SMU Science and Technology Law Review

Volume 27 | Number 1 Article 7

2024

The Fall of Z-Library: The "Burning of the Library of Alexandria" or Protection for Authors Against AI Companies

Lisa Silveira
Southern Methodist University, Dedman School of Law

THE FALL OF Z-LIBRARY: THE "BURNING OF THE LIBRARY OF ALEXANDRIA"[1] OR PROTECTION FOR AUTHORS AGAINST AI COMPANIES?

*Lisa Silveira**

ABSTRACT

The development and advancement of artificial intelligence ("AI") is changing the way we use technology while creating an ongoing battle between media and technology companies. With AI companies gathering data from the internet to train programs like ChatGPT, authors have growing concerns about unpermitted use of their work when pirated copies of their books exist illegally online through shadow libraries. This article examines the popular shadow library known as Z-Library and the views of its proponents and opponents. In addition, this article will discuss the training process AI companies use and the data sets containing content from shadow libraries. While companies like Getty Images and The New York Times filed suit against AI companies, this article specifically focuses on the class action lawsuits filed by authors for unauthorized use of their books to train AI models. Copyright law may offer a solution to protect these author's works. This article will examine the current limitations of copyright law and the difficulties of proving copyright infringement. This article attempts to explore the current legal action, claims these authors raise, and possible defenses they will have to overcome. This article will also examine solutions like agreements with authors and paying them royalties to compensate them for the use of their work. Regardless of how the court cases come out, these authors need a solution to ensure their content is not exploited by AI companies.

I. INTRODUCTION

Shadow libraries have existed online for many years—prior to the prevalence of companies using artificial intelligence. Shadow libraries, for example—Library Genesis, Z-Library, and Anna's Archive—are file-sharing platforms that store books, articles, academic and scholarly texts, and

---

https://doi.org/10.25172/smustlr.27.1.7

* Lisa Silveira is a J.D. candidate at SMU Dedman School of Law. She received a Bachelor of Arts in Communication with a minor in English from Texas A&M University in 2022.

1. @tedsbecca, Twitter (Nov. 4, 2022, 5:18 AM), https://twitter.com/tedsbecca/status/1588475791897395200.

textbooks, and they allow users to illegally download these materials for free.[2] Although some authors upload their own works onto shadow libraries, most of the materials on these websites are unapproved, pirated copies of author's works.[3] While shadow libraries are not illegal as a whole, these pirated versions of materials available for download infringe on author's copyrights.[4] Many students and teachers use these websites to access materials at no cost, or when they are unavailable elsewhere.[5] Authors can upload their own works onto shadow libraries in order to share them with a broader audience.[6] However, when shadow libraries offer works of authors without their permission, they often infringe copyrights held by authors and publishers, impacting sales in the publishing industry.[7]

Although shadow libraries are problematic to authors on their own, artificial intelligence adds a new level of concern for authors and their works.[8] AI is a broad term used to refer to "applications of technology to perform tasks that resemble human cognitive function"[9] and is the "capability of a machine to imitate intelligent human behavior."[10] For authors, AI companies threaten their work by imitating their style or regurgitating portions of their work.[11]

2. *See* Martin Schweiger, *The Actual Z-Library Case Demonstrates That Copyright Law Needs An Overhaul*, Lexology (Nov. 10, 2022), https://www.lexology.com/library/detail.aspx?g=1bf88436-55f4-498a-8bd0-cb07842b41c8 [https://perma.cc/QW7Y-ZRGZ].

3. Riddhi Setty, *Rampant 'Shadow Libraries' Drive Calls for Anti-Piracy Action*, Bloomberg Law (Oct. 19, 2022, 4:03 AM), https://news.bloomberglaw.com/ip-law/rampant-shadow-libraries-drive-calls-for-anti-piracy-action [https://perma.cc/2WV8-JF4Y].

4. *Id.*

5. Schweiger, *supra* note 2.

6. *Id.*

7. Setty, *supra* note 3.

8. *See* Ellen Glover, *AI-Generated Content and Copyright Law: What We Know*, Built In (Aug. 23, 2023), https://builtin.com/artificial-intelligence/ai-copyright [https://perma.cc/2Z8M-BBAN].

9. *Overview of Artificial Intelligence Technology*, Finra, https://www.finra.org/rules-guidance/key-topics/fintech/report/artificial-intelligence-in-the-securities-industry/overview-of-ai-tech#:~:text=The%20term%20artificial%20intelligence%20broadly,of%20computer%20systems%20able%20to [https://perma.cc/DQP3-L4H7].

10. *Artificial Intelligence*, Merriam Webster, https://www.merriam-webster.com/dictionary/artificial%20intelligence [https://perma.cc/9RXT-PXM7].

11. Glover, *supra* note 8.

Additionally, authors are concerned with the content AI companies train on.[12] As AI companies grow and more people access AI programs, the rights of authors and creators whose works exist on the internet come into question.[13] Programs that have recently gained popularity include ChatGPT, "a natural language processing tool driven by AI technology," allowing users to have conversations, ask questions, and compose written materials.[14] To train AI programs, the program learns through data like "numbers, photos, or text."[15] Once this data is gathered, "programmers choose a machine learning model to use, supply the data, and let the computer model train itself to find patterns or make predictions."[16] Since AI programs need to generate output like a written sentence, "it must first learn from the real work of actual humans."[17] Books are a good source for AI programs to train on since they contain lengthy examples of high-quality writing.[18] In recently filed lawsuits, authors allege most of the book data AI programs use to train is taken from illegal shadow libraries that house author's works.[19]

In the United States, state laws governing the right of publicity, copyright law, and the First Amendment have the potential to shape the laws governing the use of AI programs.[20] The right of publicity is used more frequently in AI cases pertaining to a person's likeness being used in media.[21] Authors, who own the copyrights to their works, are at risk when AI training programs copy

12. *See id.*

13. *See id.*

14. Sabrina Ortiz, *What is ChatGPT and Why Does it Matter? Here's What You Need to Know*, ZDNET (Sep. 25, 2023), https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/ [https://perma.cc/5E3R-WZ5W].

15. Sara Brown, *Machine Learning, Explained*, MIT Sloan School of Management (Apr. 21, 2021), https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained [https://perma.cc/NK95-E896].

16. *Id.*

17. Glover, *supra* note 8.

18. Complaint at 5, Silverman v. OpenAI, Inc., No. 3:23-CV-03416 (N.D. Cal. filed July 7, 2023); *see also* Ortiz, *supra* note 14.

19. Michelle Cheng, *"Shadow Libraries" are at the Heart of the Mounting Copyright Lawsuits Against OpenAI*, Quartz (July 10, 2023), https://qz.com/shadow-libraries-are-at-the-heart-of-the-mounting-cop-1850621671 [https://perma.cc/6KZF-9RF6].

20. Ethan Beberness, *2 Creative IP Attorneys On The Complications Of AI-Generated Art*, Above the Law (July 27, 2023), https://abovethelaw.com/2023/07/2-creative-ip-attorneys-on-the-complications-of-ai-generated-art/ [https://perma.cc/3EKD-4YUQ].

21. *See infra* Part V and notes 104–05.

their works without their consent, without giving credit to them, and without compensation.[22] The First Amendment addresses "whether a new work is transformative, newsworthy, or a parody."[23]

This comment will focus on the positive and negative views on shadow libraries and the impact these websites have on AI programs' training process. It will explore recent legal suits AI companies face and how the First Amendment, state laws, and copyright law impact AI companies. Finally, it will consider future implications and problems between AI companies and shadow libraries.

II. THE FALL OF Z-LIBRARY

Z-Library was created in 2009 as a "free file sharing platform for academic and scholarly articles," initially mirroring another shadow library website called Library Genesis.[24] Z-Library eventually grew in popularity due to its easy accessibility and modern website layout.[25] It relied on donations from its users for funding.[26] Z-Library ranked in the top 10,000 most visited websites on the internet, offering millions of free books and articles.[27]

In early November 2022, the U.S. Department of Justice seized Z-Library.[28] Despite removing this website and blocking the domain, other websites still exist that mirror the content once housed by Z-Library, just under a different domain.[29]

Shadow libraries have a number of proponents and opponents and many opinions were revealed when the government removed Z-Library, one of the largest and most popular shadow libraries.[30] Supporters shared alternative rea-

22. Complaint, *supra* note 18, at 5, 11–12.

23. Beberness, *supra* note 20.

24. Bill Toulas, *Z-Library eBook Site Domains Seized by U.S. Dept of Justice*, BleepingComputer (Nov. 4, 2022, 1:53 PM), https://www.bleepingcomputer.com/news/technology/z-library-ebook-site-domains-seized-by-us-dept-of-justice/ [https://perma.cc/5XQL-8HEA].

25. *Id.*

26. *Id.*

27. *Federal Law Enforcement Arrests and Indicts Z-Library Operators with AG's Assistance*, The Authors Guild (Nov. 16, 2022), https://authorsguild.org/news/federal-law-enforcement-indicts-z-library-operators-with-ag-assistance/#:~:text=Z%2DLibrary%2C%20which%20had%20been,websites%20on%20the%20internet%20worldwide [https://perma.cc/79M5-5CAP].

28. Schweiger, *supra* note 2.

29. *Id.*

30. *See* Allison Rumfitt, *In Defence of Z-Library and Book Piracy*, Dazed (Nov. 25, 2022), https://www.dazeddigital.com/life-culture/article/57545/1/

sons individuals choose to use shadow libraries other than simply to get a free copy.[31] Some proposed alternatives are to find copies of materials that are out of print, out of stock, or never officially published.[32] As electronic books ("e-books") become more popular, shadow libraries offer instantaneous downloads as opposed to the time it takes to purchase a print copy of a book.[33] From an economic standpoint, shadow libraries provide materials at no cost to individuals who may be unable to afford them or do not have the resources to access them.[34] Not all communities have access to libraries, let alone a library as well stocked as websites like Z-Library.[35]

On the other hand, shadow libraries are opposed by authors who do not consent to their works being posted to shadow libraries.[36] Those in the book industry argue shadow libraries steal profits away from them.[37] One study estimated "[p]irated e-books have depressed legitimate book sales by as much as 14%."[38] Authors stated they released books and saw them posted on Z-Library within the same day.[39] The websites have received thousands of take down notices sent by authors and publishers under the Digital Millennium Copyright Act, but the website domain owners often ignore these notices unless they are compelled to take down their website by court order.[40] In the United States, there is no effective remedy to combat shadow libraries, which makes it difficult to address these websites.[41] Users face harm since the pirated files often contain malware or other viruses that can harm the downloading device.[42] Overall, shadow libraries harm authors and publishers by taking away

---

in-defence-of-piracy-and-z-library-shut-down-alison-rumfitt-writer-author [https://perma.cc/TF3K-F8CM].

31. *See id.*

32. *Id.*

33. *Id.*

34. *Id.*

35. *Id.*

36. *See* Setty, *supra* note 3.

37. *Id.*

38. *Id.*

39. *Id.*

40. *Id.*

41. *Id.*

42. Dan Holloway, *Self-Publishing News: New Lawsuit Against LibGen Brings Shadow Libraries into the Light*, Alliance of Independent Authors (Sept. 19, 2023), https://selfpublishingadvice.org/self-publishing-news-libgen-lawsuit-shadow-libraries/ [https://perma.cc/QKQ9-B8NF].

the rights they have to their own published works. These complaints regarding shadow libraries are only amplified with the involvement of AI programs.

III. TRAINING PROCESS OF AI PROGRAMS THROUGH USE OF SHADOW LIBRARIES

Shadow libraries, already disliked by authors and publishers, are now being used on a more commercialized scale as AI programs train from them. For example, OpenAI (the company behind ChatGPT) trained on datasets known as "Books1" and "Books2," with Books1 containing an estimated 63,000 titles and Books2 containing 294,000 titles.[43] Although OpenAI did not disclose the specifics of the data it used to train ChatGPT, the only internet-based websites that offer that much book material are shadow libraries like Library Genesis or Z-Library.[44] In a similar manner, Shawn Presser, an independent artificial intelligence researcher, and a few of his fellow peers created "Books3" by downloading entire shadow libraries, converting the files, and creating a library of around 196,000 books "including works by popular authors like Stephen King, Margaret Atwood, and Zadie Smith."[45] Books3 was named after Books1 and Books2, and Presser was inspired by Books1 and Books2 to create a similar dataset from online shadow libraries to make available to the public.[46] In October 2020, Books3 went online as "a way to democratize access to the kind of data sets OpenAI was already using."[47] In a recent analysis of the data contained in Books3, Alex Reisner, a programmer and technical consultant, found the books stored in this dataset to be in "large, unlabeled blocks of text."[48] To identify the authors and titles within these text blocks, Reisner pulled the ISBNs from the text and searched them through a book database.[49] Reisner identified 191,000 book titles and associated author information to 183,000 of them, then compiling this data into a searchable database to assist authors in determining if their works are included in the training set.[50] This information was helpful to authors,

43. Complaint, *supra* note 18, at 6–8.

44. *Id.* at 7.

45. Kate Knibbs, *The Battle Over Books3 Could Change AI Forever*, Wired (Sep. 4, 2023, 6:00 AM), https://www.wired.com/story/battle-over-books3/ [https://perma.cc/7PS7-QK9Z].

46. *Id.*

47. *Id.*

48. Alex Reisner, *What I Found in a Database Meta Uses to Train Generative AI*, The Atlantic (Sep. 25, 2023), https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/ [https://perma.cc/XT67-6NTN].

49. *Id.*

50. *Id.*

many of whom were unaware their books were contained in the dataset.[51] AI program training practices often remain secretive and nonconsensual to those individuals whom the data is taken from, which is why a deep analysis like Reisner's was necessary to understand what components make up the datasets.[52]

While Presser's goal was to level the playing field for smaller companies, researchers, and independent people to create large language models, big companies took advantage of this dataset as well.[53] Companies like Meta and Bloomberg acknowledged they trained their large language models with Books3.[54] Thus, it is not a mystery how these datasets came about, and the evidence shows that, due to their immense data, Meta and Bloomberg sourced it from shadow libraries.[55]

The U.S. government has taken steps to address shadow libraries, like its removal of Z-Library in 2022.[56] However, these websites often appear online again under different domains and are not easy to erase from the internet entirely.[57] It's difficult to target or hold someone legally accountable for shadow libraries, since they are often websites created by individuals in other countries.[58] It is challenging to establish jurisdiction, shut them down completely, or sue the individuals creating shadow libraries.[59] One solution other countries have attempted is blocking these websites;[60] however, users can get around this easily by using a VPN to change their location.

Rather than targeting shadow libraries themselves, which as mentioned previously, may offer personal benefits to some individuals or may be difficult to accomplish, AI companies need to be held accountable for using data from these websites. It is "up to AI companies whether or not to disclose where their training sets come from."[61] Otherwise, it will be difficult for individuals to prove their data was used without their consent.[62] Europe "passed a draft

51. *Id.*

52. *Id.*

53. Knibbs, *supra* note 45.

54. *Id.*

55. *Id.*

56. Toulas, *supra* note 24.

57. *Id.*

58. Setty, *supra* note 3.

59. *Id.*

60. *Id.*

61. Knibbs, *supra* note 45.

62. *Id.*

law of AI regulations that would require increased data transparency," and the United States should follow suit to protect the rights of creators.[63]

Other solutions, like AI companies offering compensation to authors for works used or removing their books from their training set if requested, would also aid authors.[64] AI companies should pay royalties to authors when their work is being used to generate text or when their work is being used to train AI programs.[65] Just as consumers are expected to pay to read authors' works, AI companies should be held to the same standard, and authors should receive initial royalties from AI companies purchasing a copy of their work to read and train on.[66] Overall, authors need to be compensated in some form for contributing to the advancement and training of AI programs.[67]

These authors spent years thinking, researching, imagining, and writing, and had no idea that their books were being used to train machines that could one day replace them. Meanwhile, the people building and training these machines stand to profit enormously.[68]

IV. NEWEST FRONTIER OF COPYRIGHT LAW: CLAIMS AGAINST AI

Suits against AI companies are not uncommon—visual artists have sued other AI companies for copyright infringement.[69] In February of 2023, Getty Images sued Stability AI for copyright infringement of images existing on the internet that this AI company used to train its AI tools.[70] The New York Times sued Microsoft and OpenAI for copyright infringement of content from its journalists that the technology companies trained on.[71] Programmers also filed a class-action lawsuit against GitHub and OpenAI, alleging a particular AI

63. *Id.*

64. Glover, *supra* note 8.

65. *Id.*

66. *See id.*

67. *Id.*

68. Alex Reisner, *These 183,000 Books are Fueling the Biggest Fight in Publishing and Tech*, The Atlantic (Sept. 25, 2023), https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/ [https://perma.cc/97GS-WTGF].

69. Cheng, *supra* note 19.

70. Alexandra Bruell, *New York Times Sues Microsoft and OpenAI, Alleging Copyright Infringement*, The Wall Street Journal (Dec. 27, 2023, 3:22 PM), https://www.msn.com/en-us/money/companies/new-york-times-sues-microsoft-and-openai-alleging-copyright-infringement/ar-AA1m6dAn?ocid=hpmsn&cvid=aebea2bfd1b3487fa81a2977fc3dd2c2&ei=22 [https://perma.cc/SDN8-MG2V].

71. *Id.*

product "relie[d] on 'unprecedented open-source software piracy.'"[72] While some of these other cases pertain to issues with the material AI programs produce in their output, the focus here will be on the training models used to teach AI programs how to produce its final product, or how we address AI program input if it is being trained on copyrighted material.[73]

A. Class Action Lawsuits

Recently, authors filed federal lawsuits in the Northern District of California against OpenAI, the creator of ChatGPT, and Meta, which has a large language model for training on authored content without author permission.[74] A large language model (LLM) is AI software that aids the program in producing natural language.[75] LLMs are not programmed traditionally, but instead are "'trained' by copying massive amounts of text and extracting expressive information from it."[76] The natural text output is thus reliant on what material it is being trained on in the dataset.[77] LLMs "continually adjust the way they interpret and make sense of data."[78] The text data that is analyzed by LLMs is processed through a neural network, "a commonly used type of AI engine made up of multiple nodes and layers" with most LLMs using a specific type of neural network architecture known as a transformer.[79] Transformers "read vast amounts of text, spot patterns in how words and phrases related to each other, and then make predictions about what words should come next."[80] AI models are not formulating their own ideas, but are figuring out how words follow each other to mimic real thought processes.[81] These authors allege copyright infringement: that they did not authorize AI

72. Cheng, *supra* note 19.

73. *Artificial intelligence, free speech, and the First Amendment*, FIRE, https://www.thefire.org/research-learn/artificial-intelligence-free-speech-and-first-amendment [https://perma.cc/X6Y9-8FXA].

74. Cheng, *supra* note 19.

75. Complaint, *supra* note 18, at 4–5.

76. *Id.* at 1.

77. *Id.*

78. David Nield, *How ChatGPT and Other LLMs Work—and Where They Could Go Next*, WIRED (Apr. 30, 2023, 7:00 AM), https://www.wired.com/story/how-chatgpt-works-large-language-model/ [https://perma.cc/8VDZ-56V6].

79. *Id.*

80. *Id.*

81. Glover, *supra* note 8.

programs to train on their books they hold copyrights to.[82] In order for AI programs to generate output like a written sentence, they "must first learn from the real work of actual humans."[83] Therefore, if an AI generator is asked to produce text in the style of Toni Morrison, "it has to be trained with words written by Toni Morrison."[84] AI program training is problematic to these authors when they do not consent to the use of their works. Authors leading these lawsuits include Mona Awad, author of *Bunny*; Paul Tremblay, author of *The Cabin at the End of the World*; Sarah Silverman, author of *The Bedwetter*; Richard Kadrey, author of *Ararat*; and Christopher Golden, author of *Sandman Slim*.[85] These lawsuits are supported by The Authors Guild, an organization advocating "for the rights of writers by supporting free speech, fair contracts, and copyright."[86]

The First Amendment of the United States Constitution should protect the use of AI companies to "create, disseminate, and receive information."[87] This appears to apply more to the output generated by AI programs. Freedom of speech and expression that is protected by the First Amendment is likely to target more of what users generate from AI tools like ChatGPT.

Thus, state laws are potentially more protective than the First Amendment in regard to these author's claims attacking the training of OpenAI. In one of the class action complaints, the authors alleged unfair competition under the California Business and Professions Code section 17200.[88] The unlawful business practices of AI programs using author's infringed works to train ChatGPT is unfair, immoral, and unethical, among other claims made under this code.[89] This complaint also cites other claims under California common law, like negligence and unjust enrichment, which may be easier to tailor to their specific claim rather than a broad interpretation of the First

82. Cheng, *supra* note 19.

83. Glover, *supra* note 8.

84. *Id.*

85. *See* Complaint, *supra* note 18, at 2–3; Alexandra Tremayne-Pengelly, *Mona Awad and Paul Tremblay are the Latest Creatives to Sue Over A.I.*, Observer (Jul. 7, 2023 12:35 PM), https://observer.com/2023/07/mona-awad-paul-tremblay-sue-chatgpt-copyright-infringement/ [https://perma.cc/V5DT-U7QF].

86. *Supporting working writers and protecting authors' rights since 1912*, Authors Guild, https://authorsguild.org/#:~:text=Supporting%20working%20writers%20and%20protecting,fight%20for%20a%20living%20wage [https://perma.cc/9B69-D85M].

87. *Artificial intelligence, free speech, and the First Amendment*, *supra* note 73.

88. Complaint, *supra* note 18, at 13.

89. *Id.*

Amendment.[90] The authors allege negligence since the AI companies owed a duty of care to the authors based on the companies' "obligations, custom and practice, and right to control information in its possession," a duty also based on requirements the companies act in a reasonable manner toward others.[91] The companies allegedly breached this duty by "negligently, carelessly, and recklessly collecting, maintaining and controlling [authors'] Infringed Works and engineering, designing, maintaining and controlling systems—including ChatGPT—which are trained on [authors'] Infringed Works without their authorization."[92] The authors allege unjust enrichment since the authors did not consent to the unauthorized use of their infringed materials to train ChatGPT and the authors were "deprived of the benefits of their work" while the AI companies "derived profit and other benefits from the use of the Infringed Materials to train ChatGPT."[93] Authors who find violations within their state codes and statutes may have an easier time asserting these claims in court, but this makes it difficult for other authors in other states to bring similar claims since different states have different laws. It would require a lot of time, money, and research for other authors to hire attorneys and be the first in their state to bring these claims against AI companies under different state laws. Furthermore, an issue with state or local laws regulating AI companies is creating a patchwork of different laws and compliance. In response, these companies can relocate their headquarters and operate out of states where the law works more in their favor.

Fair use law currently "permits the use of copyrighted material under certain conditions without needing the permission of the owner."[94] Thus using copyrighted materials is allowed when training AI models.[95] However, currently pending lawsuits by authors are attempting to regulate the materials being used to train AI programs.[96] Authors fear AI-generated work mimicking their style may eat into their commissions or profits.[97]Authors bringing lawsuits are most likely to sue under U.S. copyright law. Copyright law relates to both the output issue—whether AI programs are creating a transformative work when it mimics an author's writing style or offers summaries of their works—and the input issue of AI program training on copyrighted material.[98]

90. *Id.* at 14.

91. *Id.*

92. *Id.*

93. *Id.* at 15.

94. Glover, *supra* note 8.

95. *Id*.

96. *Id.*

97. *Id.*

98. Beberness, *supra* note 20.

However, critics assert these are weak claims, since training AI programs "does not require 'copying' the work in question, but rather reading it," and merely reading a copyrighted work cannot constitute copyright infringement.[99] When comparing AI programs to creators in general:

> Humans read, listen, watch, learn from, and are inspired by those who came before them. And then they synthesize that with other things, and create new works, often seeking to emulate the styles of those they learned from. AI systems and LLMs [large language models] are doing the same thing. It's not infringing to learn from and be inspired by the works of others. It's not infringing to write a book report style summary of the works of others.[100]

Generally speaking, an individual using publicly available information does not violate copyright law when they are inspired by other works.[101] Therefore, authors may have a difficult time proving copyright claims in these cases. While "[i]t may be beyond the scope of copyright law to address the harms being done to authors by generative AI," authors have rights to their works which must be protected in some form.[102]

As a defense to copyright claims, companies training AI programs will likely argue fair use, "the legal doctrine that permits the use of copyrighted material under certain circumstances, enabling parody, quotation, and derivative works that enrich the culture."[103] These companies will claim "generative-AI tools do not replicate the books they've been trained on but instead produce new works, and that those new works do not hurt the commercial market for the originals."[104] While this argument is strong, it may be damaged by the fact that the books AI programs trained on were acquired without permission from an unauthorized source.[105] The intentions and knowledge of AI companies may be a relevant factor, especially if AI companies claim to have no idea where the books they trained on came from.[106] These companies should

---

99. Mike Masnick, *A Bunch Of Authors Sue OpenAI Claiming Copyright Infringement, Because They Don't Understand Copyright*, TECHDIRT, https://www.techdirt.com/2023/07/11/a-bunch-of-authors-sue-openai-claiming-copyright-infringement-because-they-dont-understand-copyright/ (July 11, 2023, 9:29 AM) [https://perma.cc/4NJW-7UHR].

100. *Id.*

101. *Artificial intelligence, free speech, and the First Amendment*, *supra* note 73.

102. Reisner, *supra* note 48.

103. Alex Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative AI*, THE ATL., https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/ (Sept. 25, 2023, 1:40 PM) [https://perma.cc/56NM-FK33].

104. *Id.*

105. *Id.*

106. *Id.*

be more transparent and disclose where they gather data used to train their systems. Companies also need to be held accountable to keep track of which authors may be harmed or which authors need to be compensated. Fair use law governing the use of unauthorized materials remains unsettled, and "previous cases [give] little indication of how a judge might rule in the future.[107] Thus, authors face the challenge of establishing whether AI programs training from illegal websites like shadow libraries is truly copyright infringement; AI programs merely reading materials that exist on the internet may not be enough to constitute copyright infringement. In direct response to one of the lawsuits, lawyers for Meta argued the case should be dismissed in part because neither the large language model nor its outputs are "substantially similar" to the authors' books.[108] This distinction of the outputs not being "substantially similar" is important since "copyright law's predominant means for determining copyright infringement" is the "substantial similarity" test.[109] Under this test, courts "assess whether an alleged infringer has taken so much of a copyright holder's protectible material as to constitute copyright infringement."[110] This is similar to the previous fair use argument, and it seems likely individuals supporting AI companies will continue to assert what they are training on and creating is considered a new work.

Depending on how these cases are decided, if it is determined AI companies are infringing on author's copyrighted works, these companies could face large penalties.[111] Under copyright law, infringing on one work can result in an award of $150,000.[112] As a result, the thousands of books AI is training off of could cost these companies a fortune.[113] Thus, it would be cheaper overall for companies to pay authors royalties for use of their works or establishing a system where they pay authors based on if their work is used to generate output, rather than take the risk of losing in court and paying much more per each infringed-upon work. Additionally, AI companies should attempt to work with and stay in the good graces of the authors since authors' works contribute to the advancement of their AI systems.

107. *Id.*

108. Reisner, *supra* note 48.

109. Clark D. Asay, *An Empirical Study of Copyright's Substantial Similarity Test*, 13 U.C. Irvine L. Rev. 35, 37 (2022).

110. *Id.*

111. Glover, *supra* note 8.

112. *Id.*

113. *See id.*

V. CONCLUSION

AI systems are being trained through illegal copies of authors' works that exist on the internet through shadow libraries. The issue of at what point AI systems' use of these works becomes copyright infringement is a growing concern.

OpenAI's training on books acquired from shadow libraries calls into question what other sources and data AI programs are using to train. Although individuals may have intended to share their ideas with others through the internet, they may not have intended to provide ideas for commercial AI programs. While authors are currently the main creators expressing these concerns through lawsuits since their works are directly involved in the data training sets like Books3, other creators should be concerned with what other materials exist on the internet that AI programs could use to train on in the future. Other websites hosting copyrighted material, like photos of artwork or movies and television shows can also be used to train AI programs. Artists may not agree to having their works on the internet used to train AI programs, and in turn their style of artwork mimicked by the program. Dialogue of television shows and movies that exist as videos on the internet may be used to train AI systems on how to write scripts. As a parallel, the ongoing Screen Actors Guild - American Federation of Television and Radio Artists (SAG-AFTRA) writer and actor strikes demand protection of writers' and actors' images and performances from appropriation by AI programs without their informed consent and fair compensation.[114] This includes AI programs being used to write scripts for films and television based on work writers have already written. In direct opposition, the Alliance of Motion Picture and Television Producers (AMPTP) wants to use an individual's likeness for any purpose without their consent and "be able to use someone's images, likenesses, and performances to train new generative AI systems without consent or compensation."[115] This is similar to the problems authors whose works are posted to shadow libraries are facing, where their current books, like the current scripts writers have written, can be used to train AI programs to produce something entirely new that threatens their careers. Thus, other creators are facing similar problems regarding AI systems trained on creator's works without their consent.

Authors and other creators are concerned with their works being used without their permission and others profiting off their works while they receive no compensation. Additionally, ghostwriting, one person writing in the name of another individual without receiving credit that is tied to the authors' style

114. *Why We Strike*, SAG-AFTRA, https://www.sagaftrastrike.org/why-we-strike (last visited Feb. 2, 2024) [https://perma.cc/SMH7-8FSS].

115. *Id.*

has existed for many years.[116] AI programs could potentially make ghostwriting even more normalized and replace authors or threaten their contracts.

While a culture of piracy is often the norm in the age of the internet, and AI programs seem to be taking a natural step to further this notion, this culture has existed through "mostly personal use by individual people."[117] It is an entirely new trend for AI programs to exploit pirated books for profit, "with the goal of replacing the writers whose work was taken."[118] Authors deserve protection and have rights to the works they create, and AI programs should not take advantage of their skills and exploit them.

116. *What Is Ghostwriting—And What Does It Mean Today?*, GOTHAM GHOSTWRITERS, https://gothamghostwriters.com/what-is-ghostwriting-and-what-does-it-mean-today/ (last visited Feb. 2, 2024) [https://perma.cc/Z27Z-B29A].

117. Reisner, *supra* note 103.

118. *Id.*