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.

Civil Action No. _______________

COMPLAINT

JURY TRIAL DEMANDED

1. Plaintiff The Intercept Media, Inc., through its attorneys Loevy & Loevy, for its Complaint against the OpenAI Defendants, 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 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 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 not to acknowledge or respect copyright, not to notify ChatGPT 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.

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 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 invested billions of dollars in OpenAI Global LLC and will own a 49% stake in the company after its investment has been repaid.

21. Microsoft provides the data center and supercomputing infrastructure used to train ChatGPT.

22. Upon information and belief based on the relationship between Defendants, 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

23. 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.

24. 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, on 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.

25. 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.

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

27. 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.

28. Defendants consented to venue in this Court in at least *Authors Guild v. OpenAI Inc.*, 23-cv-08292.

DEFENDANTS' DMCA VIOLATIONS

29. Defendants have kept secret the specific content used to train all versions of ChatGPT beginning with GPT-4. Plaintiff's allegations 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.

30. Earlier versions of ChatGPT[1] (prior to GPT-4) were trained using at least the following training sets: WebText, WebText2, and Common Crawl. These training sets range from collections of links posted on the website Reddit to a scrape of most of the internet.

31. WebText and WebText2 were created by the OpenAI Defendants. Common Crawl originated elsewhere, but was adapted and utilized by Defendants for inclusion in ChatGPT

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

training sets. Upon information and belief, both the OpenAI Defendants and Defendant Microsoft have created Common Crawl datasets, as opposed to copying a dataset already created by someone other than Defendants.

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

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.

34. ChatGPT offers a product to its customers that provides responses to questions or other prompts. ChatGPT's ability to provide these responses is the key value proposition of its product, one which it is able to sell to its customers for enormous sums of money, soon likely to be in the billions of dollars.

35. At least some of the time, ChatGPT provides or has 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.

36. At least some of the time, ChatGPT provides or has 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 about a current event or the results of a work of investigative journalism, ChatGPT 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.

37. ChatGPT does not have any independent knowledge of the information provided in its responses. Rather, to service Defendants' paying customers, ChatGPT instead repackages,

among other material, the copyrighted journalism work product developed by Plaintiff and others at their expense.

38. Various sources have recreated approximations of the Common Crawl and WebText training sets based on publicly available information discussing the methodologies used to create them. Those sources have made these recreated data sets, or instructions on how to derive them, available on the internet. Thousands of Plaintiff's works are contained in the recreated versions of these data sets without the author, title, copyright notice, and terms of use information found in Plaintiff's original publications.

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

40. When ChatGPT provides responses to users, it generally does not provide the author, title, copyright notice, or terms of use information applicable to the works on which its responses are based. Upon information and belief, in the instances in which author or title information is included in a response, it is because other material used in a training set references the author or title in the text of such material (e.g., a Wikipedia article discussing the underlying works of journalism).

41. When providing responses, ChatGPT gives the impression that it is 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 simply mimics.

42. Based on the publicly available 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.

43. Based on the publicly available information described above, the OpenAI Defendants intentionally removed author, title, copyright notice, and terms of use information from Plaintiff's copyrighted works in creating ChatGPT training sets.

44. 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 Bing Copilot, Defendant Microsoft intentionally removed author, title, copyright notice, and terms of use information from Plaintiff's copyrighted works in creating ChatGPT and Bing Copilot training sets.

45. 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.

46. 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.

47. Defendants had reasonable grounds 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.

48. 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.

49. 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, 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.

50. 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 they 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.

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

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

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

53. 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.

54. 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.

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

56. 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.

57. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiff's works of journalism without author, title, copyright, 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.

58. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiff's works of journalism without author, title, copyright, 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.

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

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

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

62. The OpenAI Defendants have acknowledged that use of copyright-protected works to train ChatGPT requires a license to that content and, in some instances, have entered licensing agreements with large copyright owners such as Associated Press and Axel Springer. The OpenAI Defendants are also in licensing talks with other copyright owners in the news industry, but have offered no compensation to Plaintiff.

63. 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, which is enabled, facilitated, and concealed by the OpenAI Defendants' removal of author, title, copyright, and terms of use information from their training sets.

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

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

65. 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.

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

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

67. 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.

68. 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.

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

70. 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.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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

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

77. 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.

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, and terms of use information was removed from their training sets and any other repositories;

(iii) 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*

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

February 28, 2024