UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

THE INTERCEPT MEDIA, 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.	No. 1:24-cv-01515-SHS **SECOND AMENDED COMPLAINT** **JURY TRIAL DEMANDED**

1. Plaintiff The Intercept Media, Inc., through its attorneys Loevy & Loevy, for its Complaint against the OpenAI Defendants and Defendant Microsoft, alleges the following:

2. The Copyright Clause of the U.S. Constitution empowers Congress to protect works of human creativity. The resulting legal protections encourage people to devote effort and resources to all manner of creative enterprises by providing confidence that creators' works will be shielded from unauthorized encroachment.

3. In recognition that emerging technologies could be used to evade statutory protections, Congress passed the Digital Millennium Copyright Act 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.

4. Generative artificial intelligence (AI) systems and large language models (LLMs) are trained using works created by humans. AI systems and LLMs ingest massive amounts of human creativity and use it to mimic how humans write and speak. These training sets have included hundreds of thousands, if not millions, of works of journalism.

5. Defendants are the companies responsible for the creation and development of the highly lucrative ChatGPT and Copilot AI products. According to the award-winning website Copyleaks, nearly 60% of the responses provided by Defendants' GPT-3.5 product in a study conducted by Copyleaks contained some form of plagiarized content, and over 45% contained text that was identical to pre-existing content.

6. When they populated their training sets with works of journalism, Defendants had a choice: they could train ChatGPT and Copilot using works of journalism with the copyright management information protected by the DMCA intact, or they could strip it away. Defendants chose the latter, and in the process, trained ChatGPT and Copilot not to acknowledge or respect copyright, not to notify ChatGPT and Copilot users when the responses they received were protected by journalists' copyrights, and not to provide attribution when using the works of human journalists.

7. Plaintiff The Intercept Media, Inc., is a news organization, and brings this lawsuit seeking actual damages and Defendants' profits, or statutory damages of no less than $2500 per violation.[1]

[1] The Intercept acknowledges that the Court has dismissed Counts II, III, and IV. *See* ECF No. 127. The Intercept does not seek to revive those claims via this complaint, but continues to plead those claims, and all associated allegations, to preserve its appellate rights.

PARTIES

8. The Intercept is an award-winning news organization dedicated to holding the powerful accountable through fearless, adversarial journalism. Its in-depth investigations and unflinching analysis focus on politics, war, surveillance, corruption, the environment, technology, criminal justice, the media, and other issues. The Intercept has been recognized for its reporting on the U.S. drone program, criminal behavior in a major metropolitan police department, and toxic Teflon chemicals, among other work.

9. The Intercept is a Delaware, non-stock, nonprofit organization. Its headquarters are located in New York, NY.

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

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

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

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

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

20. 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."[2]

21. 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.

22. Microsoft 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.

[2] 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.

23. 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."[3]

24. Upon information and belief based on the relationship between Defendants and the statements discussed above, Microsoft hosts ChatGPT training sets and provides access to those training sets to one or more of the OpenAI Defendants, and some of those training sets were created by the OpenAI Defendants and provided to Microsoft.

JURISDICTION AND VENUE

25. 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., as amended by the Digital Millennium Copyright Act.

26. 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 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 case by not raising any such challenge in their Motions to Dismiss.

27. Because Plaintiff's principal place of business is in this District, Defendants could reasonably foresee that the injuries alleged in this Complaint would occur in this District.

[3] 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.

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

29. 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, on information and belief, were engaged in the activities alleged in this Complaint.

30. Defendants consented to venue in this Court in at least *Authors Guild v. OpenAI Inc.*, 23-cv-08292. They further waived any challenge to venue in this case by not raising any such challenge in their Motions to Dismiss.

PLAINTIFF'S COPYRIGHT-PROTECTED WORKS OF JOURNALISM

31. Plaintiff's copyrighted works of journalism are published on Plaintiff's website, theintercept.com, and are conveyed to the public with author, title, copyright, and terms of use information.

32. Plaintiff owns copyrights to all the articles listed in Exhibit 1.

33. Plaintiff's copyright-protected works are the result of significant investments by Plaintiff in the human and other resources necessary to report on the news.

DEFENDANTS' INCLUSION OF PLAINTIFF'S WORKS IN THEIR TRAINING SETS AND REMOVAL OF COPYRIGHT MANAGEMENT INFORMATION

34. Defendants' generative AI 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 given the potentially billions of input examples used to train it.

35. 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.

36. 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."

37. 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.[4] 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.

38. Microsoft has built its own AI product, called Copilot, which uses Microsoft's Prometheus technology. Prometheus combines the Bing search product with the OpenAI 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.[5]

39. 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.

40. 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."[6] On

[4] Plaintiff collectively refers to all versions of ChatGPT as "ChatGPT" unless a specific version is specified.
[5] https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing
[6] 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

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.

41. 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, 6,484 distinct URLs from Plaintiff's web domain were included in WebText.[7]

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

43. 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 Defendants did in creating WebText.[8] They published the results online. A data scientist employed by Plaintiff's counsel then analyzed those results. OpenWebText contains 5,026 distinct URLs from Plaintiff's web domain. A list of these URLs and a description of the analysis is attached as Exhibit 2.

44. Upon information and belief, there are different numbers of Plaintiff's articles in WebText and OpenWebText at least in part because the scrapes occurred on different dates.

45. OpenAI has explained that, in developing WebText, it used sets of algorithms called Dragnet and Newspaper to extract text from websites.[9] Upon information and belief,

[7] https://github.com/openai/gpt-2/blob/master/domains.txt.
[8] https://github.com/jcpeterson/openwebtext/blob/master/README.md.
[9] 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.

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.

46. 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."[10] Dragnet is also unable to extract author and title information. Put differently, copies of news articles made by Dragnet necessarily do not contain author, title, copyright notices, and footers.

47. 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 unable to extract author and title information.

48. 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.

49. 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

[10] Matt McDonnell, Benchmarking Python Content Extraction Algorithms (Jan. 29, 2015), https://moz.com/devblog/benchmarking-python-content-extraction-algorithms-dragnet-readability-goose-and-eatiht.

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.

50. 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.

51. A data scientist employed by Plaintiff's counsel applied the Dragnet code to three of Plaintiff's URLs contained in OpenWebText. The results are attached as Exhibit 3. 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.

52. A data scientist employed by Plaintiff's counsel also applied the Newspaper code to three of Plaintiff's 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 4. 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.

53. 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.

54. 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.

55. Common Crawl is a data set that consists of a scrape of most of the internet created by a non-profit research institute, also called Common Crawl. ChatGPT was trained on a version of Common Crawl, in addition to the WebText and WebText2 training sets.

56. 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.[11]

57. 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.[12]

58. A data scientist employed by Plaintiff's counsel analyzed this recreation. It contains 2,753 distinct URLs from Plaintiff's web domain. The vast majority of these URLs

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

contain The Intercept's copyright-protected news articles. None of the news articles contains copyright notice or terms of use information. The vast majority lack both author and title information. In some cases, the articles are reproduced entirely verbatim, while in others a small number of paragraphs is omitted.

59. As a representative sample, the text of three of the articles as they appear in the C4 set is attached as Exhibit 5. None of these articles contains the author, title, copyright notice, or terms of use information with which it was conveyed to the public. In each case, the article's text in C4 is substantively identical to the text from Plaintiff's website.

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

61. Defendants' actions in downloading thousands of Plaintiff's articles without permission infringes Plaintiff's copyright, more specifically, the right to control reproductions of copyright-protected works.

DEFENDANTS' REGURGITATION, MIMICKING, AND ABRIDGEMENT OF COPYRIGHT-PROTECTED WORKS OF JOURNALISM

62. 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.

63. To train ChatGPT, the OpenAI Defendants retain users' chat histories with ChatGPT unless the user takes the affirmative step of disabling that feature.[13] Thus, upon

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

information and belief, the OpenAI Defendants possess a repository of regurgitations of Plaintiff's works apart from those whose storage users have affirmatively disabled.

64. 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 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).

65. 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.

66. At least some of the time, ChatGPT memorizes and regurgitates material.[14] The OpenAI Defendants have publicly admitted their knowledge of this fact. 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.

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

67. 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 6.

68. At least some of the time, ChatGPT provides or has provided responses to users that abridge copyright-protected works of journalism without providing any author, title, or copyright notice information conveyed in connection with those works. Examples of such abridgements are included in Exhibit 11 to the First Amended Complaint in *The Center for Investigative Reporting, Inc. v. OpenAI, Inc.*, No. 24-cv-4872 (S.D.N.Y. Sept. 9, 2024), ECF No. 88-14. For instance, in the fourth example, the ChatGPT abridgement reproduces, verbatim, nine consecutive paragraphs of text (minus one sentence) from the original article, which can be found at https://www.motherjones.com/politics/2024/01/100-bill-crime-corruption-treasury-taxevasion/.

69. 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.

70. When Copilot and later versions of ChatGPT abridge a copyright-protected news article in response to a user prompt, they may find the previously downloaded article inside a database called a search index using a method called synthetic searching or retrieval-augmented generation ("RAG"). 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.

71. Plaintiffs' 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.

72. An ordinary observer of a ChatGPT abridgement of copyright-protected news articles would conclude that the abridgements were derived from the articles being abridged.

73. In response to prompts seeking an abridgement of an article, ChatGPT or 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, ChatGPT 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.

74. 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 ChatGPT or Copilot itself. See Exhibit 11 to the First Amended Complaint in *The Center for Investigative Reporting, Inc. v. OpenAI, Inc*., No. 24- cv-4872 (S.D.N.Y. Sept. 9, 2024), ECF No. 88-14. These abridgements lack copyright notice information conveyed in connection with the work, and sometimes lack author information. They sometimes link to webpages that do not belong to the news organization that owns the article and that do not contain the news organization's copyright management information.

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

76. Further, the abridgements from Copilot and later versions of ChatGPT lack copyright notice 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 results by employing RAG.

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

77. 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 that was developed and created by Plaintiff, and others, at often considerable their expense.

78. 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.

79. 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.

80. 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.

81. 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

82. 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.

83. 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.

84. 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.

85. 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

86. The OpenAI Defendants have acknowledged that use of copyright-protected works to train ChatGPT requires a license to that content. 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.

87. The OpenAI Defendants are also in licensing talks with other copyright owners in the news industry, but offered no compensation to Plaintiff prior to taking Plaintiff's content, removing its copyright management information, and using it train their AI bots.

88. 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.

[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/.

89. 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 without permission or license is copyright infringement.

90. 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.

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

92. Upon information and belief, Defendants had reason to know that the removal of author, title, copyright notice, and terms of use information from copyright-protected works used in synthetic searching would result in their products' providing response to their 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, title, and copyright notice information.

93. 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.

94. 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.

95. 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.

96. 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

[16] https://www.microsoft.com/en-us/licensing/news/microsoft-copilot-copyright-commitment.

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.

97. Defendants intend in part for ChatGPT and Copilot to replicate how ordinary English speakers express themselves. When ordinary English speakers are not conveying copyright-protected works, they do not include copyright management information—especially copyright notices and terms of use. Had ChatGPT and Copilot been trained on Plaintiff's and others' copyright-protected works that include this copyright management information, they would have falsely learned that ordinary English speakers convey copyright management information in situations when they do not. To avoid this result, Defendants had a choice between removing the copyright management information at the outset or retraining ChatGPT and Copilot not to emit the copyright management information after they had incorrectly learned how English speakers normally express themselves. Upon information and belief, Defendants chose to remove the copyright management information at the outset, at least because doing so involves fewer computational resources and therefore is far less expensive than retraining. Thus, because Defendants infringed Plaintiff's copyright by using Plaintiff's articles to train ChatGPT and Copilot, Defendants removed Plaintiff's copyright management information from its copyright-protected articles knowing, or having reasonable grounds to know, that doing so would facilitate their own training-based infringing conduct.

98. Defendants' unauthorized copying of Plaintiffs' works into Defendants' training sets and search indices is facilitated by the removal of author, title, and copyright notice information because copying less data requires fewer computational and storage resources.

[17] https://openai.com/policies/service-terms/.

DEFENDANTS' CONTINUING VIOLATIONS

99. Upon information and belief, Defendants have continued to unlawfully copy, regurgitate, abridge, and remove author, title, and copyright notice information from Plaintiff's copyright-protected works up to the present date, or at least until Plaintiff implemented the exclusion protocols on January 18, 2024, that the OpenAI Defendants released in August 2023 allowing websites to opt out of OpenAI's web crawling.

100. ChatGPT and Copilot have emitted significant material from copyright-protected works of journalism that significantly postdate the WebText and WebText2 training sets. Examples are contained in Exhibit 11 to the First Amended Complaint in *The Center for Investigative Reporting, Inc. v. OpenAI, Inc*., No. 24-cv-4872 (S.D.N.Y. Sept. 9, 2024), ECF No. 88-14. ChatGPT and Copilot could not have produced this material without Defendants' copying the original articles and storing them in computer memory, including in training sets created by ChatGPT 3.5-turbo and earlier, and search indices for RAG purposes.

101. In addition, each successive GPT model has had orders of magnitude more parameters than the last. For instance, GPT-4 is reported to have 1.8 trillion parameters,[18] a tenfold increase from the 175 billion parameters used to train GPT-3.[19] Because adding more parameters requires training on more data, it is unlikely that Defendants would have foregone including Plaintiff's articles in their more recent training sets. Thus, upon information and belief, Defendants continue to include Plaintiff's articles in their training sets up to the present date.

[18] Maximilian Schreiner, GPT-4 architecture, datasets, costs and more leaked, *The Decoder* (July 11, 2023), https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.
[19] Tom B. Brown et al, Language Models are Few-Shot Learners, 5 (July 22, 2020), https://arxiv.org/pdf/2005.14165.

102. Further, the OpenAI Defendants' adoption of a tool in late 2023 to allow website owners to block web crawling would have been unnecessary if OpenAI was not continuing to copy content from the internet, including Plaintiff's copyright-protected works, as it had done in the past.

103. According to OpenAI's Chief of Intellectual Property and Content, each of OpenAI's models is "trained from scratch."[20] Thus, when assembling new training sets, OpenAI recrawls the same articles it included in past training sets.

104. As alleged above, 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. Thus, upon information and belief, they have continued to remove author, title, copyright notice, and terms of use information from Plaintiff's copyright-protected articles up to the present, including but not limited to Plaintiff's articles that are contained in Defendants' training sets created in the past three years.

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

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

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

107. 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.

[20] 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/.

108. 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.

109. 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.

110. 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.

111. 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 and abridged or regurgitated copyright-protected works verbatim or nearly verbatim.

112. 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 notice, or terms of use information.

113. 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.

114. 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.

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

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

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

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

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

119. 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.

120. 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.

121. 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.

122. 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.

123. Defendant Microsoft had reason to know that inclusion in training sets of Plaintiff's works of journalism without author, title, copyright, 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.

124. Defendant Microsoft had reason to know that inclusion in training sets of Plaintiff's works of journalism without author, title, copyright, 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, or terms of use information.

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

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

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

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

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

129. Upon information and belief, Defendant Microsoft shared copies of Plaintiff's works without author, title, copyright, 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 Plaintiff's copyrighted works from which author, title, copyright, or terms of use information was removed from their training sets and any other repositories;

(iii) An injunction prohibiting the unlawful conduct alleged above;

(iv) An injunction ordering the destruction of all GPT or other LLMs and training sets that incorporate Plaintiff's works from which author, title, copyright notice, or terms of use information have been removed; and

(v) Attorney fees and costs.

JURY DEMAND

Plaintiff demands a jury trial.

RESPECTFULLY SUBMITTED,

/s/ Stephen Stich Match

Jon Loevy*
Michael Kanovitz*
Matthew Topic*
Stephen Stich Match (No. 5567854)
Thomas Kayes*
Steven Art*
Kyle Wallenberg*
Shelley Geislzer*

LOEVY & LOEVY
311 North Aberdeen, 3rd Floor
Chicago, IL 60607
312-243-5900
jon@loevy.com
mike@loevy.com
match@loevy.com
matt@loevy.com
kayes@loevy.com
steve@loevy.com
wallenberg@loevy.com
geiszler@loevy.com

**pro hac vice*

April 18, 2025