UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

THE CENTER FOR INVESTIGATIVE REPORTING, INC.,

Plaintiff,

v.

OPENAI, INC., OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, OPENAI HOLDINGS, LLC, and MICROSOFT CORPORATION

Defendants.

Civil Action No. _____

COMPLAINT

JURY TRIAL DEMANDED

1. Plaintiff The Center for Investigative Reporting, Inc. ("CIR"), through its attorneys Loevy & Loevy, for its complaint against Defendants Microsoft Corporation ("Microsoft") and OpenAI, Inc., OpenAI GP LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, (collectively "OpenAI" and, with Microsoft, "Defendants") alleges the following:

NATURE OF THIS ACTION

2. Independent, nonprofit news reporting is a critical and unique voice in the United States media landscape. Founded in 1976, CIR is the oldest nonprofit newsroom in the country. CIR's sole purpose is to benefit the public by reporting investigative stories about underrepresented voices in our democracy. For decades CIR has published valuable, one-of-a-kind, award-winning reporting that highlights diverse communities that are often overlooked. In just the last few months, CIR was awarded the George Polk Award, a Peabody Award, a Webby Award, and Robert F. Kennedy Human Rights Award for its unique reporting on diverse subjects,

including prosecution of alleged sexual assault victims, abuse in the Mormon Church, and police procedures that injure families.

3. To sustain itself in today's notoriously challenging media market, CIR has worked especially hard to survive while continuing to tell stories that are usually untold and left unseen. CIR has developed ways to gain revenue for its reporting, including license, advertising, and affiliate revenue, and has created partnership agreements and programs compatible with its mission to bring in new revenue. CIR has dedicated staff to develop streams of revenue to fund its reporting, including staff dedicated to licensing, advertising, revenue, and partnerships.

4. Defendants are companies responsible for the creation and development of the highly lucrative ChatGPT and Copilot artificial intelligence (AI) products, which are built on uncompensated and unauthorized use of the creative works of humans. According to the award-winning website Copyleaks, nearly 60% of the responses provided by Defendants' GPT-3.5 product contained some form of plagiarized content, and over 45% contained text that was identical to pre-existing content.

5. These systems, and the large language models (LLMs) that power them, are trained using human works. In particular, AI systems and LLMs ingest and use human-made journalism to attempt to mimic how humans write and speak in an effort to compete for the attention of consumers to generate profits. These training sets have included hundreds of thousands, if not millions, of works of journalism, including works created by CIR.

6. Defendants copied, used, abridged, and displayed CIR's valuable content without CIR's permission or authorization, and without any compensation to CIR. Defendants' products undermine and damage CIR's relationship with potential readers, consumers, and partners, and

deprive CIR of subscription, licensing, advertising, and affiliate revenue, as well as donations from readers.

7. At the same time, Defendants greatly benefit from CIR's distinct voice in the marketplace, as CIR provides a unique perspective, especially regarding investigative topics impacting diverse communities. If limited to a homogenous dataset, Defendants' large language models would be stunted in growth and power. Their success depends on content creators like CIR and other members of the news media that are unique in their style and voice.

8. Protecting these unique voices is one of the fundamental purposes of copyright law. Since the founding of the United States, the Copyright Clause of the U.S. Constitution promises to "promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." The Copyright Act similarly empowers Congress to protect works of human creativity that persons have worked hard to create, encouraging people to devote substantial effort and resources to all manner of creative enterprises by providing confidence that creators' works will be shielded from unauthorized encroachment and that creators will be properly compensated.

9. Further recognizing that emerging technologies could be used to evade statutory protections, Congress passed the Digital Millennium Copyright Act (DMCA) in 1998. The DMCA prohibits the removal of author, title, copyright, and terms of use information from protected works where there is reason to know that it would induce, enable, facilitate, or conceal a copyright infringement. Unlike copyright infringement claims, which require copyright owners to incur significant and often prohibitive registration costs as a prerequisite to enforcing their copyrights, a DMCA claim does not require registration.

10. When they populated their training sets with works of journalism, Defendants had a choice: to respect works of journalism, or not. Defendants chose the latter. They copied copyrighted works of journalism when assembling their training sets. Their LLMs memorized and at times regurgitated those works. They distributed those works and abridgements of them to each other and the public. They contributed to their users' own unlawful copying. They removed the works' copyright management information. They trained ChatGPT not to acknowledge or respect copyright. And they did this all without permission.

11. CIR brings this lawsuit seeking actual damages and Defendants' profits, or statutory damages of no less than $750 per infringed work and $2,500 per DMCA violation.

PARTIES

12. The Center for Investigative Reporting, Inc. is the nation's oldest nonprofit investigative newsroom. It is the product of a merger between Mother Jones, founded in 1976 by the esteemed author Adam Hochschild, publishing executive Richard Parker, and others; and the Center for Investigative Reporting, founded in 1977 by three esteemed investigative news reporters, Lowell Bergman, Dan Noyes, and David Weir. CIR has evolved to a diversified multi-media nonprofit organization that reaches millions of listeners, and readers producing on all three major platforms—audio, video and print—to produce investigative stories. CIR runs, *inter alia*, the brands Mother Jones, Reveal and CIR Studios.

13. Mother Jones is a reader-supported news magazine and website known for ground-breaking investigative and in-depth journalism on issues of national and global significance. Since its start in 1976, Mother Jones has won many awards for its reporting, illustration, photography, videos, and social media. Since 2001, it has been selected a finalist for the prestigious National Magazine Award for General Excellence 11 times, winning on three occasions.

14. Reveal produces investigative journalism for the Reveal national public radio show and the Reveal podcast. Its radio show is listened to by millions of public radio listeners around the country and nearly 3 million podcast listeners every month. Reveal also operates an online news site. Reveal has received countless awards for its investigatory reporting, including multiple George Foster Peabody Awards, George Polk awards, Emmy awards, Alfred I. duPont-Columbia University Awards, IRE Awards, and Edward R. Murrow Awards and has been a finalist for various other awards, including the Pulitzer Prize.

15. The Center for Investigative Reporting, Inc. is a 501(c)(3) California corporation headquartered in San Francisco, CA with offices in New York, NY, and Washington, DC.

16. Defendants are the organizations responsible for the creation, training, marketing, and sale of ChatGPT AI products.

17. Some of the Defendants consist of interrelated OpenAI entities, referred to herein collectively as the OpenAI Defendants. These include the following:

18. OpenAI Inc. is a Delaware nonprofit corporation with a principal place of business in San Francisco, CA.

19. OpenAI OpCo LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. OpenAI OpCo LLC is the sole member of OpenAI, LLC. Previously, OpenAI OpCo was known as OpenAI LP.

20. OpenAI GP, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. It is the general partner of OpenAI OpCo and controls OpenAI OpCo.

21. OpenAI, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. It owns some of the services or products operated by OpenAI.

22. OpenAI Global LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. Its members are OAI Corporation LLC and Microsoft Corporation.

23. OAI Corporation, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. Its sole member is OpenAI Holdings, LLC.

24. OpenAI Holdings, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. Its sole members are OpenAI, Inc. and Aestas Corporation.

25. Microsoft Corporation is a Washington corporation with a principal place of business and headquarters in Redmond, Washington.

26. Microsoft has described itself as being in partnership with OpenAI. In a 2023 interview, Microsoft CEO Satya Nadella said that "ChatGPT and GPT family of models … is something that we've been partnered with OpenAI deeply now for multiple years."[1]

27. This tight-knit relationship is also borne out financially. Microsoft has invested billions of dollars in OpenAI Global LLC and will own a 49% stake in the company after its investment has been repaid.

28. Microsoft also provides the data center and bespoke supercomputing infrastructure used to train ChatGPT, which it created in collaboration with, and exclusively for, the OpenAI Defendants. It also offers to the public its own AI product called Copilot that is powered by OpenAI's GPT models.

[1] Microsoft CEO Satya Nadella's Big Bet on AI, *WSJ Podcasts* (Jan. 18, 2023), https://www.wsj.com/podcasts/the-journal/microsoft-ceo-satya-nadella-big-bet-on-ai/b0636b90-08bd-4e80-9ae3-092acc47463a.

29. In a 2023 interview, Microsoft's CEO stated that, "[i]f OpenAI disappeared tomorrow," Microsoft could still "continue the innovation" alone because, among other reasons, "we have the data, we have everything."[2]

JURISDICTION AND VENUE

30. The Court has subject matter jurisdiction under 28 U.S.C. §§ 1331 and 1338(a) because this action arises under the Copyright Act of 1976, 17 U.S.C. § 101, et seq., including as amended by the Digital Millennium Copyright Act.

31. Jurisdiction over Defendants is proper because they have purposefully availed themselves of New York to conduct their business. Defendants maintain offices and employ staff in New York who, upon information and belief, were engaged in training and/or marketing of ChatGPT and/or Copilot, and thus in the removal of Plaintiff's copyright management information as discussed in this Complaint and/or the sale of products to New York residents resulting from that removal. Defendants consented to personal jurisdiction in this Court in at least *Authors Guild v. OpenAI Inc.*, 23-cv-08292. They further waived any challenge to personal jurisdiction in this District by declining to contest it in their first pre-answer motions in *The New York Times Company v. Microsoft Corporation*, 23-cv-11195, *Raw Story Media, Inc. v. OpenAI, Inc.*, No. 24-cv-01514 (OpenAI Defendants only), *The Intercept Media, Inc. v. OpenAI, Inc.*, No. 24-cv-01515, and *Daily News v. Microsoft Corporation*, No. 24-cv-03285.

32. CIR also has one of its main offices in this District in New York, NY, further demonstrating that the injuries occurred in this District.

[2] Intelligencer Staff, Satya Nadella on Hiring the Most Powerful Man in AI, *Intelligencer*, (Nov. 21, 2023), https://nymag.com/intelligencer/2023/11/on-with-kara-swisher-satya-nadella-on-hiring-sam-altman.html.

33. Venue is proper under 28 U.S.C. § 1400(a) because Defendants or their agents reside or may be found in this District.

34. Venue is also proper under 28 U.S.C. § 1391(b)(2) because a substantial part of the acts or omissions giving rise to Plaintiff's claims occurred in this District. Specifically, Defendants employ staff in New York who, upon information and belief, were engaged in the activities alleged in this Complaint.

35. Defendants consented to venue in this District in at least *Authors Guild v. OpenAI Inc.*, 23-cv-08292. They further waived any challenge to venue in this District by declining to contest it in their first pre-answer motions in *The New York Times Company v. Microsoft Corporation*, 23-cv-11195, *Raw Story Media, Inc. v. OpenAI, Inc.*, No. 24-cv-01514 (OpenAI Defendants only), *The Intercept Media, Inc. v. OpenAI, Inc.*, No. 24-cv-01515, and *Daily News v. Microsoft Corporation*, No. 24-cv-03285.

PLAINTIFF'S COPYRIGHT-PROTECTED WORKS OF JOURNALISM

36. Plaintiff owns the exclusive copyright to all works published in the Mother Jones magazine since 1978 (the "Registered Works"), and (with a few exceptions) has registered monthly and bimonthly issues of its magazines with the Copyright Office since 1978. A list of copyright registrations applicable to the Registered Works is attached as Exhibit 1.

37. Registered Works dating back to the January/February 1995 issue are available on Mother Jones' website, https://www.motherjones.com/customer-service/back-issues/back-issues-1995/. In addition, pursuant to an agreement between Plaintiff and Google, Registered Works are available online through Google Books, https://books.google.com/books/serial/ISSN:03628841?rview=1. All the Registered Works are available online on at least one of these websites.

38. Mother Jones has published its journalism online since 1993, and digitized older material at significant cost.

39. Plaintiff owns copyright to additional articles under the Mother Jones brand and the Reveal brand that are not registered with the Copyright Office. These articles are published, respectively, on the web domains motherjones.com and revealnews.org. Plaintiff owns copyright to at least 95 percent of the articles on motherjones.com (including but not limited to all the Registered Works), the rest of which are licensed and published with permission. Plaintiff owns copyright to all of the articles on revealnews.org. Plaintiff's online-only works are not registered because the Copyright Office does not currently offer an economically feasible way to register works that are published only online.

40. Plaintiff's copyright-protected works are the result of significant investments by Plaintiff through its reporters and other resources necessary to tell important stories. Making award-winning investigative journalism is harder and more expensive than ever, which is why many news outlets have recently abandoned investigative reporting teams in their newsrooms. Most investigative stories require months to report and some even take years, costing significant sums of money to tell stories that otherwise go untold.

41. The protection of CIR's intellectual property is critical to its continued ability to fund its nonprofit public interest journalism. Without control of its copyrighted content for revenue purposes, nonprofits news organizations like CIR will have even fewer journalists able to tell the ever-more dwindling number of investigative news stories that are already disappearing at an alarming rate in the today's paltry media landscape. With fewer investigative news stories told, the cost to democracy will be enormous. Indeed, the vital importance of investigative reporting to democracy is why CIR maintains two websites and a digital archive at significant cost.

42. CIR depends on its exclusive rights of reproduction, adaptation, publication, performance, and display under copyright law to resist these forces. That is why Mother Jones has registered its copyright since inception, dating back nearly fifty years. CIR has also maintained an ongoing licensing program, and implemented a terms of service that set limits on the use of its content. For instance, CIR requires third parties to provide notice and obtain permission before using CIR content or trademarks for commercial purposes, and for decades CIR has licensed its content under negotiated licensing agreements to other news media outlets.

43. In addition, newsrooms, including CIR, have long had licensing programs sharing content with one another, at a cost, to create new reporting. For instance, newsrooms often contact CIR for its decades-old archive of video footage and online articles, which CIR sells at market rate. Publishers also frequently contact CIR to license articles from the Mother Jones archives. CIR has dedicated staff to manage this service and readymade contracts to easily assist these requests.

44. Unlike countless other publishers in the industry that have contacted CIR to license material, Defendants never contacted CIR to license its work in any way. Instead, Defendants simply took what they wanted with no regard for the intellectual property rights of CIR or other publishers.

45. CIR has never given permission to any entity, including Defendants, to use its content for GenAI purposes.

DEFENDANTS' UNAUTHORIZED USE OF PLAINTIFF'S WORKS IN THEIR TRAINING SETS

46. OpenAI was formed in December 2015 as a "non-profit artificial intelligence research company" but quickly became a multi-billion-dollar for-profit business built on the exploitation of copyrighted works belonging to creators around the world, including CIR. Unlike

CIR, OpenAI shed its exclusive nonprofit status just three years after its founding and created OpenAI LP in March 2019, a for-profit company dedicated to its for-profit activities including product development and raising capital from investors.

47. Defendants' GenAI products utilize a "large language model," or "LLM." The different versions of GPT are examples of LLMs. An LLM, including those that power ChatGPT and Copilot, take text prompts as inputs and emit outputs to predict responses that are likely to follow a given the potentially billions of input examples used to train it.

48. LLMs arrive at their outputs as the result of their training on works written by humans, which are often protected by copyright. They collect these examples in training sets.

49. When assembling training sets, LLM creators, including Defendants, first identify the works they want to include. They then encode the work in computer memory as numbers called "parameters."

50. Defendants have not published the contents of the training sets used to train any version of ChatGPT, but have disclosed information about those training sets prior to GPT-4.[3] Beginning with GPT-4, Defendants have been fully secret about the training sets used to train that and later versions of ChatGPT. Plaintiff's allegations about Defendants' training sets are therefore based upon an extensive review of publicly available information regarding earlier versions of ChatGPT and consultations with a data scientist employed by Plaintiff's counsel to analyze that information and provide insights into the manner in which AI is developed and functions.

51. Microsoft has built its own AI product, called Copilot, which uses Microsoft's Prometheus technology. Prometheus combines the Bing search product with the OpenAI

[3] Plaintiff collectively refers to all versions of ChatGPT as "ChatGPT" unless a specific version is specified.

Defendants' GPT models into a component called Bing Orchestrator. When prompted, Copilot responds to user queries using Bing Orchestrator by providing AI-rewritten abridgements or regurgitations of content found on the internet.[4]

52. Earlier versions of ChatGPT (prior to GPT-4) were trained using at least the following training sets: WebText, WebText2, and sets derived from Common Crawl.

53. WebText and WebText2 were created by the OpenAI Defendants. They are collections of all outbound links on the website Reddit that received at least three "karma."[5] On Reddit, a karma indicates that users have generally approved the link. The difference between the datasets is that WebText2 involved scraping links from Reddit over a longer period of time. Thus, WebText2 is an expanded version of WebText.

54. The OpenAI Defendants have published a list of the top 1,000 web domains present in the WebText training set and their frequency. According to that list, 16,793 distinct URLs from Mother Jones's web domain appear in WebText.[6]

55. Defendants have a record, and are aware, of each URL that was included in each of their training sets.

56. Joshua C. Peterson, currently an assistant professor in the Faculty of Computing and Data Sciences at Boston University, and two computational cognitive scientists with PhDs from U.C. Berkeley, created an approximation of the WebText dataset, called OpenWebText, by also scraping outbound links from Reddit that received at least three "karma," just like the OpenAI

[4] https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing
[5] Alec Radford et al, Language Models are Unsupervised Multitask Learners, 3 https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
[6] https://github.com/openai/gpt-2/blob/master/domains.txt.

Defendants did in creating WebText.[7] They published the results online. A data scientist employed by Plaintiff's counsel then analyzed those results. OpenWebText contains 17,019 distinct URLs from motherjones.com and 415 from revealnews.org. A list of the Mother Jones works contained in OpenWebText is attached as Exhibit 2. A list of the Reveal works contained in OpenWebText is attached as Exhibit 3.

57. Upon information and belief, there are slightly different numbers of Mother Jones articles in WebText and OpenWebText at least in part because the scrapes occurred on different dates.

58. OpenAI has explained that, in developing WebText, it used sets of algorithms called Dragnet and Newspaper to extract text from websites.[8] Upon information and belief, OpenAI used these two extraction methods, rather than one method, to create redundancies in case one method experienced a bug or did not work properly in a given case. Applying two methods rather than one would lead to a training set that is more consistent in the kind of content it contains, which is desirable from a training perspective.

59. Dragnet's algorithms are designed to "separate the main article content" from other parts of the website, including "footers" and "copyright notices," and allow the extractor to make further copies only of the "main article content."[9] Dragnet is also unable to extract author and title information from the header or byline, and extracts it only if it happens to be separately contained in the main article content. Put differently, copies of news articles made by Dragnet are designed

[7] https://github.com/jcpeterson/openwebtext/blob/master/README.md.
[8] Alec Radford et al., Language Models are Unsupervised Multitask Learners, 3 https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
[9] Matt McDonnell, Benchmarking Python Content Extraction Algorithms (Jan. 29, 2015), https://moz.com/devblog/benchmarking-python-content-extraction-algorithms-dragnet-readability-goose-and-eatiht.

not to, contain author, title, copyright notices, and footers, and do not contain such information unless it happens to be contained in the main article content.

60. Like Dragnet, the Newspaper algorithms are incapable of extracting copyright notices and footers. Further, a user of Newspaper has the choice to extract or not extract author and title information. On information and belief, the OpenAI Defendants chose not to extract author and title information because they desired consistency with the Dragnet extractions, and Dragnet is typically unable to extract author and title information.

61. In applying the Dragnet and Newspaper algorithms while assembling the WebText dataset, the OpenAI Defendants removed Plaintiff's author, title, copyright notice, and terms of use information, the latter of which is contained in the footers of Plaintiff's websites.

62. Upon information and belief, the OpenAI Defendants, when using Dragnet and Newspaper, first download and save the relevant webpage before extracting data from it. This is at least because, when they use Dragnet and Newspaper, they likely anticipate a possible future need to regenerate the dataset (*e.g.*, if the dataset becomes corrupted), and it is cheaper to save a copy than it is to recrawl all the data.

63. Because, by the time of its scraping, Dragnet and Newspaper were publicly known to remove author, title, copyright notices, and footers, and given that OpenAI employs highly skilled data scientists who would know how Dragnet and Newspaper work, the OpenAI Defendants intentionally and knowingly removed this copyright management information while assembling WebText.

64. A data scientist employed by Plaintiff's counsel applied the Dragnet code to three Reveal URLs contained in OpenWebText. The results are attached as Exhibit 4. The resulting copies, whose text is substantively identical to the original (*e.g.*, identical except for the seemingly

random addition of an extra space between two words, or the exclusion of a description associated with an embedded photo), lack the author, title, copyright notice, and terms of use information with which they were conveyed to the public, except in some cases where author information happened to be contained in the main article content. The Dragnet code failed when the data scientist attempted to apply it to Mother Jones articles, further corroborating the OpenAI Defendants' need for redundancies referenced above.

65. A data scientist employed by Plaintiff's counsel also applied the Newspaper code to three Mother Jones and three Reveal URLs contained in OpenWebText. The data scientist applied the version of the code that enables the user not to extract author and title information based on the reasonable assumption that the OpenAI Defendants desired consistency with the Dragnet extractions. The results are attached as Exhibit 5. The resulting copies, whose text is substantively identical to the original, lack the author, title, copyright notice, and terms of use information with which they were conveyed to the public, except in some cases where author information happened to be contained in the main article content.

66. The absence of author, title, copyright notice, and terms of use information from the copies of Plaintiff's articles generated by applying the Dragnet and Newspaper codes—codes OpenAI has admitted to have intentionally used when assembling WebText—further corroborates that the OpenAI Defendants intentionally removed author, title, copyright notice, and terms of use information from Plaintiff's copyright-protected news articles.

67. Upon information and belief, the OpenAI Defendants have continued to use the same or similar Dragnet and Newspaper text extraction methods when creating training sets for every version of ChatGPT since GPT-2. This is at least because the OpenAI Defendants have admitted to using these methods for GPT-2 and have neither publicly disclaimed their use for later

version of ChatGPT nor publicly claimed to have used any other text extraction methods for those later versions.

68. The other repository the OpenAI Defendants have admitted to using, Common Crawl, is a scrape of most of the internet created by a third party.

69. To train GPT-2, OpenAI downloaded Common Crawl data from the third party's website and filtered it to include only certain works, such as those written in English.[10]

70. Google has published instructions on how to replicate a dataset called C4, a monthly snapshot of filtered Common Crawl data that Google used to train its own AI models. Upon information and belief, based on the similarity of Defendants' and Google's goals in training AI models, C4 is substantially similar to the filtered versions of Common Crawl used to train ChatGPT. The Allen Institute for AI, a nonprofit research institute launched by Microsoft cofounder Paul Allen, followed Google's instructions and published its recreation of C4 online.[11]

71. A data scientist employed by Plaintiff's counsel analyzed this recreation. It contains 26,178 URLs originating from motherjones.com. The vast majority of these URLs contain Plaintiff's copyright-protected news articles. None contain terms of use information. None contain copyright notice information as to Plaintiff's copyright-protected news articles. The majority also lack author and title information. In some cases, the articles are substantively identical, while in others a small number of paragraphs are omitted.

72. This recreation also contains 451 articles originating from revealnews.org. The vast majority of these URLs contain Plaintiff's copyright-protected news articles. None of the news articles contains copyright notice or terms of use information. The majority also lack author

[10] Tom B. Brown et al, Language Models are Few-Shot Learners, 14 (July 22, 2020), https://arxiv.org/pdf/2005.14165.
[11] https://huggingface.co/datasets/allenai/c4.

and title information. In some cases, the articles are substantively identical, while in others a small number of paragraphs is omitted.

73. As a representative sample, the text of three Mother Jones and three Reveal articles as they appear in the C4 set is attached as Exhibit 6. None of these articles contains the author, title, copyright notice, or terms of use information with which they were conveyed to the public.

74. Plaintiff has not licensed or otherwise permitted Defendants to include any of its works in their training sets.

75. Downloading tens of thousands of Plaintiff's articles without permission infringes Plaintiff's copyrights, more specifically, the right to control reproductions of copyright-protected works.

DEFENDANTS' UNAUTHORIZED REGURGITATION OF COPYRIGHT-PROTECTED WORKS OF JOURNALISM

76. ChatGPT and Copilot provide responses to questions or other prompts. Their ability to provide these responses is the key value proposition of Defendants' products, which they are able to sell to their customers for enormous sums of money, soon likely to be in the billions of dollars.

77. To train ChatGPT, the OpenAI Defendants retain users' chat histories with ChatGPT unless the user takes the affirmative step of disabling that feature.[12] Thus, the OpenAI Defendants possess a repository of every regurgitation of Plaintiff's works apart from those whose storage users have affirmatively disabled.

78. At least some of the time, ChatGPT and Copilot provide or have provided responses to users that regurgitate verbatim or nearly verbatim copyright-protected works of journalism

[12] New ways to manage your data in ChatGPT (Apr. 25, 2023), https://openai.com/index/new-ways-to-manage-your-data-in-chatgpt/.

without providing author, title, copyright, or terms of use information contained in those works. Examples of such regurgitations are included in Exhibit J to the Complaint in *Daily News, LP v. Microsoft Corporation*, No. 24-cv-03285 (S.D.N.Y. Apr. 30, 2024).

79. At least some of the time, ChatGPT and Copilot provide or have provided responses to users that mimic significant amounts of material from copyright-protected works of journalism without providing any author, title, copyright, or terms of use information contained in those works. For example, if a user asks ChatGPT or Copilot about a current event or the results of a work of investigative journalism, ChatGPT or Copilot will provide responses that mimic copyright-protected works of journalism that covered those events, not responses that are based on any journalism efforts by Defendants.

80. At least some of the time, ChatGPT memorizes and regurgitates material. The OpenAI Defendants have publicly admitted their knowledge of this fact.[13] The OpenAI Defendants have also effectively admitted that regurgitation of copyrighted works is infringement: when Plaintiff attempted to obtain the same regurgitations set forth in the *Daily News* case using the same methodology, Plaintiff received in one instance a message stating, "I'm sorry, but I can't generate the original ending for the article or any copyrighted content." Thus, upon information and belief, the OpenAI Defendants have recently changed ChatGPT to reduce regurgitations for copyright reasons.

81. Nonetheless, ChatGPT has produced regurgitations of Plaintiff's copyright-protected works. Examples of three such regurgitations, along with the prompts that generated them, are attached as Exhibit 7.

[13] OpenAI and journalism (Jan. 8, 2024), https://openai.com/index/openai-and-journalism/.

82. Such memorization and regurgitation constitute unauthorized copies or derivative works of the Plaintiff's work. Defendants directly engage in the unauthorized public display of CIR's articles as part of generative output provided by their products built on the GPT models.

DEFENDANTS' UNLAWFUL ABRIDGEMENT OF PLAINTIFFS' WORKS

83. In response to prompts by users, ChatGPT and Copilot provide highly detailed abridgements of copyright-protected news articles, including articles published by Plaintiff.

84. When earlier versions of ChatGPT (up to and including ChatGPT 3.5-turbo) abridge a copyright-protected news article in response to a user prompt, they draw from their training on the article. During training, the patterns of all content, including copyright-protected news articles, are imprinted onto the model. That imprint allows the model to abridge the article.

85. When Copilot and later versions of ChatGPT abridge a copyright-protected news article in response to a user prompt, they find the previously downloaded article inside a database called a search index using a method called synthetic searching. Upon information and belief, they make another copy of the article in the memory of their computing system and use their LLM or other programming to generate an abridgement by applying the model or other programming to the text of the article.

86. Plaintiff's articles are not merely collections of facts. Rather, they reflect the originality of their authors in selecting, arranging, and presenting facts to tell compelling stories. They also reflect the authors' analysis and interpretation of events, structuring of materials, marshaling of facts, and the emphasis given to certain aspects.

87. An ordinary observer of a ChatGPT or Copilot abridgement of one of Plaintiff's articles would conclude that the abridgements were derived from the articles being abridged, at

least because ChatGPT and Copilot expressly link to the article they abridge and explain that they are searching Plaintiffs' website in the course of generating a response.

88. In response to prompts seeking an abridgement of an article, ChatGPT and Copilot will typically provide a general abridgement of such an article, on the order of a few paragraphs. In some instances, the initial response will summarize the article in substantial detail. Further, when prompted by the user to provide more information about one or more aspects of that abridgement, ChatGPT or Copilot will provide additional details, often in the format of a bulleted list of main points. If prompted by the user to provide more information on one of more of those points, Chat GPT or Copilot will provide additional details. In some instances, however, ChatGPT or Copilot will provide a bulleted list of main points in response to an initial prompt seeking an abridgement.

89. A ChatGPT or Copilot user is capable of obtaining a substantial abridgement of a copyright protected news article through such series of prompts, and in some instances, further prompts designed to elicit further summary are even suggested by Copilot or ChatGPT itself. As a representative sample, a series of abridgements by ChatGPT and Copilot is attached as Exhibit 8.

90. These abridgements lack copyright notice or terms of use information conveyed in connection with the work, and sometimes lack author information.

91. Thus, upon information and belief, abridgements from earlier versions of ChatGPT lack copyright notice, terms of use, and typically author information because Defendants intentionally removed that information from the ChatGPT training sets.

92. Further, the abridgements from Copilot and later versions of ChatGPT lack copyright notice, terms of use, and typically author information. Upon information and belief,

this is because Defendants intentionally removed them either when initially storing them in computer memory or when generating the synthetic search results.

93. Defendants' abridgements, rewritten from copyright-protected news articles, harm the market for those articles by reducing the incentives for users to go to the original source, thus reducing Plaintiff's subscription, licensing, advertising, and affiliate revenue. This allows Defendants to monetize copyright owners' content at the expense of copyright owners who created the works ChatGPT has abridged.

94. Defendants' abridgements do not add anything new to, or further any purpose or character different from, that of Plaintiff's articles. They simply take the text of the articles and rewrite them into abridgements, including, when prompted, into detailed abridgements of the entire articles. Those abridgements often serve as a substitute for the original articles even when they are not complete, as evidenced by a study showing that only 51% of consumers read the entire text of a typical news article.[14]

95. Defendants' abridgements of Plaintiff's articles violates Plaintiff's copyrights.

DEFENDANTS' INTENTIONAL REMOVAL OF COPYRIGHT MANAGEMENT INFORMATION FROM PLAINTIFF'S WORKS IN THEIR TRAINING SETS

96. ChatGPT and Copilot do not have any independent knowledge of the information provided in their responses. Rather, to service Defendants' paying customers, ChatGPT and Copilot instead repackage, among other material, the copyrighted journalism work product developed by Plaintiff, and others, at often considerable expense.

[14] *See* Sharing on Social Media Makes Us Overconfident in Our Knowledge, UT News (Aug. 30, 2022), https://news.utexas.edu/2022/08/30/sharing-on-social-media-makes-us-overconfident-in-our-knowledge/#:~:text=Recent%20data%20from%20the%20Reuters,headline%20or%20a%20few%20lines.

97. When providing responses, ChatGPT and Copilot give the impression that they are an all-knowing, "intelligent" source of the information being provided, when in reality, the responses are frequently based on copyrighted works of journalism that ChatGPT and Copilot simply mimic.

98. If ChatGPT and Copilot were trained on works of journalism that included the original author, title, copyright notice, and terms of use information, they would have learned to communicate that information when providing responses to users unless Defendants trained them otherwise.

99. Based on the information described above, thousands of Plaintiff's copyrighted works were included in Defendants' training sets without the author, title, copyright notice, and terms of use information that Plaintiff conveyed in publishing them.

100. Based on the information above, including the OpenAI Defendants' admission to using the Dragnet and Newspaper extraction methods, which remove author, title, copyright notice, and terms of use information from copyright-protected news articles published online, the OpenAI Defendants intentionally removed author, title, copyright notice, and terms of use information from Plaintiff's copyrighted works in creating ChatGPT training sets.

DEFENDANTS' COLLABORATION IN INFRINGING PLAINTIFF'S COPYRIGHT, UNLAWFULLY REMOVING COPYRIGHT MANAGEMENT INFORMATION, AND UNLAWFULLY DISTRIBUTING PLAINTIFF'S WORKS WITH COPYRIGHT MANAGEMENT INFORMATION REMOVED

101. Based on the publicly available information described above, including the admission from Microsoft's CEO that "we have the data, we have everything," Defendant Microsoft has created, without Plaintiff's permission, its own copies of Plaintiff's copyright-protected works of journalism, including but not limited to the Registered Works.

102. Based on the publicly available information described above, including information showing that Defendant Microsoft created and hosted the data centers used to develop ChatGPT and information regarding Microsoft's own Copilot, Defendant Microsoft intentionally removed author, title, copyright notice, and terms of use information from Plaintiff's copyrighted works in creating ChatGPT and Copilot training sets.

103. Based on publicly available information regarding the relationship between Defendant Microsoft and the OpenAI Defendants, and Defendant Microsoft's provision of database and computing resources to the OpenAI Defendants, Defendant Microsoft has shared copies of Plaintiff's works from which author, title, copyright notice, and terms of use information had been removed, with the OpenAI Defendants as part of Defendants' efforts to develop ChatGPT and Copilot.

104. Based on publicly available information regarding the working relationship between Defendant Microsoft and the OpenAI Defendants, including the creation of training sets by the OpenAI Defendants such as WebText and WebText2, the OpenAI Defendants have shared copies of Plaintiff's works from which author, title, copyright notice, and terms of use information had been removed, with Defendant Microsoft as part of Defendants' efforts to develop ChatGPT and Copilot.

DEFENDANTS' ACTUAL AND CONSTRUCTIVE KNOWLEDGE OF THEIR VIOLATIONS

105. The OpenAI Defendants have acknowledged that use of copyright-protected works to train ChatGPT requires a license to that content. and, in some instances. Recognizing that obligation, the OpenAI Defendants have entered into agreements with large copyright owners such as Associated Press, the Atlantic, Axel Springer, Dotdash Meredith, Financial Times, News Corp,

and Vox Media to obtain licenses to include those entities' copyright-protected works in Defendants' LLM training data.

106. The OpenAI Defendants are also in licensing talks with other copyright owners in the news industry, but have offered no compensation to Plaintiff.

107. In a May 29, 2024 interview, OpenAI's Chief of Intellectual Property and Content, Tom Rubin, stated that these deals focus on "the display of news content and use of the tools and tech," and are thus "largely not" about training.[15] This admission, while qualified, confirms that these deals involve training, at least in part.

108. The OpenAI Defendants created tools in late 2023 to allow copyright owners to block their work from being incorporated into training sets. This further corroborates that the OpenAI Defendants had reason to know that use of copyrighted material in their training sets is copyright infringement.

109. The creation of such tools also corroborates that the OpenAI Defendants had reason to know that their copyright infringement is enabled, facilitated, and concealed by their removal of author, title, copyright, and terms of use information from their training sets.

110. Defendants had reason to know that the removal of author, title, copyright notice, and terms of use information from copyright-protected works and their use in training ChatGPT would result in ChatGPT providing responses to ChatGPT users that abridged or regurgitated material verbatim from copyrighted works in creating responses to users, without revealing that those works were subject to Plaintiff's copyrights. This is at least because Defendants were aware that ChatGPT responses are the product of its training sets and that ChatGPT generally would not

[15] Charlotte Tobitt, OpenAI content boss: 'Incumbent' on us to help small publishers, not just the giants, *PressGazette* (May 30, 2024), https://pressgazette.co.uk/platforms/openai-tom-rubin-publishers-news/.

know any author, title, copyright notice, and terms of use information that was not included in training sets.

111. Upon information and belief, Defendants had reason to know that the removal of author, copyright notice, and terms of use information from copyright-protected works used in synthetic searching would result in ChatGPT or Copilot providing responses to ChatGPT users that abridged or regurgitated material verbatim from copyrighted works in creating responses to users, without revealing that those works were subject to Plaintiff's copyrights. This is at least because Defendants were aware that Copilot's and later versions of ChatGPT's responses to prompts are the product of the articles encoded in their computer memory, from which, upon information and belief, Defendants removed author, copyright notice, and terms of use information.

112. Defendants had reason to know that users of ChatGPT would further distribute the results of ChatGPT responses. This is at least because Defendants promote ChatGPT as a tool that can be used by a user to generate content for a further audience.

113. Defendants had reason to know that users of ChatGPT would be less likely to distribute ChatGPT responses if they were made aware of the author, title, copyright notice, and terms of use information applicable to the material used to generate those responses. This is at least because Defendants were aware that at least some likely users of ChatGPT respect the copyrights of others or fear liability for copyright infringement.

114. Defendants had reason to know that ChatGPT would be less popular and would generate less revenue if users believed that ChatGPT responses violated third-party copyrights or if users were otherwise concerned about further distributing ChatGPT responses. This is at least because Defendants were aware that Defendants derive revenue from user subscriptions, that at

least some likely users of ChatGPT respect the copyrights of others or fear liability for copyright infringement, and that such users would not pay to use a product that might result in copyright liability or did not respect the copyrights of others.

115. If a commercial user of Defendants' ChatGPT and Copilot products is sued for copyright infringement, Defendants have committed to paying the user's costs in defending against the infringement claim, and to indemnifying the user for an adverse judgment or settlement. These commitments apply only if the user uses the product as advertised. In particular, Microsoft's "Copilot Copyright Commitment" applies only if the user "used the guardrails and content filters we have built into our products,"[16] and OpenAI's "Copyright Shield" does not apply if the user "disabled, ignored, or did not use any relevant citation, filtering or safety features or restrictions provided by OpenAI."[17] Thus, Defendants know or have reason to know that ChatGPT and Copilot users are capable of infringing and likely to infringe copyright even when used according to terms specified by Defendants.

Count I – Violation of 17 U.S.C. § 501 by OpenAI Defendants

116. The above paragraphs are incorporated by reference into this Count.

117. Plaintiff owns the exclusive rights to the Registered Works under 17 U.S.C. § 106.

118. The OpenAI Defendants infringed Plaintiff's exclusive rights in the Registered Works by downloading those works from the internet.

119. The OpenAI Defendants infringed Plaintiff's exclusive rights in the Registered Works by encoding the Registered Works in computer memory.

[16] https://www.microsoft.com/en-us/licensing/news/microsoft-copilot-copyright-commitment.
[17] https://openai.com/policies/service-terms/.

120. Upon information and belief, the OpenAI Defendants infringed Plaintiff's exclusive rights in the Registered Works by regurgitating those works verbatim or nearly verbatim in response to prompts by ChatGPT users.

121. Upon information and belief, the OpenAI Defendants infringed Plaintiff's exclusive rights in the Registered Works by producing significant amounts of material from those works in response to prompts by ChatGPT users.

122. The OpenAI Defendants infringed Plaintiff's exclusive rights in the Registered Works by abridging those works in response to prompts by ChatGPT users.

123. Upon information and belief, the OpenAI Defendants' infringements were willful and with full knowledge of Plaintiff's rights in the Registered Works.

Count II: Violation of 17 U.S.C. § 501 by Defendant Microsoft

124. The above paragraphs are incorporated by reference into this Count.

125. Plaintiff owns the exclusive rights to the Registered Works under 17 U.S.C. § 106.

126. Upon information and belief, Defendant Microsoft infringed Plaintiff's exclusive rights in the Registered Works by downloading those works from the internet.

127. Upon information and belief, Defendant Microsoft infringed Plaintiff's exclusive rights in the Registered Works by encoding the Registered Works in computer memory.

128. Upon information and belief, Defendant Microsoft infringed Plaintiff's exclusive rights in the Registered Works by regurgitating those works verbatim or nearly verbatim in response to prompts by Copilot users.

129. Upon information and belief, Defendant Microsoft infringed Plaintiff's exclusive rights in the Registered Works by producing significant amounts of material from those works in response to prompts by ChatGPT users.

130. Defendant Microsoft infringed Plaintiff's exclusive rights in the Registered Works by abridging those works in response to prompts by Copilot users.

131. Upon information and belief, Defendant Microsoft's infringements were willful and with full knowledge of Plaintiff's rights in the Registered Works.

Count III: Contributory Copyright Infringement by All Defendants

132. The above paragraphs are incorporated by reference into this Count.

133. In the alternative, to the extent a user may be liable as a direct infringer based on output of ChatGPT and/or Copilot, Defendants materially contributed to and directly assisted with the direct infringement by those users by jointly developing LLMs capable of distributing unlicensed copies and abridgements of the Registered Works, building and training LLMs using the Registered Works, and deciding what content is emitted by their products through the process of training them and developing them to conduct synthetic searching.

134. Defendants knew or had reason to know of the direct infringement by their users because Defendants undertake extensive efforts in developing, testing, or troubleshooting their models, (as to the OpenAI Defendants) have admitted that their products regurgitate material in response to user prompts, and have agreed to defend and indemnify certain of their users for copyright violations only when the users are using the products according to terms specified by Defendants.

Count IV – Violation of 17 U.S.C. § 1202(b)(1) by OpenAI Defendants

135. The above paragraphs are incorporated by reference into this Count.

136. Plaintiff is the owner of copyrighted works of journalism that contain author, title, copyright notice, and terms of use information.

137. Upon information and belief, the OpenAI Defendants created copies of Plaintiff's works of journalism with author information removed and included them in training sets used to train ChatGPT.

138. Upon information and belief, the OpenAI Defendants created copies of Plaintiff's works of journalism with title information removed and included them in training sets used to train ChatGPT.

139. Upon information and belief, the OpenAI Defendants created copies of Plaintiff's works of journalism with copyright notice information removed and included them in training sets used to train ChatGPT.

140. Upon information and belief, the OpenAI Defendants created copies of Plaintiff's works of journalism with terms of use information removed and included them in training sets used to train ChatGPT.

141. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would induce ChatGPT to provide responses to users that incorporated material from Plaintiff's copyright-protected works or regurgitated copyright-protected works verbatim or nearly verbatim.

142. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would induce ChatGPT users to distribute or publish ChatGPT responses that utilized Plaintiff's copyright-protected works of journalism that such users would not have distributed or published if they were aware of the author, title, copyright, or terms of use information.

143. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would enable copyright infringement by ChatGPT and ChatGPT users.

144. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would facilitate copyright infringement by ChatGPT and ChatGPT users.

145. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would conceal copyright infringement by Defendants, ChatGPT, and ChatGPT users.

Count V – Violation of 17 U.S.C. § 1202(b)(3) by OpenAI Defendants

146. The above paragraphs are incorporated by reference into this Count.

147. Upon information and belief, the OpenAI Defendants shared copies of Plaintiff's works without author, title, copyright notice, and terms of use information with Defendant Microsoft in connection with the development of ChatGPT and Copilot.

Count VI – Violation of 17 U.S.C. § 1202(b)(1) by Defendant Microsoft

148. The above paragraphs are incorporated by reference into this Count.

149. Upon information and belief, Defendant Microsoft created copies of Plaintiff's works of journalism with author information removed and included them in training sets used to train ChatGPT and Bing AI products.

150. Upon information and belief, Defendant Microsoft created copies of Plaintiff's works of journalism with title information removed and included them in training sets used to train ChatGPT and Bing AI products.

151. Upon information and belief, Defendant Microsoft created copies of Plaintiff's works of journalism with copyright notice information removed and included them in training sets used to train ChatGPT and Bing AI products.

152. Upon information and belief, Defendant Microsoft created copies of Plaintiff's works of journalism with terms of use information removed and included them in training sets used to train ChatGPT and Bing AI products.

153. Defendant Microsoft had reason to know that inclusion in training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would induce ChatGPT and Bing AI products to provide responses to users that incorporated material from Plaintiff's copyright-protected works or regurgitated copyright-protected works verbatim or nearly verbatim.

154. Defendant Microsoft had reason to know that inclusion in training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would induce ChatGPT and Bing AI product users to distribute or publish responses that utilized Plaintiff's copyright-protected works of journalism that such users would not have distributed or published if they were aware of the author, title, copyright notice, or terms of use information.

155. Defendant Microsoft had reason to know that inclusion in training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would enable copyright infringement by ChatGPT, Bing AI, and ChatGPT and Bing AI users.

156. Defendant Microsoft had reason to know that inclusion in training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would facilitate copyright infringement by ChatGPT, Bing, AI, and ChatGPT and Bing AI users.

157. Defendant Microsoft had reason to know that inclusion in training sets of Plaintiff's works of journalism without author, title, copyright notice, and terms of use information would conceal copyright infringement by Defendants, ChatGPT, Bing AI, and ChatGPT and Bing AI users.

Count VII – Violation of 17 U.S.C. § 1202(b)(3) by Defendant Microsoft

158. The above paragraphs are incorporated by reference into this Count.

159. Upon information and belief, Defendant Microsoft shared copies of Plaintiff's works without author, title, copyright notice, and terms of use information with the OpenAI Defendants in connection with the development of ChatGPT and Copilot.

PRAYER FOR RELIEF

Plaintiff seeks the following relief:

(i) Either statutory damages or the total of Plaintiff's damages and Defendants' profits, to be elected by Plaintiff;

(ii) An injunction requiring Defendants to remove all copies of the Registered Works from their training sets and any other repositories;

(iii) An injunction requiring Defendants to remove all copies of Plaintiff's copyrighted works from which author, title, copyright, or terms of use information was removed from their training sets and any other repositories;

(iv) Attorney fees and costs.

JURY DEMAND

Plaintiff demands a jury trial.

RESPECTFULLY SUBMITTED,

/s/ Stephen Stich Match

Jonathan Loevy*
Michael Kanovitz*
Lauren Carbajal*
Stephen Stich Match (No. 5567854)
Matthew Topic*
Thomas Kayes*
Steven Art*

LOEVY & LOEVY
311 North Aberdeen, 3rd Floor
Chicago, IL 60607
312-243-5900 (p)
312-243-5902 (f)
jon@loevy.com
mike@loevy.com
carbajal@loevy.com
match@loevy.com
matt@loevy.com

*pro hac vice forthcoming

June 27, 2024