer apply to the output of the model.[28]

Benjamin Sobel has raised the worry that the fair use doctrine of copyright law may be ill-suited to governing the use of copyrighted works of authorship in training datasets.[29] He worries this could act as a drag, or even thwart, progress in AI if copyright law ended up denying training sets the data they need to succeed.[30] On the other side of the risk ledger, James Grimmelmann has raised the alarm that a permissive interpretation of copyright law and the fair use defense, if it allowed copyrighted works of authorship to be used to train AIs, could give rise to unintended consequences by empowering superintelligent AIs to gain too much power over humanity, possibly endangering it.[31]

Beyond the United States' legal context, several authors have reviewed the issues raised under European laws by the incorporation of copyrighted works of authorship into training datasets for training AI systems. Eleonora Rosati has examined whether European

25. *Id.* at 748–49 (footnotes omitted).

26. *Id.* at 750.

27. Peter Henderson et al., *Foundation Models and Fair Use*, 23 J. MACH. LEARNING RSCH. (forthcoming 2023).

28. *Id.*

29. *See* Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 45–46 (2017).

30. *Id.* at 80.

31. James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 677–78 (2016).

laws, including copyright laws, help or hinder the development of AI systems.[32] She opines that despite two new exceptions to the text and data mining ("TDM") restrictions imposed by the Digital Single Market ("DSM") Directive, "copyright restrictions might continue affecting and restricting significantly the possibility of undertaking TDM activities in Europe."[33]

Rossana Ducato and Alain Strowel also review the DSM Directive and its potential influence on the evolution of AI systems, offering a critique of the Directive and suggestions on how it might be improved.[34] In addition, they have proposed a legal right of "machine legibility" to ensure that text-mining and data-mining are enabled in the contexts of smart disclosure systems and training datasets for AI.[35] In her Master's thesis, Gabriella Svensson provides an appraisal of how European Union (EU) copyright laws constrain text-mining and data-mining.[36]

Outside of the United States and Europe, Martin Senftleben has surveyed a variety of national copyright legal regimes with particular focus on TDM and compliance with the Berne Convention for the Protection of Literary and Artistic Works ("Berne Convention").[37] He argues that:

> TDM does not concern a traditional category of use that could have been contemplated at the diplomatic conferences leading to the current texts of the Berne Convention, the TRIPS Agreement and the WIPO Copyright Treaty. It is an automated, analytical type of use that does not affect the expressive core of literary and artistic works. Arguably, TDM constitutes a new category of copying that falls outside the scope of international copyright harmonization altogether.[38]

---

32. *See* Eleonora Rosati, *Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity*, 27 ASIA PAC. L. REV. 198, 198–99 (2019).

33. *Id.*

34. Rossana Ducato & Alain M. Strowel, *Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out*, 43 EUR. INTELL. PROP. REV. 322, 322–24 (2021).

35. Rossana Ducato & Alain M. Strowel, *Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility'*, 50 INT'L REV. INTELL. PROP. & COMPETITION L. 649, 652 (2019).

36. GABRIELLA SVENSSON, TEXT AND DATA MINING IN EU COPYRIGHT LAW 3 (2020), https://tinyurl.com/25wuxvt3 [https://perma.cc/DU2R-YSPZ].

37. Berne Convention for the Protection of Literary and Artistic Works, July 14, 1967, 828 U.N.T.S. 221; *see* Martin Senftleben, *Compliance of National TDM Rules with International Copyright Law: An Overrated Nonissue?*, 53 INT'L REV. INTELL. PROP. & COMPETITION L. 1477, 1478–79 (2022).

38. Senftleben, *supra* note 37, at 1477.

He suggests that some national copyright systems go beyond the requirements of the Berne Convention.[39]

Amanda Levendowski has described how concerns about potential copyright infringement liability could funnel a skewed sample of disproportionately low-risk works of authorship into training data, resulting in potential for heightened bias in any AI trained on such skewed data.[40] Nevertheless, she identifies the fair use doctrine of copyright law as offering some mitigation for such bias.[41] Addressing similar worries, Thomas Margoni and Martin Kretschmer assess the effects the DSM Directive may have on the development of AI systems.[42] One of their sobering conclusions is that:

> [T]he provisions of the []DSM Directive paradoxically favour the development of biased AI systems due to price and accessibility conditions for training data that offer the wrong incentives. To avoid licensing, it may be economically attractive for developers to train their algorithms on older, less accurate, biased data, or import AI models already trained on unverifiable data.[43]

Christian Handke, Lucie Guibault, and Joan-Josep Vallbé have addressed broad challenges scholarly researchers face in navigating TDM for purposes of research.[44] Rachael Samberg and Cody Hennesy have written a thoughtful guide to "computational text analysis" that is aimed at scholars who might use this approach.[45] Among other topics, they address legal pitfalls arising from copyright law, contract law, and database law that scholars should be careful of.[46]

---

39. *See id.* ("[C]ountries in the EU rely on a more restrictive regulation that is based on specific copyright exceptions.").

40. *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 579–82 (2018).

41. *See id.* at 619–20.

42. *See* 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, 685–86 (2022).

43. *Id.* at 700.

44. *See generally* Christian Handke, Lucie Guibault & Joan-Josep Vallbé, *Copyright's Impact on Data Mining in Academic Research*, 42 MANAGERIAL & DECISION ECON. 1999 (2021).

45. *See generally* Rachael Gayza Samberg & Cody Hennesy, *Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis*, *in* COPYRIGHT CONVERSATIONS: RIGHTS LITERACY IN A DIGITAL WORLD 289 (Sara R. Benson ed., 2019).

46. *See id.* at 289–308.

II. THE NEED FOR TRAINING DATA IN AI

AI algorithms rely on large amounts of data to "learn" how to perform tasks and make decisions.[47] This data, referred to as "training data," is used to train AI algorithms to recognize patterns and make predictions based on those patterns.[48] The accuracy of the AI algorithm is directly dependent on the quality and quantity of the training data that it is exposed to.[49]

For example, a machine learning algorithm trained to recognize images of cats must be exposed to a large number of images of cats to learn what a cat looks like and how to distinguish it from other objects. In a similar manner, a large language model like OpenAI's GPT-3 must be exposed to large quantities of written text to learn the patterns of language and how to generate coherent and contextually appropriate responses to user inputs.[50]

A problem with training data is that it often contains copyrighted works, such as images, written works, and sounds.[51] This raises the question of whether the unlicensed use of copyrighted works in training sets for AI constitutes a fair use, or fair dealing, under the law, or if it constitutes misappropriation of intellectual property.[52] Several scholars have already contributed substantially to our understanding of copyright and training data.

Given the critical role that training data plays in AI development, it is important to understand the legal implications of using copyrighted works in AI training sets. The answer to this question has far-reaching implications for AI development and the future of AI technology. In the following Parts, we will examine the arguments for and against viewing the use of copyrighted works in training sets for AI as fair use, fair dealing, or "fair training."

---

47. *See* Amal Joby, *What is Training Data? How It's Used in Machine Learning*, LEARN G2 (July 30, 2021), https://tinyurl.com/mss9hf25 [https://perma.cc/GPB2-PGEA] (discussing the building blocks of machine learning, training data, and artificial intelligence).

48. *Id.*

49. *Id.*

50. *See* Will Douglas Heaven, *ChatGPT is Everywhere. Here's Where it Came From*, MIT TECH. REV. (Feb. 8, 2023), https://tinyurl.com/e6yxjcfe [https://perma.cc/XK2C-GCGL] (describing how GPT-3 functions and its capabilities).

51. Vincent, *supra* note 12 ("Most systems are trained on huge amounts of content scraped from the web; be that text, code, or imagery.").

52. *See id.*

III. THE DEMOCRATIZATION OF AI THROUGH OPENAI AND OTHER COMPANIES

AI was once a field that was limited to computer scientists and researchers. However, this changed dramatically in 2022 with the release of AI tools that were easy enough for people without formal computer science training to use.[53] Originally set up as a bulwark against unethical applications of AI, the company OpenAI was at the forefront of this democratization of AI, releasing graphics generators like DALL-E2, text generators like GPT-3.5, and chatbots like ChatGPT, which could carry on fluent, engaging, and sometimes even compelling conversations with humans.[54]

OpenAI's contributions to the democratization of AI were accompanied by those of several other companies, such as Stability AI and Discord, which made AI tools even more accessible to the public.[55] With these tools, almost anyone could create and experiment with AI; from artists and musicians to journalists and small businesses, AI entered a new phase of popular accessibility.

The democratization of AI has had a profound impact on society by creating an environment in which AI is used in a more practical, everyday mode by orders of magnitude more people than ever before. There is now a new class of AI users who are not computer scientists but rely on AI to perform a range of tasks limited largely by human imagination.[56]

Democratization of AI has also created a new set of legal challenges, as AI algorithms must be trained on vast quantities of published images, written works, and sounds, all of which are within the core subject matter of copyright.[57] The legal implications of using copyrighted works in AI training sets must be understood and addressed to ensure the continued growth and development of AI technology.

---

53. *See* Siemens, *supra* note 1.

54. *See* Arianna Johnson, *Here's What To Know About Open AI's ChatGPT—What It's Disrupting and How To Use It*, FORBES (Dec. 12, 2022, 12:23 PM), https://tinyurl.com/5n8fyfz2 [https://perma.cc/ETJ5-4V3B].

55. *See* Guinness, *supra* note 6.

56. *See, e.g.*, Megan Cerullo, *Here's How Professionals in 3 Different Fields Are Using ChatGPT for Work*, CBS NEWS (Feb. 9, 2023, 5:00 AM), https://tinyurl.com/28ncfawe [https://perma.cc/C882-YUWH] (detailing how professionals in real estate, finance, and the medical field use ChatGPT); *see also* Nick Bilton, *ChatGPT Made Me Question What It Means To Be a Creative Human*, VANITY FAIR (Dec. 9, 2022), https://tinyurl.com/2m3zufj8 [https://perma.cc/5B5X-RDCB] (describing how ChatGPT is also used to produce various forms of creative content including jokes, haikus, and screenplays).

57. *See* Xiang, *supra* note 10; Brittain, *supra* note 10.

IV. THE RELATIONSHIP BETWEEN AI AND COPYRIGHT LAW

The use of copyrighted works as training sets for AI algorithms is a new and rapidly evolving issue that has yet to be fully addressed by copyright law.[58] On one hand, some argue that the use of copyrighted works in AI training sets is a transitory and non-consumptive use that does not materially interfere with owners' copyrights, and therefore should be considered a particular form of fair use under U.S. law, or fair dealing in other countries.[59] This is the concept of "fair training."

On the other hand, others argue that the incorporation of copyrighted works into training sets for AI constitutes an unauthorized misappropriation of owners' intellectual property, and is decidedly not fair use, fair dealing, or "fair training" under the law.[60] This disagreement has led to conflicting interpretations of copyright law, and a lack of clarity regarding the legal status of AI training sets.[61]

It is important to consider both the legal and ethical implications of using copyrighted works in AI training sets. This includes considering the impact on copyright owners, as well as the benefits to society, and the advancement of AI technology.

To address these questions, we must examine the current state of copyright law, as well as consider possible solutions that may reconcile the conflicting interests of copyright owners and AI developers. This Part will provide an overview of the relationship between AI and copyright law, including the legal implications of using copyrighted works in AI training sets, and the ongoing debate over the fairness of such uses.

In this Essay, we propose the concept of "fair training" for AI. We argue that the use of copyrighted works as training data for AI should be considered a lawful, non-consumptive, and transformative use.[62] To understand why, it's helpful to consider how humans interact with and learn from copyrighted content.

Just like AI algorithms, humans consume, process, and store information contained within copyrighted works, such as books,

---

58. *See* Vincent, *supra* note 12.

59. *Id.* ("The justification used by AI researchers, startups, and multibillion-dollar tech companies alike is that using these images is covered (in the US, at least) by fair use doctrine, which aims to encourage the use of copyright-protected work to promote freedom of expression.").

60. *See* Jessica L. Gillotte, Note, *Copyright Infringement in AI-Generated Artworks*, 53 U.C. DAVIS L. REV. 2655, 2679–91 (2020) (discussing the circuit courts that do find infringement when AI uses copyrighted works).

61. *See id.*

62. For an in-depth discussion on fair use factors, see generally Neil Weinstock Netanel, *Making Sense of Fair Use*, 15 LEWIS & CLARK L. REV. 715 (2011).

music, and movies. This consumption and learning process may not infringe on the authors' copyright in some cases, because humans have the ability to engage in transformative uses of copyrighted content.[63] For example, if a human were reading a book, she might take notes, and then summarize the book's contents, which may rise to the level of a transformative use that does not infringe on the copyright of the author.[64] Similarly, a DJ sampling tiny bits of copyrighted songs at a dance party to generate a fun musical pastiche may sometimes amount to a transformative use that does not infringe the original author's copyright.[65]

We argue that training AI algorithms should also be considered a transformative use and therefore, should be considered "fair training." The use of copyrighted works as training data is crucial for the development of AI, as it allows the algorithms to learn, understand, and improve upon the information they are processing. When the AI algorithm uses this training set, they are transforming the data into new and unique forms of knowledge and not producing copies of the original works. Because the AI algorithms are transforming the original work, this use should not be considered a violation of the creators' copyright. Instead, such uses by AI algorithms should be protected under a "fair training" exception.[66]

In this light, "fair training" becomes a necessary concept for the democratization and continued development of AI. The "fair training" exception will balance the rights of copyright owners with the AI's ability to learn and grow.

V. ARGUMENTS IN FAVOR OF "FAIR TRAINING"

"Fair training" is necessary for the continued development of AI and for society to fully realize the benefits that come from AI. AI learns in a comparable way to how humans learn, by exposure to a variety of works without necessarily violating copyright. Exposure to these sources is necessary for AI to develop the ability to recognize and understand the nuances of language, images, and sounds. This exposure ensures that AI can learn and become more sophisticated. Furthermore, "fair training" does not consume the data upon which

63. *See* David E. Shipley, *A Transformative Use Taxonomy: Making Sense of the Transformative Use Standard*, 63 WAYNE L. REV. 267, 279–311 (2018) (defining transformative use and discussing various types of transformative use).

64. *See id.*

65. *See id.*

66. *Id.* at 280. ("The use of a copyrighted work need not alter or augment the work to be transformative in nature. Rather, it can be transformative in function or purpose without altering or actually adding to the original work.") (quoting A.V. v. iParadigms, LLC., 562 F.3d 630, 639 (9th Cir. 2009)).

it trains, but leaves this data unaltered and intact once its training is completed.

"Fair training" protects the AI's ability to learn and develop because the use of copyrighted works is crucial to training AI. The use of copyrighted works is a necessary step towards creating a more advanced AI that will benefit society. Once trained, AI transforms the copyrighted works and generates new and original works based on what it has learned, rather than copying or replicating existing works. Because the new and original works are not copies of the original, there is no commercial exploitation.

Additionally, the development of AI-powered tools and applications will lead to the creation of new jobs, the growth of existing industries, and innovative technologies.[67] These benefits will drive economic growth and benefit society in numerous ways. Moreover, the widespread adoption of AI will lead to improved efficiency and accuracy in various fields, such as healthcare, finance, and education.[68]

An underappreciated consequence of excluding source texts from training data sets is the exacerbation of bias. This could occur if certain source texts were not allowed, due to the copyright law, to be included in training under an interpretation of the fair use defense—in this case, fair training—that allowed authors or owners to forbid their works from being part of training data sets. If maximalist copyright interpretations were to prevail, training data would consist only of the subset of works allowed by their owners to be included. The more inclusive the training set, the less vulnerable the resulting AI would be to bias. On the other hand, the less inclusive the training set is, the more likely the AI trained on it would exhibit bias. Consequently, there is a strong public policy interest in ensuring that the fair use defense is not used to exclude works from training data sets. Fair use means fair AI.[69]

In summary, the arguments in favor of "fair training" are centered around the idea that using copyrighted works as training sets for AI is a non-consumptive, necessary, and beneficial use that promotes the advancement of AI and the growth of society. As such, it should not be considered copyright infringement.

---

67. *See, e.g.*, Adi Gaskell, *AI Creates Job Disruption Not Job Destruction*, FORBES (Jan. 18, 2022, 8:45 AM), https://tinyurl.com/28hc2zje [https://perma.cc/G8W4-DR8B] (discussing AI's influence in the workplace).

68. *See* Q.ai, *Artificial Intelligence's New Role in Medicine, Finance and Other Industries—How Computer Learning is Changing Every Corner of the Market*, FORBES (Feb. 2, 2023, 12:49 PM), https://tinyurl.com/en5awc2a [https://perma.cc/P42F-8H32] (discussing AI's impact in healthcare, finance, and education).

69. *See generally* Levendowski, *supra* note 40.

VI. ARGUMENTS AGAINST "FAIR TRAINING"

Critics of "fair training" argue that using copyrighted works as training sets for AI does materially interfere with owners' copyrights and is not a transformative or non-consumptive use.[70] They view the incorporation of copyrighted works into training sets for AI as a misappropriation of owners' intellectual property and not a fair use, fair dealing, or "fair training" under the law.

One argument is that AI algorithms are designed to mimic human thought processes, so the use of copyrighted works in training sets may result in AI that creates similar or identical works, which would infringe on the original creators' rights.

Another argument is that the use of copyrighted works in AI training sets creates derivative works, which are protected under copyright law. This would mean that the training of AI algorithms would require permission from the copyright holders, even if the AI-generated outputs are not identical to the original works.

Additionally, critics might argue that the use of copyrighted works in AI training sets could lead to market harm, as AI-generated outputs could compete with or replace the original works. The harm to the copyright holders' market can be justified as fair use, fair dealing, or "fair training."

In conclusion, those who argue against "fair training" believe that the use of copyrighted works in AI training sets is an infringing use that holds the potential to harm copyright holders, and, as with derivative works, cannot be justified as fair use, fair dealing, or "fair training" under the law.

VII. INTERNATIONAL APPROACHES TO AI AND COPYRIGHT

In this Part, we will examine the approach to AI and copyright law in various international jurisdictions. Different countries have different legal systems and cultural attitudes towards AI and copyright, which have influenced their approach to the issue. Some countries may adopt a more permissive approach, which may allow for greater use of copyrighted works for AI training without permission, while others may adopt a more restrictive approach, which might require explicit permission for such use.

---

70. *See, e.g.*, James Vincent, *Getty Images Sues AI Art Generator Stable Diffusion in the US for Copyright Infringement*, VERGE (Feb. 6, 2023, 10:56 AM), https://tinyurl.com/yne995e8 [https://perma.cc/9QYE-2WFT] (discussing the claims in the Getty Images lawsuit including copyright infringement and transformative use arguments).

In the EU, the legal framework for AI and copyright is established by the 2001 Information Society Directive and the 2019 Directive on Copyright in the Digital Single Market.[71] This directive provides a harmonized legal framework for the protection of copyrighted works in the digital environment.[72] However, it is silent on the specific issue of AI and copyright.[73] As a result, EU member states have some latitude in interpreting the directive and in developing their own laws in this area.[74]

In the United Kingdom (UK), whether or not a particular instance of copying constitutes fair dealing (the equivalent to fair use in the United States) would be the legal inquiry used to assess if a particular use of copyrighted works for AI training is permissible.[75] British copyright law employs a flexible approach.[76] It takes into account factors such as, but not limited to: the purpose and character of the use, the nature of the copyrighted work, and the portion of the work used.[77] When weighed together, these factors help decide whether the use is copyright infringement or fair dealing.[78]

In Canada, the concept of fair dealing may also be used to determine the legality of the use of copyrighted works for AI training.[79] Formerly, Canadian law appeared to possess less flexibility than UK law and had established a more limited set of circumstances in which fair dealing applies, but Canada's strictures have loosened recently.[80]

---

71. *See generally* Federico Ferri, *The Dark Side(s) of the EU Directive on Copyright and Related Rights in the Digital Single Market*, 7 CHINA-EU L. J. 21 (2021) (broadly discussing the 2019 and 2001 Directives and copyright law in the EU).

72. *See id.*

73. There are not currently any laws regulating AI in the EU, copyright or otherwise. *See* Luke Hurst, *ChatGPT in The Spotlight as The EU Steps Up Calls For Tougher Regulation. Is Its New AI Act Enough?,* EURONEWS.NEXT, (Feb. 6, 2023), https://tinyurl.com/nhjhc7ds [https://perma.cc/4LJ6-6ALJ] (discussing proposed draft rules to regulate AI in the EU).

74. *See* Barry Scannel, *When Irish AIs are Smiling: Could Ireland's Legislative Approach Be A Model For Resolving AI Authorship for EU Member States?*, 17 J. INTELL. PROP. L. & PRAC. 727, 731–32 (2022) (discussing the different EU member state approaches to authorship in copyright law and how they may apply to AI, indicating that member states can fill in the gaps when EU Directives do not provide the legal framework).

75. *See* Giuseppina D'Agostino, *Healing Fair Dealing? A Comparative Copyright Analysis of Canada's Fair Dealing to U.K. Fair Dealing and U.S. Fair Use*, 53 MCGILL L.J. 309, 337–45 (2008) (providing an overview of fair dealing in the UK).

76. *See id.* at 338 ("The U.K.'s enumerated purposes [to determine fair dealing] are said to be liberally construed.").

77. *See id.* at 342–43.

78. *See id.* at 343 (discussing the hierarchy of these factors with the market impact being the most important factor in UK courts).

79. *See id.* at 317–19 (discussing Canada's fair dealing statute broadly).

80. *See generally* Niva Elkin-Koren & Neil Weinstock Netanel, *Transplanting Fair Use Across the Globe: A Case Study Testing the Credibility of U.S. Opposition*, 72 HASTINGS L.J. 1121 (2021). The authors explain:

In Australia, the Copyright Act of 1968 established the legal framework for AI and copyright law.[81] This act contains provisions relating to the use of copyrighted works for research and study, among other uses, which may be relevant to the use of copyrighted works for AI training.[82] However, the exact scope of these provisions has not been clearly defined, and their applicability to AI training is uncertain.[83]

In conclusion, the approach to AI and copyright varies greatly between international jurisdictions, reflecting differences in legal systems and cultural attitudes. As AI continues to grow in significance, it will be important for the international community to develop a consistent and harmonized approach to the relationship between AI and copyright.

VIII. COPYRIGHT, AI, AND COURTS

The interaction between AI and copyright law is a relatively new area of legal inquiry, and there have been few court cases addressing the issue of the use of copyrighted works in AI training sets. As the use of AI continues to proliferate and expand, it is likely that more cases will be brought that test the limits of copyright law as it applies to AI.

Some of the few existing cases have dealt with questions related to the infringement of copyrighted works, such as the unauthorized use of images in machine-learning algorithms.[84] These cases have

---

> Canada's fair dealing exception was long thought to provide a closed list of uses that could qualify for the exception. But beginning in 2004, the Canadian Supreme Court has ruled that the specific permitted uses enumerated in Canada's fair dealing statute must be given a large and liberal interpretation and thus impose a low threshold, and that, in determining fairness, courts are to apply factors that overlap with those of U.S. fair use. Those rulings, together with Canadian Parliament's addition of parody, satire, and education to the list of enumerated uses, has brought a leading Canadian copyright scholar to conclude that "the current Canadian fair dealing regime now more closely resembles a flexible, open-ended fair use model."

Id. at 1125 n. 12 (quoting Michael Geist, *Fairness Found: How Canada Quietly Shifted from Fair Dealing to Fair Use*, *in* THE COPYRIGHT PENTALOGY: HOW THE SUPREME COURT OF CANADA SHOOK THE FOUNDATIONS OF CANADIAN COPYRIGHT LAW 157, 159 (Michael Geist ed., 2013)).

81. Copyright Act 1968 (Cth) (Austl.).

82. *Id.* pt III div 3 s 40.

83. *See* Madeleine Lezon, *Reforming 'Fair Dealing': An Analysis of Approaches to Copyright Exceptions in the United States and Australia*, ANU JOLT (Apr. 1, 2022) https://tinyurl.com/2mhynenr [https://perma.cc/SSF6-RV4J] (discussing the current lack of clarity in Australian fair dealing case law).

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

tended to focus on the commercial nature of the use and the amount of the copyrighted work that was used in the training process.[85]

Another notable case dealt with the reproduction of song lyrics in search engine results.[86] The court found the plaintiff, a supplier of song lyrics, had failed to state a claim upon which relief could be granted despite the fact that the defendant had reproduced song lyrics in results from user searches.[87]

In light of these cases, it appears that courts are still grappling with the appropriate balance between protecting the rights of copyright owners and allowing for the development and use of AI technologies. As AI's use continues to evolve, it will be interesting to see how courts balance these competing interests and whether they will recognize the concept of "fair training" as a valid defense in copyright infringement cases.

IX. A PROPOSAL TO RECOGNIZE A "FAIR TRAINING EXCEPTION" TO COPYRIGHT INFRINGEMENT

Copyright law in the United States contemplates uses of copyrighted works that, although carried out without permission from authors or owners, are nevertheless acceptable and do not trigger infringement. This fair use concept is enshrined in the Copyright Act at 17 U.S.C. §107,[88] which provides as follows:

> Notwithstanding the provisions of §§ 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—
> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
> (2) the nature of the copyrighted work;
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and

85. *See id.* at 214–25 (holding that each of the fair use factors "supported finding [Google's] activities were protected by fair use").

86. *See generally* Genius Media Group v. Google LLC & Lyricfind, 2020 U.S. Dist. LEXIS 173196, (E.D.N.Y Aug. 10, 2020).

87. *Id*.

88. Copyright Act of 1976, 17 U.S.C. § 107.

> (4) the effect of the use upon the potential market for or value of the copyrighted work.
>
> The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.[89]

Many court decisions have interpreted the requirements and application of § 107. In practice, courts apply each of the four enumerated factors in the statute to the facts of cases in which copyright infringement has been alleged.[90] The "purpose and character of the use" involves consideration of whether copying has been carried out to further a business or commercial purpose, or whether the copying has not implicated the making of a profit.[91] Copyrighted works come in many different forms (e.g., from extremely expressive to highly factual), and the particular "nature" of a copied work can be important in determining whether or not particular instances of copying are "fair."[92] Whether a large amount of a work of authorship is copied, rather than just a modest fraction, is another important factor in determining fair use. In general, the more of a work that has been copied, the less likely it is that such copying will be found to be fair use.[93] The fourth factor has much in common with the first. Copying that does not harm the ability of an owner to make money from their copyrighted work is more likely to constitute fair use.[94] On the other hand, copying that appropriates profits for the copyist which would otherwise have been available to the copyright owner tend not to constitute fair use.[95]

Once all four factors have been thoroughly evaluated, courts then typically undertake a balancing analysis.[96] There is no hard and fast rule about how this balancing test is to be resolved. Rather, a court will carefully evaluate the facts of the particular instance of copying, consider which factors weigh in favor of each party, take into

89. *Id.*

90. *See* Jacquelyn M. Creitz, Google LLC v. Oracle America Inc.*: The Court's New Definition of "Transformative" Expands the Fair Use Defense*, 17 J. BUS. & TECH. L. 317, 323 (2022).

91. *Id.* at 323–324.

92. *Id.* at 325 (2022) ("This factor recognizes that some works are more protected than others under copyright law because they fulfill the purpose of copyrights, to 'promote the sciences and the arts.'").

93. *Id.* at 326.

94. *Id.* at 326–27.

95. *Id.* at 327 ("If the reproduced work is commercial in nature, the work is presumed to be unfair.").

96. *Id.* at 323 ("The court will also consider each factor in relation to the other factors rather than by itself.").

account previous relevant court decisions, and then decide whether the copying amounted to fair use or not.[97] Though the laws of different countries differ in their particulars (e.g., whether they recognize the concept of fair dealing), the general contours of this sort of analysis are similar, resulting in a conclusion on whether a particular instance of copying is justified or not.

An overriding purpose of fair use or fair dealing is to ensure that society benefits from the copyright system.[98] Society benefits in one way when copyright owners feel secure in their rights, because this creates incentives for the creation of new works of authorship.[99] On the other hand, when copyright is too strictly protected, non-owners who might make valuable uses of owned works may be reluctant to engage in such uses, resulting in lost benefits to society.[100] Fair use attempts to maximize the net benefits (that is, the benefits minus the costs) that society gains from the copyright system.[101]

A *sine qua non* of most AI is the need for a training set.[102] Copyrighted works can be valuable components of a training set capable of helping an AI produce excellent new works for its users.[103] For example, an AI that generates new images based on users' queries will generally require access to a large number of existing—and often copyrighted—images for its training.[104] Even if the final images produced by this AI differ substantially from the images on which it trained, its need to train on copyrighted images may be crucial.[105] A similar example might involve written work made possible through a training set of existing, and copyrighted, writing. It is important to point out that, despite AI's need to use copyrighted works in training sets, the products of a generative AI tend to be substantially different from any of the individual copyrighted works that are part of its training set.[106] Moreover, generative AIs are usually designed not to copy or plagiarize the expressive elements of copyrighted works in a training set, but, rather, to make use of facts and patterns to

97. *Id.*

98. Jasmine Abdel-khalik, *Visual Appropriation Art, Transformativeness, and Fungibility*, 48 AIPLA Q.J. 171, 178–81 (2020).

99. *Id.*

100. *Id.*

101. *Id.*

102. *See* Jenny Quang, *Does Training AI Violate Copyright Law?*, 36 BERKELEY TECH. L.J. 1407, 1429–30 (2021).

103. *Id.*

104. *Id.*

105. *Id.* at 1410–12.

106. *See* Vincent, *supra* note 12 ("If the [text-to-image] model is training on millions of images and used to generate novel pictures, it's extremely unlikely that this constitutes copyright infringement. The training data has been transformed in the process, and the output does not threaten the market for the original art.").

compose new works.[107] Since copyright protects expressive, not factual, components of works of authorship, generative AIs will usually, and should, avoid copying elements of works having strong copyright protection.[108]

We propose that AI offers tremendous potential benefits for society. These benefits may be maximized by exposing AI to vast training sets that include works protected by copyright. The principle of fair use could be applied to training sets to determine whether inclusion of copyrighted works in a set used to train an AI constitutes "fair training." The existing fair use analysis could be adapted for training sets. We believe that, in general, such use of copyrighted materials in training sets would pass muster under a fair use-like analysis. Consequently, a "fair training" analysis would tend to allow the inclusion of copyrighted works in training sets used to improve AI. There could be cases in which inclusion of copyrighted works would fail the "fair training" test, such as where the AI itself, once trained, retained and reproduced substantial portions of works found in its training data set; in these circumstances, infringers would have to compensate owners, and sometimes be legally precluded from using owners' copyrighted works in training sets.[109] However, we believe a rigorous "fair training" analysis would allow most copyrighted material to be used for AI training, yielding generous benefits to society.

There may be technical or social mechanisms which creators could use to make their preferences known to AI systems. For example, creators who prefer not to have their works harvested for training sets could set a digital flag, similar to noindex for websites, that would ask AI systems not to use their content in training.[110] Alternatively, or complementarily, groups of creators could pool their works

---

107. *Id.*

108. *Compare* 17 U.S.C. § 102(a) ("Copyright protection subsists . . . in original works of authorship fixed in any tangible medium of expression."), *with* Vincent, *supra* note 12 ("If the [text-to-image] model is training on millions of images and used to generate novel pictures, it's extremely unlikely that this constitutes copyright infringement. The training data has been transformed in the process, and the output does not threaten the market for the original art."). If the AI output is transformative of the original works of art, then it would avoid copyright infringement and should instead have copyright protection.

109. Sarah Ligon Pattishall, *AI Can Create Art, but Can It Own Copyright in It, or Infringe?*, LEXIS PRAC. GUIDANCE J. (Mar. 1, 2019), https://tinyurl.com/5fvb5pz3 [https://perma.cc/8HGY-PHPY] ("If an AI-artist sells or displays AI-art that is substantially similar to the underlying work, it is unlikely the AI-artist will be able to rely on fair use.")

110. *See generally Block Search Indexing with 'Noindex'*, GOOGLE SEARCH CENT., https://tinyurl.com/mr2tfh95 [https://perma.cc/C6X3-SAEQ] (last visited Aug. 20, 2023) (describing how users can use a noindex meta tag "to prevent indexing content by search engines that support the noindex rule").

to produce known, licensable training sets, available for a fee, similar to an ASCAP license for music.[111]

Currently, AI systems are produced by both major corporations, academic institutions, and individuals; these entities may not have access to similar levels of personnel or financial resources. Given these disparities, we suggest that AI training sets be made available more readily to organizations and individuals without large amounts of money, so that those organizations and individuals may ensure diverse contributions to the development of future AI systems.

In sum, we believe that the potential good that AI can do is vast, and that training sets are necessary for many forms of AI to flourish. We encourage the legal and creative communities to work together with technologists to develop viable processes for large-scale training sets to be widely available, especially to AI systems that do not have substantial financial backing.

X. FUTURE IMPLICATIONS AND THE ROAD AHEAD

The debate on the compatibility of AI and copyright law continues to evolve. As AI technology continues to advance, the use of copyrighted material in AI training sets will likely become more widespread. Therefore, it is important to consider the potential implications of this development and determine a clear legal framework for AI and copyright.

The concept of "fair training" has yet to be fully tested in the court system, and the outcome of any legal challenges will have significant consequences for the future of AI development. In addition, the international approach to AI and copyright is still fragmented, with some countries taking a more lenient view of the use of copyrighted material in AI training, while others take a stricter stance.

As AI becomes increasingly integrated into our daily lives, it is crucial to find a balance between the protection of copyright holders and the advancement of AI technology. The road ahead will likely involve ongoing debates, legislative action, and, potentially, legal challenges that will determine the future of AI and its relationship with copyright law.

111. *See generally ASCAP Licensing, ASCAP: Frequently Asked Questions*, AM. SOC'Y OF COMPOSERS, AUTHORS, & PUBLISHERS https://tinyurl.com/mr2rk46m [https://perma.cc/557S-BTMX] (last visited Aug. 24, 2023) (describing how ASCAP licensing works).

CONCLUSION: BALANCING AI AND COPYRIGHT PROTECTIONS

The use of copyrighted works as training sets for AI is a complex issue that raises important questions about the balance between the rights of copyright owners and the potential societal benefits of AI. On one hand, proponents of "fair training" suggest that the use of copyrighted works in AI training sets is a transitory and non-consumptive use that does not materially interfere with owners' copyrights. Because the use of copyrighted works in training AIs is transformative, such use should be considered a form of fair use under U.S. law, or fair dealing in some other countries, that qualifies as "fair training."[112] On the other hand, opponents argue that incorporating copyrighted works into training sets for AI is a misappropriation of owners' intellectual property and not fair use under the law.[113]

International approaches to AI and copyright have varied, and case law specifically applicable to training sets of data is thus limited, making it difficult to predict the outcome of pending and future disputes. Many believe an overriding goal should be to balance promoting the development and use of AI for the benefit of society against the rights of copyright owners, while others are skeptical that such balancing is a justifiable goal.

112. *See supra* notes 40–47 and accompanying text.
113. *See supra* note 34 and accompanying text.

***