**UNITED STATES DISTRICT COURT**
**SOUTHERN DISTRICT OF NEW YORK**

RAW STORY MEDIA, INC., ALTERNET MEDIA, INC.,

Plaintiffs,

v.

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

Defendants.

No. 1:24-cv-01514-CM

**PLAINTIFFS' REPLY IN SUPPORT OF THEIR MOTION FOR LEAVE TO AMEND COMPLAINT OR, IN THE ALTERNATIVE, TO CONTINUE TAKING JURISDICTIONAL DISCOVERY**

# TABLE OF CONTENTS

# TABLE OF AUTHORITIES

***Page***

***Cases***



***Statutes***

# I. INTRODUCTION

Lacking any strong argument against amendment, Defendants oppose only by conflating Plaintiffs' current standing theory with one the Court rejected, ignoring inconvenient Second Circuit precedent under *TransUnion*, and neglecting key allegations in the Proposed First Amended Complaint on scienter. They also ignore that, when faced with a largely identical amended complaint in *Intercept*, Judge Rakoff found standing under the DMCA and held that the plaintiff stated a CMI removal claim against OpenAI. Indeed, the only material differences between the amended complaints are that Plaintiffs here allege facts supporting an ***additional*** standing theory—unjust enrichment—and two further grounds for scienter. The Court should reach the same conclusion and allow the proposed amendments.

# II. ARGUMENT

## A. The proposed amendments establish Plaintiffs' standing.

### 1. Plaintiffs' injuries are closely analogous to copyright infringement on the specific facts of this case.

Plaintiffs argued that they have alleged standing analogous to copyright infringement because, on the specific facts of this case, Defendants removed their CMI through a technical process that constitutes actual *prima facie* infringement. P. Br. at 3-6, ECF No. 119. In resisting that analogy, Defendants misconstrue this Court's prior opinion, Plaintiffs' standing theory, and the Supreme Court's holding in *TransUnion*.

Defendants cast this Court as holding that dissemination is an absolute prerequisite for a concrete injury. *See* D. Opp. at 7, ECF No. 122 ("[T]he Court has already decided as a matter of law that standing requires dissemination of Plaintiffs' works without CMI."). But the Court held no such thing. It required dissemination only after rejecting the sole alternative proposal on offer:

that violations of the DMCA are *per se* analogous to "injury for interference with property." Order at 6, ECF No. 117 ("Order"). It did not preclude standing theories the parties did not raise.

The Court, moreover, neither considered nor rejected the standing theory Plaintiffs now advance. In opposing Defendants' motion to dismiss, Plaintiffs argued that having one's CMI removed is itself an Article III injury because removal *per se* interferes with property. *See* ECF No. 70 at 6-7. That argument entails standing for ***all*** plaintiffs who allege Section 1202(b)(1) violations, since CMI removal—hence interference with property—is an element of the claim. *See* 17 U.S.C. § 1202(b)(1). The Court rejected that argument on its terms, holding that "the mere removal of identifying information from a copyrighted work" does not confer standing. Order at 7. But the current theory entails standing for only ***some*** Section 1202(b)(1) plaintiffs, namely those whose CMI was removed through an act that constitutes *prima facie* copyright infringement. Those plaintiffs have suffered an injury analogous to the historical injury of copyright infringement: their copyright has literally been infringed. The Court has held nothing contrary.[1]

Defendants mount no real resistance to the adequacy of Plaintiffs' allegations that they removed Plaintiffs' CMI through *prima facie* infringement.[2] They argue instead that the allegations are irrelevant because they go to Defendants' conduct rather than Plaintiffs' harm. *See*

[1] For this reason, Defendants are wrong to accuse Plaintiffs of seeking "reconsideration in disguise," D. Opp. at 8, as that accusation wrongly conflates Plaintiffs' standing theories. For avoidance of doubt, Plaintiffs are not now asking the Court to reconsider its rejection of their initial standing theory, though they reserve all rights as to that theory in any eventual appeal.
[2] The closest they come is to point out that Plaintiffs have not brought a copyright infringement claim because they have not alleged that they registered their copyrights with the Copyright Office. *See* D. Opp. at 9. But while registering a copyright might be a prerequisite to an infringement suit, it has no effect on the relevant question here: whether OpenAI has *prima facie* infringed Plaintiffs' copyright. *See Palmer/kane LLC v. Gareth Stevens Publ'g*, No. 15-cv-7404, 2016 WL 6238612, at *1 (S.D.N.Y. Oct. 24, 2016) (holding that "registration of a copyright claim is not a condition of copyright protection" but that it is "a prerequisite to bringing a civil copyright infringement action).

D. Opp. at 9. But the allegations go to both, and establish an Article III harm because infringement has always been a harm in itself. *See* P. Br. at 5 (citing authorities that infringement alone is a harm). So because infringement itself is an Article III injury, and because Plaintiffs' injuries are closely analogous to infringement, Plaintiffs have identified the required "close relationship to a harm traditionally recognized as providing a basis for a lawsuit in American courts." *TransUnion LLC v. Ramirez*, 594 U.S. 413, 424 (2021) (citation omitted). Just as the statutory violations in *TransUnion* sufficed for standing when they also ***resembled*** defamation, so too are the statutory violations here sufficient when the not only resemble, but ***constitute***, infringement of copyright, a historically recognized harm. Article III requires no more.

**2. Plaintiffs' injuries are also closely analogous to unjust enrichment.**

Plaintiffs argued that Defendants injured them by unlawfully profiting from removing Plaintiffs' CMI—thus, by realizing profits that rightly belong to Plaintiffs. Further, this injury is analogous to unjust enrichment, which has a strong common-law pedigree under *TransUnion*, and which has historically led to disgorgement of profits made in violation of a plaintiff's legally protected rights—here, rights conferred by the DMCA—whether or not they led to an additional economic loss. These premises entail standing under *TransUnion*. *See* P. Br. at 6-9.

Defendants do not dispute any of this directly. Instead, they argue that federal courts lack jurisdiction over claims alleging unlawful profits rightly belonging to the plaintiff unless those profits come with an additional economic loss. D. Opp. at 12. But they fail to even acknowledge the post-*TransUnion* Second Circuit case cited in Plaintiffs' opening brief that holds the opposite. *See* P. Br. at 6 (citing *Packer on behalf of 1-800-Flowers.Com, Inc. v. Raging Cap. Mgmt., LLC*, 105 F.4th 46, 51-56 (2d Cir. 2024), cert. denied, No. 24-408, 2024 WL 4743106 (Nov. 12, 2024)). Beyond that, none of their three sub-arguments avails.

First, Defendants argue that Plaintiffs conflate the elements of unjust enrichment—defined by state law—with federal standing law under Article III. *See* D. Opp. at 12. But there is nothing to conflate. After all, historical state-law claims ***determine*** federal standing law. Article III is satisfied when a plaintiff asserts an injury with a "'close relationship' to a harm 'traditionally' recognized as providing a basis for a lawsuit in American courts," and specifically when the plaintiff's injury has a "close historical or common-law analogue." *TransUnion*, 594 U.S. at 424 (quoting *Spokeo, Inc. v. Robins*, 578 U.S. 330, 341 (2016)). And these common-law analogues will typically be state-law claims. *TransUnion* itself is a prime example. Citing the Restatement of Torts, the Court held that the plaintiffs had standing if and only if they satisfied the "publication" element of common-law defamation—a state-law claim. *See id*. at 434. So this argument fails.

Second, Defendants try to liken Plaintiffs' arguments to those made by the dissents in *TransUnion* and *Thole v. U. S. Bank N.A*, 590 U.S. 538 (2020). *See* D. Opp. at 10-12. But neither majority opinion considered the relationship between Article III and unjust enrichment, and Defendants supply no citation to the contrary. *TransUnion* was about an analogy to defamation, and Justice Thomas's dissent mentions unjust enrichment only in passing. *See TransUnion*, 594 U.S. at 459 (Thomas, J., dissenting). In *Thole*, the dissent raised unjust enrichment to argue that the plaintiffs had standing as beneficiaries of a defined benefit plan under ERISA akin to the beneficiary of a private trust. *See Thole*, 590 U.S. at 558-59 (Sotomayor, J., dissenting). But as the dissent observed, the Court rejected that argument because beneficiaries of a defined benefit plan "are not similarly situated to the beneficiaries of a private trust," *id*. at 560 (Sotomayor, J. dissenting) (quoting *Thole*, 590 U.S. at 542), not because Article III bars some actions analogous to unjust enrichment. *Thole* thus has no bearing on this case.

Third, Defendants cite out-of-jurisdiction cases they portray to hold that standing for unjust enrichment or analogous claims requires an additional economic loss. But at least one of the cases, *Baehr v. Creig Northrop Team, P.C.*, 953 F.3d 244 (4th Cir. 2020), if anything, holds the opposite. There the Fourth Circuit rejected an analogy between unjust enrichment and statutory damages for violations of the Real Estate Settlement Procedures Act ("RESPA"). *See id*. at 257-58. Unlike unjust enrichment, RESPA's statutory damages do not disgorge the defendant's unlawful gain, but are instead pegged to the plaintiff's loss. *See* 12 U.S.C. § 2607(d)(2) (providing treble damages). In light of that difference, the court held, RESPA does not protect against injuries analogous to unjust enrichment—namely, a defendant's unlawful gain. *See Baehr*, 953 F.3d at 257-58. But it took pains to note that the defendant's same actions "might give rise to liability in a lawsuit brought under the unjust enrichment cause of action" rather than RESPA, confirming that Article III standing does not require an additional economic loss. *Id*. at 257. And the DMCA's statutory damages provision is analogous to unjust enrichment, not RESPA, since the DMCA's provision is one of disgorgement. *See* P. Br. at 7-8 (making this argument). So if anything, *Baehr* confirms Plaintiffs' standing. More, one of Defendants' other cases, *Del Vecchio v. Amazon.com, Inc*., No. 11-cv-366, 2012 WL 1997697 (W.D. Wash. June 1, 2012), does not discuss standing at all.[3]

Defendants' other out-of-jurisdiction cases all predate *TransUnion* and *Spokeo* and thus did not apply contemporary principles. *See* D. Opp. at 12-13 & n.2 (citing cases from 2015 and earlier).[4] In particular, they did not consider whether the alleged injuries existed historically or at

[3] Defendants cite *Del Vecchio* for the proposition that unjust enrichment (under Washington law) supposedly requires an additional monetary loss. D. Br. at 13 n.2. But *Del Vecchio* cites no authority holding that. The Washington case it cites, *Dragt v. Dragt/DeTray, LLC*, 161 P.3d 473, 576 (2007), adopts Section 1 of the Third Restatement, which states that the plaintiff need suffer no "observable loss." Restatement (Third) of Restitution and Unjust Enrichment, § 1 cmt. a.

[4] One of Defendants' cases, *McNamara v. City of Chicago*, 138 F.3d 1219 (7th Cir. 1998), also says nothing remotely related to unjust enrichment. The plaintiffs challenged the

common law, or were analogous to such injuries. Plus, they cannot overcome the Second Circuit's application of contemporary principles in *Packer* to find standing based solely on a defendant's unlawful gain without an additional economic loss.

**B. In the alternative, jurisdictional discovery is warranted.**

If the Court holds that standing requires dissemination, it should allow discovery into whether OpenAI has disseminated Plaintiffs' articles. As an initial matter, Defendants have no response to the fact that such discovery would have been complete by now had they timely responded to Plaintiffs' RFPs, and for that reason is plainly warranted here. *See* P. Br. at 11-12.

Defendants' responses also do not persuade. They argue that Plaintiffs only "speculat[e]" that ChatGPT might have disseminated their articles, citing the Court's holding that it "seems remote" that ChatGPT plagiarized Plaintiffs' content. D. Opp. at 20 (citing Order at 9). But the Court was applying the standard for jurisdiction over a claim for injunctive relief, which requires a "substantial" risk of imminent harm. Order at 9. That standard is far more stringent than the low bar for jurisdictional discovery. *See, e.g.*, *Ayyash v. Bank Al-Madina*, No. 04-cv-9201, 2006 WL 587342, at *5 (S.D.N.Y. Mar. 9, 2006) (holding jurisdictional discovery warranted where the plaintiff has made a "threshold showing that there is some basis for the assertion of jurisdiction"). Indeed, the discovery standard must be lower than that for alleging jurisdiction, for otherwise every plaintiff entitled to discovery on jurisdiction will have established jurisdiction, and jurisdictional discovery would be an empty vessel. Here, Plaintiffs cannot allege with certainty that ChatGPT disseminated its articles. But the presence of tens of thousands of its articles in the training sets, plus ChatGPT's propensity to plagiarize, at least makes the requisite start. *See* P. Br. at 11.

---

constitutionality of an affirmative action plan used to promote firefighters. The court held that plaintiffs would typically be required to allege that they would have been promoted absent the allegedly unlawful plan. *See id*. at 1221.

Defendants next argue that Plaintiffs do not need discovery because other parties were able to identify disseminations without it. D. Opp. at 20. That argument is disingenuous. Those parties ***created*** the disseminations but did not identify disseminations produced by ChatGPT in response to prompts by ***others***, which is what Plaintiffs' discovery seeks. PFAC ¶ 77, ECF No. 119-1. While Plaintiffs did not create disseminations, they plausibly alleged that that was due to recent changes OpenAI made to ChatGPT and does not reflect the state of the product years ago, when OpenAI was more cavalier about copyright issues. PFAC ¶ 67.

Defendants last argue that "Plaintiffs' inability to recognize their own injury without discovery only confirms that their purported harm is speculative." D. Opp. at 21. If accepted, that argument would prove that plaintiffs can never get jurisdictional discovery to establish injury in fact. But that is not the law. *See, e.g.*, *Dunne v. Ricolcol*, No. 21-56254, 2024 WL 5088112, at *1 (9th Cir. Dec. 12, 2024) (permitting court on remand to order jurisdictional discovery into injury in fact); *Marsh & McLennan Agency LLC v. Williams*, No. 22-cv-8920, 2023 WL 4534984, at *1 (S.D.N.Y. July 13, 2023) (ordering jurisdictional discovery into injury in fact).

**C. The proposed amendments satisfy Rule 12(b)(6).**

Defendants do not dispute that the proposed amendments resolve two of their initial grounds for dismissal: identification of the works at issue, and allegations supporting Defendants' intent to remove CMI. And in *Intercept*, Judge Rakoff resolved the other two issues in the plaintiff's favor based on a materially identical amended complaint. *Compare* Match Decl. Ex. 1 *with* ECF No. 119-1; *see* Match Decl. Ex. 2.[5] This Court should do the same.

[5] Because Judge Rakoff has thus far only issued a bottom-line order, it is unclear which scienter theories he accepted. But since he allowed the case to go forward, he accepted at least one. That was necessarily one of the first three theories presented in this brief and Plaintiffs' opening brief, since the *Intercept* plaintiff did not raise the other two.

As to the third issue, statutory standing, OpenAI adds little of substance to its prior briefing. It cites *CoxCom, Inc. v. Chaffee*, 536 F.3d 101 (1st Cir. 2008), which it mistakenly casts as a Second Circuit case, for the proposition that the "question of constitutional standing is not the same as establishing [] standing to sue … as a 'person injured' under the DMCA." D. Opp. at 14 (quoting *CoxCom*, 536 F.3d at 107 n.7). But the First Circuit was simply noting that Congress cannot create standing when Article III forbids it. *See CoxCom*, 536 F.3d at 107 n.7 ("Congress cannot erase Article III's standing requirements by statutorily granting the right to sue to a plaintiff who would not otherwise have standing."). It did not endorse Defendants' proposition—that the bar for injury is higher under the DMCA than the Constitution—since the defendants made no arguments under the DMCA's injury provision at all. *See id*. *CoxCom* also involved claims under a different section of the DMCA—Section 1201—and so has no bearing here. *See id*. at 104.

On the last issue, the second scienter element, Defendants' responses all fail.

***Concealing Defendants' training-based infringement from their users***. Defendants' only response is to misconstrue the theory. Defendants make the banal observation that removing CMI cannot conceal "a non-public dataset." D. Opp. at 15. But Plaintiffs have not alleged otherwise. The point, rather, is that Defendants concealed their own infringement via the process of reproducing Plaintiffs' articles to ChatGPT users. A reproduction that incorporates Plaintiffs' works ***and*** Plaintiffs' CMI would communicate to users that the reproduction resulted from an infringing copy in the training set, and thus that OpenAI committed training-based infringement. Conversely, one that incorporates Plaintiffs' works ***without*** Plaintiffs' CMI conveys no such thing. So omitting CMI from the reproduction works a concealment. And critically, Plaintiffs alleged,

removing CMI from the training set causes it to be absent from the reproduction. *See* PFAC ¶¶ 81, 89. Defendants have no direct response to this theory properly construed.[6]

***Concealing Defendants' output-based infringement from their users***. Defendants at least describe this theory mostly correctly,[7] yet their rebuttals are wrong. They argue that their knowledge of the "general phenomenon of regurgitation" does not equate to knowledge that removing CMI will conceal an infringement. D. Opp. at 16-17. But Defendants' reason to know this derives not only from their expertise in regurgitation, but the fact that OpenAI designed the product at issue, and that removing CMI in training causes CMI to be absent from ChatGPT outputs—clearly a plausible inference at the pleading stage given the low bar for scienter. *See* P. Br. at 14-15 (collecting cases on scienter pleading standards). Defendants also say that Plaintiffs fail to allege this causal connection. D. Opp. at 17. But that is just false. *See* PFAC ¶ 81 ("If ChatGPT was trained on works of journalism that included the original author, title, and copyright information, ChatGPT would have learned to communicate that information."). And while Defendants point out that Plaintiffs have not alleged dissemination of their articles, D. Opp. at 17, they fail to contend with Plaintiffs' argument that this is unnecessary under the DMCA, which does not require an infringement to have occurred. *See* P. Br. at 15-16 (citing *Murphy v. Millennium Radio Grp. LLC*, No. 08-cv-1743, 2015 WL 419884, at *4-5 (D.N.J. Jan. 30, 2015)).

***Inducing, enabling, or facilitating users to infringe***. Apart from objections already addressed above, Defendants dispute this theory by assailing as "wholly conclusory" the

[6] Defendants' citation to *Tremblay v. OpenAI, Inc.*, 716 F. Supp. 3d 772 (N.D. Cal. 2024) confounds. The plaintiffs there argued that removing CMI would enable users to infringe, which is completely different from the issue Defendants cite it for. *See id*. at 778.

[7] Defendants suggest Plaintiffs to be alleging that Defendants removed the CMI with a purpose to conceal infringement. *See* D. Opp. at 16. But the second scienter element requires only knowledge, not purpose. *See* 17 U.S.C. § 1202(b).

proposition that users would refrain from distributing works they know to be copyrighted. D. Opp. at 17-18. But it is common sense—and certainly plausible at the pleading stage—that people are less likely to perform conduct they know is illegal.

***Facilitating Defendants' training-based infringement***. Defendants do not dispute the adequacy of Plaintiffs' allegation that removing CMI facilitates LLM training. They instead argue that Plaintiffs have not shown how LLM training constitutes infringement. D. Opp. at 18-19. They suggest that the infringement must be one of several discrete parts of the training process, such as the act of downloading Plaintiffs' works. But that defines the infringement too narrowly. Training itself is the infringement, as Defendants have recognized in other cases. *See, e.g.*, Memorandum of Law in Support of OpenAI Defendants' Motion to Dismiss, at 2-3, *The New York Times Company v. Microsoft Corp.*, No. 23-cv-11195 (S.D.N.Y. Feb. 26, 2024), ECF No. 52 (defining the "genuinely important issue at the heart of this lawsuit" as "whether it is fair use under copyright law to use publicly accessible content to train generative models").

***Facilitating Defendants' copying***. Defendants complain that Plaintiffs supposedly have not explained why removing CMI facilitates copying, D. Opp. at 19, but that is false. Articles without CMI take up fewer computational and storage resources, which frees up those resources for further copying of other articles. PFAC ¶ 96. And *contra* Defendants, this theory does not contradict Plaintiffs' allegation that Defendants first download an article and then extract CMI. D. Opp. at 19 (citing PFAC ¶ 46). Removing CMI from already-downloaded articles saves resources compared to leaving the CMI on them.

## III. CONCLUSION

The Court should grant Plaintiffs leave to amend their Complaint. In the alternative, it should permit them to continue taking discovery of Defendants' dissemination of their works.

RESPECTFULLY SUBMITTED,

*/s/ Stephen Stich Match*

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

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

January 21, 2025

**UNITED STATES DISTRICT COURT**
**SOUTHERN DISTRICT OF NEW YORK**

RAW STORY MEDIA, INC., ALTERNET MEDIA, INC.,

Plaintiffs,

v.

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

Defendants.

No. 1:24-cv-01514-CM

**DECLARATION OF STEPHEN STICH MATCH**

I, Stephen Stich Match, declare as follows:

1. I am an attorney at Loevy & Loevy, which represents Plaintiffs Raw Story Media, Inc. and AlterNet Media, Inc. in this case.

2. A true and correct copy of the First Amended Complaint in *The Intercept Media, Inc. v. OpenAI, Inc.*, No. 24-cv-01515 (S.D.N.Y. June 21, 2024), ECF No. 87 is attached as Exhibit 1. Its exhibits are omitted.

3. A true and correct copy of an order in *The Intercept Media, Inc. v. OpenAI, Inc.*, No. 24-cv-01515 (S.D.N.Y. Nov. 21, 2024), ECF No. 122 is attached as Exhibit 2.

I declare under penalty of perjury that the foregoing is true and correct. Executed on January 21, 2025.

*/s/ Stephen Stich Match*

Stephen Stich Match

# EXHIBIT 1

**UNITED STATES DISTRICT COURT**
**SOUTHERN DISTRICT OF NEW YORK**

| | |
|---|---|
| THE INTERCEPT MEDIA, INC.,<br><br>Plaintiff,<br><br>v.<br><br>OPENAI, INC., OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, OPENAI HOLDINGS, LLC, and MICROSOFT CORPORATION<br><br>Defendants. | No. 1:24-cv-01515-JSR<br><br>**FIRST AMENDED COMPLAINT**<br><br>**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.

**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."[1]

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.

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."[2]

---

[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.

[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.

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.

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 a 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.[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.

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.[4]

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."[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.

---

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

[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

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.[6]

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.[7] 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.[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

[6] https://github.com/openai/gpt-2/blob/master/domains.txt.
[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.

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."[9] 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 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.

[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.

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.[10]

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.[11]

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 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

[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.

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 AND MIMICKING 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.[12] Thus, upon information and belief, the OpenAI Defendants possess a repository of every regurgitation of Plaintiff's works apart from those whose storage users have affirmatively disabled.

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

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.[13] 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.

[13] 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.

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

68. 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.

69. 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.

70. 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.

71. 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.

72. 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**

73. 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.

74. 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.

75. 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.

76. 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

77. 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.

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

79. 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.[14] This admission, while qualified, confirms that these deals involve training, at least in part.

80. 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.

[14] 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/.

81. 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.

82. 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.

83. 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.

84. 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.

85. 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.

86. 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,"[15] 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."[16] 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. § 1202(b)(1) by OpenAI Defendants**

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

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

89. 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.

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

90. 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.

91. 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.

92. 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.

93. 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.

94. 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.

95. 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.

96. 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.

97. 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.

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

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

99. 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**

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

101. 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.

102. 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.

103. 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.

104. 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.

105. 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.

106. 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.

107. 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.

108. 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.

109. 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**

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

111. 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, 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 (*pro hac vice*)
Michael Kanovitz (*pro hac vice*)
Lauren Carbajal (*pro hac vice*)
Stephen Stich Match (No. 5567854)
Matthew Topic (*pro hac vice*)

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

June 21, 2024

# EXHIBIT 2

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

THE INTERCEPT MEDIA, INC.,

Plaintiff,

-v-

OPENAI, INC., et al.,

Defendants.

24-cv-1515 (JSR)

ORDER

JED S. RAKOFF, U.S.D.J.:

On June 21, 2024, plaintiff The Intercept Media, Inc. ("The Intercept") filed an amended complaint in compliance with the schedule set out in the Court's order issued on June 6, 2024. See ECF Nos. 81, 87. On July 8, 2024, defendants OpenAI[1] and Microsoft Corporation ("Microsoft") renewed their previously filed motions to dismiss and submitted supplemental briefs in further support of their motions. See ECF Nos. 88, 89. One week later, The Intercept submitted a supplemental brief in opposition to defendants' motions. See ECF No. 90.

On November 1, 2024, the Court held oral argument on defendants' motions and advised counsel for all parties that it would issue a bottom-line order by November 22, 2024. Accordingly, the Court (1) grants Microsoft's motion in full and with prejudice, and (2) grants OpenAI's motion in part, dismissing The Intercept's claim under 17

[1] The Intercept sued OpenAI, Inc.; OpenAI GP, LLC; OpenAI, LLC; OpenAI OpCo LLC; OpenAI Global LLC; OAI Corporation, LLC; and OpenAI Holdings, LLC. Because The Intercept's allegations do not distinguish among these entities, this Order generally refers to "OpenAI."

U.S.C. § 1202(b)(3) with prejudice but allowing The Intercept's claim under 17 U.S.C. § 1202(b)(1) to proceed past the motion-to-dismiss stage.

An Opinion explaining the reasons for this ruling will issue in due course. The Clerk of Court is respectfully directed to close docket entry numbers 49 and 52.

SO ORDERED.

New York, NY
November 21, 2024

JED S. RAKOFF, U.S.D.J.