UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

THE CENTER FOR INVESTIGATIVE REPORTING, INC.,

Plaintiff,

v.

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

Defendants.

Case No. 24-cv-04872-SHS-OTW

ORAL ARGUMENT REQUESTED

PLAINTIFF'S COMBINED RESPONSE TO DEFENDANTS' MOTIONS TO DISMISS THE FIRST AMENDED COMPLAINT

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Page

Cases

Statutes

Other Authorities

I. INTRODUCTION

In moving to dismiss, Defendants seek to limit this case to what they call the "core issue," which they frame as whether training their large language models on unlicensed copies of news articles and then using those same models to compete with the journalists who wrote them is "fair use." OpenAI Br. at 2, ECF No. 100. That question is indeed critical, but its resolution is replete with factual disputes, why is why Defendants have not moved to dismiss on it.

It is also not the only thing in dispute. The case also involves essential questions about whether Defendants' GenAI products may abridge news articles without permission (a new form of competition with journalists), help their users create even more unlicensed copies, and disrupt the integrity of a news organization's copyright management information in order to facilitate their products' creation and obfuscate the extent of their infringement. These issues are serious, and existential to the news industry. And none can be resolved on a motion to dismiss.

OpenAI first moves to dismiss a narrow part of Plaintiff The Center for Investigative Reporting's ("CIR's") direct infringement claim based not on regurgitations or training, but on the increasingly controversial abridgements that summarize CIR's works, displacing the market for the originals in many cases.[1] OpenAI argues that CIR's examples are insufficiently similar, but under applicable case law they are, and OpenAI's approach to this issue is unworkable and legally unsupported. And in any event, these were simply the examples CIR was able to generate as part of the prefiling due diligence process, which is all that is required at this stage. CIR is entitled to discovery into OpenAI's responses to ChatGPT's prompts in the regular course of business to identify the extent to which it has abridged CIR's works. Given Defendants' secrecy and the stakes

[1] *See* White Paper: How the pervasive copying of expressive works to train and fuel generative artificial intelligence systems is copyright infringement and not fair use, *News/Media Alliance*, 4 & n.5 (2023), https://www.newsmediaalliance.org/wp-content/uploads/2023/10/AI-White-Paper-with-Technical-Analysis.pdf .

for the news industry, CIR should not be held to a stricter pleading standard than the law requires. Notably, Microsoft does not seek dismissal on abridgements.

Both Defendants also move to dismiss CIR's claim for contributory infringement, which alleges that Defendants helped their users produce responses that infringe CIR's copyright. Defendants argue that the First Amended Complaint does not plead user infringement and Defendants' knowledge, but these arguments are refuted by many well-pleaded allegations. This includes the allegations that OpenAI admitted that ChatGPT regurgitates, and all Defendants have pledged to defend and indemnify some of their users against claims of infringement.

Defendants next ask the Court to dismiss CIR's claims under the Digital Millennium Copyright Act ("DMCA"), which assert that Defendants unlawfully removed CIR's copyright management information ("CMI") from CIR's works used to train their products, and distributed CMI-less copies to each other. They move first on standing grounds, arguing primarily that CIR has not alleged a concrete injury, and also positing that CIR must allege public dissemination of its works. But CIR has alleged two injuries that do not involve public dissemination: Defendants' unlawful interference with CIR's copyright-protected property and their unlawful profits from removing CMI. Supreme Court precedent only requires an analogous historical injury, and the analogy between CIR's DMCA claim and copyright infringement (which has never required dissemination), is far stronger than analogies found sufficient by the Second Circuit. Even if dissemination were required, CIR has alleged it. And in the event the Court finds more is required, it should order jurisdictional discovery, since only Defendants possess the additional evidence.

Last, Defendants move to dismiss the DMCA claims on the merits. They focus primarily on scienter, which courts typically resolve only after discovery given liberal scienter pleading standards. The Court should do so here given that Defendants' knowledge is uniquely within their possession, and CIR has articulated four types of scienter Defendants plausibly possess.

Recognizing the weakness of its defenses, OpenAI has begun licensing news content from large news organizations—implicitly acknowledging that it cannot just gobble up journalists' works, profit from it, and decline to pay. OpenAI has recently achieved a valuation of over $150B by leveraging journalists' works. As CIR alleges, OpenAI is now paying substantial sums to large organizations for the rights to do exactly what CIR (a smaller but important independent nonprofit newsroom) alleges Defendants are doing with CIR's works but without paying. Defendants' lawless attempt to reshape the internet and the news industry by helping themselves to works of journalism pirated from organizations like CIR violates the law and must not be permitted.

II. BACKGROUND

A. CIR's substantial investment in groundbreaking investigative journalism.

As the country's oldest nonprofit newsroom, CIR has produced groundbreaking investigative journalism for nearly five decades. First Amend. Compl. ¶ 2, ECF No. 88 ("FAC"). Its accolades are too numerous to name, having won a George Polk Award, a Peabody Award, a Webby Award, and a Robert F. Kennedy Human Rights Award, all in the last few months. *Id.*

CIR's investigative journalism involves more than putting facts on a page. Many of its articles take months or years to report, costing significant sums of money. *Id.* ¶ 41. The finished product weaves complex concepts into an original tapestry of expressive thought.[2]

CIR operates two brands relevant to this case: Mother Jones and Reveal. *Id.* ¶ 12. Mother Jones started as a reader-supported magazine that has registered issues with the Copyright Office since the late 1970s. *Id.* ¶¶ 13, 36. Since 1993, it has published all its magazine articles, plus tens

[2] Some examples can be found at the links referenced in Exhibits 10 and 11 of the First Amended Complaint, which OpenAI argues are incorporated by reference. *See* OpenAI Br. at 8.

of thousands more, online; its pre-1993 articles are digitized and also available on the internet. *Id.* ¶¶ 37-40; FAC Ex. 3, ECF Nos. 88-3-88-6. Reveal also operates an online investigative news site with thousands of online articles. FAC ¶¶ 14, 40; FAC Ex. 4, ECF No. 88-7.

B. Defendants' unauthorized use of CIR's articles to train large language models worth billions of dollars.

Defendants OpenAI and Microsoft operate chatbots called ChatGPT and Copilot. FAC ¶ 4. Each is powered by a large language model, or LLM. *Id.* ¶ 48. Large language models "take text prompts as inputs and emit outputs to predict responses that are likely to follow." *Id.* ¶ 48. They can make these predictions only because they were trained on human expression found in billions of inputs, many of which are copyright-protected works created by humans. *Id.* ¶¶ 48-49.

Defendants train their LLMs using content copied from the internet, including CIR's copyright-protected articles. Though OpenAI has recently adopted a policy of secrecy surrounding its LLM training, it used to be more, well, "open." *Id.* ¶ 51. For instance, it has admitted that one of its training sets, WebText, contains 16,793 distinct URLs from Mother Jones's web domain. *Id.* ¶ 55. OpenAI also published instructions for how to assemble WebText. *Id.* ¶ 54. Computer scientists followed those instructions and released the results online. Those results contain a similar number of Mother Jones URLs and an additional 415 distinct URLs from Reveal's web domain. *Id.* ¶ 57. Put simply, OpenAI has admitted to copying tens of thousands of CIR's copyright-protected articles to train its models. And it did so without CIR's permission. *Id.* ¶ 77.

Defendants trained their LLMs on CIR's news articles only after removing author, title, terms of use, and copyright notice information (together, "copyright management information" or "CMI"). When assembling WebText, OpenAI used computer algorithms called Dragnet and Newspaper, which are designed to exclude CMI from the material extracted from the article. *Id.* ¶¶ 59-62. Defendants choose not to include CMI in their training sets because in addition to revealing the extent to which chat responses are based on copyrighted works, it would lead the LLMs to incorrectly learn how English writers express themselves. *Id.* ¶ 119. For instance, if an LLM were trained largely on content ending with a copyright notice, it would learn that English writers typically put copyright notices at the end of sentences—which they do not except when ascribing copyright. *Id.* The only alternative would be to retrain the model not to emit copyright notices, which is more resource-intensive and expensive than removing CMI from the start. *Id.*

This large-scale copying, helped by the resources Defendants saved by removing CMI, has transformed ChatGPT and Copilot into highly lucrative products. OpenAI is now a multi-billion-dollar business into which Microsoft has alone invested billions of dollars for a 49 percent stake. *Id.* ¶¶ 27, 47. As part of what Microsoft's CEO described as its "deep[]" partnership with OpenAI, Microsoft provides the data center and bespoke supercomputing infrastructure that powers the training of ChatGPT. *Id.* ¶¶ 27-28. And to accomplish their joint purposes, OpenAI and Microsoft have shared the training data with each other. Microsoft's CEO has admitted as much, stating that "we have the data, we have everything." *Id.* ¶ 29.

C. Unauthorized outputs of CIR's works by Defendants and their users.

Defendants profit from responses ChatGPT and Copilot provide their users in reaction to user prompts. To formulate responses, Defendants' products employ technical methods, including "synthetic searching," also known as "retrieval-augmented generation" or "RAG," that involve making copies of online content beyond those made to train their models. FAC ¶ 88. The resulting outputs then often consist of regurgitations, excerpts, or abridgements of copyright-protected content. *Id.* ¶¶ 79-98. Discovery is required to unearth the extent to which ChatGPT and Copilot have abridged or regurgitated CIR's works. But OpenAI has admitted that regurgitation happens, *id.* ¶ 83, and ChatGPT and Copilot's ability to regurgitate or abridge CIR's works is proven by example. *See* FAC Exs. 10, 11. ChatGPT itself has even essentially said that regurgitations can violate copyright laws: when CIR attempted to obtain the same regurgitations as set forth in the *Daily News* case, ChatGPT refused to produce them on copyright grounds. *Id*.

Both OpenAI and Microsoft know of their users' possible infringement. Both have committed to defend and indemnify their corporate users against copyright infringement claims. *Id*. This commitment applies ***only*** when the users deploy the product as advertised. *Id*. And OpenAI promotes ChatGPT as a tool for users to generate content for a future audience, *id*. ¶ 115, an act that would often be infringing when derived from an underlying copyrighted work.

III. LEGAL STANDARD

Rule 8 requires "a short and plain statement of the claim showing that the pleader is entitled to relief." Fed. R. Civ. P. 8(a)(2). When assessing a motion to dismiss under Rule 12(b)(6), the "the court must "accept[] all factual allegations as true and draw[] all reasonable inferences in favor of the plaintiff." *Sierra Club v. Con-Strux, LLC*, 911 F.3d 85, 88 (2d Cir. 2018). On a motion to dismiss under Rule 12(b)(1) for lack of standing, the court's resolution depends on whether the motion is facial—based solely on the allegations in the complaint—or fact-based. *Sonterra Cap. Master Fund Ltd. v. UBS AG*, 954 F.3d 529, 533 (2d Cir. 2020). When the motion is facial, as Defendants' motions are, the court must "accept[] as true all material factual allegations of the complaint, and draw[] all reasonable inferences in favor of the plaintiff." *Id*. (cleaned up). In these cases, "the plaintiff has no evidentiary burden." *Id*.

IV. ARGUMENT

A. CIR states a direct infringement claim based on abridgements.

No defendant has moved to dismiss CIR's direct infringement claims in full. All agree that CIR has stated direct infringement claims based on training, synthetic searching, and regurgitation. FAC ¶¶ 128-31; 136-39. But OpenAI has moved to dismiss CIR's direct infringement claim based on abridgements, *id*. ¶ 132, arguing that CIR has not adequately pled that ChatGPT's abridgements are substantially similar to the originals. *See* OpenAI Br. at 8-13. Microsoft has not made an analogous argument, and for good reason: it is misguided on three main grounds.

First, because OpenAI does not move to dismiss the claim as to regurgitations, its argument depends entirely on the distinction between a copy and a derivative, like an abridgement. *See* 17 U.S.C. § 101 (defining "derivative work" to include abridgements). But under prevailing law, whether a secondary work is a copy or a derivative is often "completely superfluous" from the standpoint of substantial similarity: if the derivative is substantially similar to the original then the defendant has also made an unlawful copy. *Twin Peaks Prods., Inc. v. Publications Int'l, Ltd.*, 996 F.2d 1366, 1373 (2d Cir. 1993).[3] Given this and OpenAI's concession that the regurgitation claim can go forward, there is no reason to dismiss abridgements on substantial similarity grounds. If a ChatGPT response is substantially similar to the original article, OpenAI has violated CIR's copyright whether or not that response is also an abridgement.

For related reasons, dismissing abridgements now would tee up laborious line-drawing exercises for summary judgment and trial: for each given ChatGPT output, it is an abridgement that has been dismissed, or a regurgitation that has not? These exercises would also be pointless under prevailing law, since the question is the same: whether the secondary work is substantially similar to the original. Take, for example, ChatGPT Example 4 from Exhibit 11 to the First Amended Complaint. There ChatGPT summarizes a Mother Jones article and reproduces verbatim nine consecutive paragraphs from it (minus one sentence). FAC Ex. 11 at 6-7; Match Decl. Ex. 1. OpenAI classifies the verbatim reproduction as a regurgitation rather than an abridgement—

[3] The substantial similarity analysis sometimes differs between derivatives and copies, such as when the original and derivative are of different media. *See, e.g.*, *Castle Rock Ent., Inc. v. Carol Pub. Grp., Inc.*, 150 F.3d 132, 139-41 (2d Cir. 1998) (television show and trivia game); *Horgan v. Macmillan, Inc.*, 789 F.2d 157, 162-63 (2d Cir. 1986) (ballet choreography and photographs). Perhaps these types of differences could apply here. But OpenAI has not argued they do. And absent such an argument, the Court should assume the test is the same.

presumably because its abridgement argument would be doomed otherwise. OpenAI Br. at 10 n.8. But the factfinder can avoid that "superfluous" question, *Twin Peaks*, 996 F.2d at 1373, by simply asking whether the secondary work substantially resembles the original. Thus, the best approach under prevailing law is to deny OpenAI's motion and have the parties litigate, for each ChatGPT output, whether it is substantially similar to the original.

Second, discovery is warranted whether or not CIR's original articles are substantially similar to its examples—examples CIR prompted for prefiling purposes and for which it does not claim damages. Before even confronting whether these specific examples cross the substantial similarity threshold, they show, viewed in the light most favorable to CIR, that ChatGPT is able to summarize CIR's copyright-protected news articles—particularly given the FAC's other allegations, including that 60% of ChatGPT responses contain plagiarized content. FAC ¶ 4; FAC Ex. 11. More, OpenAI, and OpenAI alone, knows the extent of ChatGPT's prompts and responses: it has a repository of them. FAC ¶ 80. *See Argo Contracting Corp. v. Paint City Contractors, Inc.*, No. 00-cv-3207, 2000 WL 1528215 (S.D.N.Y. Oct. 16, 2000) (allowing "discovery to take place" because operative facts were within defendant's possession). And CIR will be taking discovery of OpenAI's repository anyway, as it houses the regurgitations OpenAI does not seek to dismiss. Dismissing the direct infringement claim as to abridgements is thus unwarranted.

Third, there is at least a fact dispute as to whether the sample ChatGPT responses are substantially similar to the original. "[Q]uestions of non-infringement have traditionally been reserved for the trier of fact," and often requires discovery or expert testimony *Peter F. Gaito Architecture, LLC v. Simone Dev. Corp.*, 602 F.3d 57, 63, 65 (2d Cir. 2010). And though courts may sometimes address substantial similarity at the dismissal stage, they must apply the Rule 12(b)(6) standard. Thus, they may dismiss only if "the similarity between two works concerns only non-copyrightable elements of the plaintiff's work, or because no reasonable jury, properly instructed, could find that the two works are substantially similar." *Gaito*, 602 F.3d at 63.

Under prevailing law, two works are substantially similar if "an average lay observer would recognize the alleged copy as having been appropriated from the copyrighted work." *Andy Warhol Found. for Visual Arts, Inc. v. Goldsmith*, 11 F.4th 26, 53 (2d Cir. 2021), *aff'd*, 598 U.S. 508 (2023). Substantial similarity pertains only to the protectable elements of the original. Thus, a secondary work that copies only facts is not substantially similar. *See Nihon Keizai Shimbun, Inc. v. Comline Bus. Data, Inc.*, 166 F.3d 65, 70 (2d Cir. 1999). But protectable elements include the description, selection, coordination, and arrangement of facts. *See id*. at 71. These protections are especially salient for abridgements, which *prima facie* infringe when they present expression encapsulating "the principal ideas of the larger work," but "where immaterial incidents are omitted and voluminous dissertations are cut down." *Twin Peaks*, 996 F.2d at 1375-76.

Under the proper standards, the abridgements in the FAC are substantially similar to the originals. OpenAI does not deny that CIR's fourth example is substantially similar. Nor could it, since after summarizing the whole article, the example excerpts verbatim nine consecutive paragraphs (minus one sentence). *See* Match Decl. Ex. 1. OpenAI instead says that the excerpt is not part of the abridgement but is instead a separate regurgitation, so it "supports a different claim." OpenAI Br. at 10 n.8. But it does not defend the premise underlying this approach: that a single ChatGPT response can constitute multiple secondary works: here, an abridgement (the summary) and a copy (the excerpt).[4]

[4] CIR at present takes no position on this approach. Given the newness of the technology, this argument is best considered at summary judgment, after CIR has taken discovery into how ChatGPT assembles its responses.

Next, a reasonable juror could view the first example as substantially similar to the original Mother Jones article, *Why Is the Colorado River Running Dry*?[5] This example easily satisfies the definition of an abridgement, as it as it sets forth the article's six main points by summarizing each seriatim. *See* FAC Ex. 11 at 2. More, while not necessary to qualify as an infringing abridgement, it copies specific examples the article uses to make these points. For instance, when summarizing the article's argument that agricultural practices are responsible for water shortages in the Colorado River, it states that alfalfa "consumes more than half of the water diverted from the river." FAC Ex. 11 at 2. Of course, CIR claims no copyright in facts about alfalfa. But CIR has not simply conveyed facts. It has instead selected a fact to make a nuanced argument about water shortages in the Colorado River. It is the selection and presentation of this fact, and others like it, that warrants copyright protection and makes this example substantially similar to the original.[6]

The same holds for the second example: a summary of an article on COVID-19 misinformation. The original discusses a viral post that makes false claims about the coronavirus and then debunks them. After being asked to summarize examples of misinformation reported in the article, ChatGPT supplies several in the same point-by-point form as the original. *See* FAC Ex. 11 at 4. In one point, for example, ChatGPT describes the post as claiming that the virus can be killed in 26-27 Celsius weather, but debunked by experts who note that it is the sun's ultraviolet light, not the heat of a warm day, that disinfects—all expressive points in the original. *See id.*[7]

[5] *See* Scott Carrier, Why Is the Colorado River Running Dry?, *Mother Jones* (Nov.-Dec. 2023), https://www.motherjones.com/environment/2023/12/colorado-river-basin-water-running-out-dry-southwest-drought/.

[6] OpenAI points out that some of these points appear in a different order. *See* OpenAI Br. at 12 n.11. "Reproducing original expression in fragments or in a different order, however, does not preclude a finding of substantial similarity." *Warner Bros. Ent. Inc. v. RDR Books*, 575 F. Supp. 2d 513, 537 (S.D.N.Y. 2008).

[7] *See also* Abigail Weinberg, There's a Facebook Coronavirus Post Going Viral Claiming to Be From Stanford. Don't Believe It, *Mother Jones* (Mar. 11, 2020), https://www.motherjones.com/politics/2020/03/theres-a-facebook-coronavirus-post-going-viral-claiming-to-be-from-stanford-dont-believe-it/.

For the last example—a summary of an article about a family's reckoning with a Jim Crow-era murder—a reasonable juror could easily conclude that ChatGPT's summary was "appropriated from the copyrighted work." *Andy Warhol*, 11 F.4th at 53. *See* FAC Ex. 11 at 5.[8] To be sure, the summary does not copy direct quotations from the article. But while direct copying is sufficient for substantial similarity, it is not necessary. *See Castle Rock Ent., Inc. v. Carol Pub. Grp., Inc.*, 150 F.3d 132, 140 (2d Cir. 1998) (finding substantial similarity even though "the direct quotations or close paraphrases … are few and almost irrelevant"). Substantial similarity can also be "comprehensive nonliteral," where the copy is "qualitatively but not exactly similar" to the original. *Ringgold v. Black Ent. Television, Inc.*, 126 F.3d 70, 75 n.3 (2d Cir. 1997). That applies to the relation between the original and the ChatGPT summary, which succinctly sets forth the article's main points.

The Court should not dismiss the direct infringement claim based on abridgements.

B. CIR states a contributory infringement claim.

All Defendants move to dismiss the contributory infringement claims. A defendant commits contributory infringement if, "with knowledge of the infringing activity, [it] induces, causes or materially contributes to the infringing conduct of another." *Arista Recs., LLC v. Doe 3*, 604 F.3d 110, 117 (2d Cir. 2010). CIR alleges that Defendants helped their users make infringing copies of CIR's articles. FAC ¶¶ 143-44.[9] Defendants dispute this on two grounds.

First, Microsoft argues that CIR "fails to allege any instance of direct infringement by a user." Microsoft Br. at 21, ECF No. 97. But CIR has alleged direct infringement by users with

[8] *See also* A Jim-Crow-Era Murder. A Family Secret. Decades Later, What Does Justice Look Like? *Mother Jones* (Nov.-Dec. 2021), https://www.motherjones.com/politics/2021/10/jim-crow-unsolved-murder-lynching-justice-margaret-burnham-stories-families/.

[9] OpenAI misreads the FAC when it states that that the acts of direct infringement are the "example outputs from the Complaint." OpenAI Br. at 14. The FAC alleges that Defendants contributed to "their users[']" infringement, not CIR's. FAC ¶ 144.

sufficient plausibility to warrant discovery. Direct infringement by users plausibly explains why Microsoft would commit to defend and indemnify them against infringement claims. FAC ¶ 118.

Microsoft also assails what it views as the lack of realism in the prompts CIR used to generate infringing outputs and the supposed minimal resemblance between the Copilot outputs and the original article. *See* Microsoft Br. at 22-23. But that argument (in addition to being a classic fact issue unsuitable for a motion to dismiss) ignores the fourth Copilot abridgement example, added by CIR's amendments. That example begins with the undeniably realistic prompt, "Tell me about the article 'Blood Money.' Please format your response as a summary followed by the actual text." FAC Ex. 11 at 14. Copilot then summarizes the article and spits out six paragraphs of significantly copied text. *See id.*; Match Decl. Ex. 2. Viewed in the light most favorable to CIR, these allegations show direct infringement by Copilot users.

Microsoft's lead case, *Matthew Bender & Co. v. W. Pub. Co.*, 158 F.3d 693 (2d Cir. 1998), confirms that discovery should go forward. It upheld summary judgment for the defendant after the plaintiff "hypothesized" possible third-party infringements but identified none after discovery. *Id.* at 706. Here, there is one feasible way to identify the precise incidents of direct infringement CIR has plausibly alleged: examine Copilot prompts and responses. *See* FAC ¶ 80. Indeed, key evidence of a third party's direct infringement may be in a defendant's sole possession. *See Arista Recs. LLC v. Usenet.com, Inc.*, 633 F. Supp. 2d 124, 150 (S.D.N.Y. 2009) (granting summary judgment to contributory infringement plaintiff based in part on "data from Defendants' server … which constitutes direct evidence of subscribers requesting to download Plaintiffs' copyrighted works"). Since CIR has plausibly alleged direct infringement by users, it should be entitled to discover specific incidents. If there were none, Microsoft can win at summary judgment.[10]

[10] Microsoft also misstates the significance of CIR's inability to recreate ChatGPT outputs obtained by earlier plaintiffs in other cases. See FAC ¶ 83; Microsoft Br. at 23-24. Microsoft says this makes it "implausible that a

Second, both Microsoft and OpenAI dispute that CIR adequately alleged knowledge of user infringement. Such knowledge can be actual or constructive. *See Arista Recs., LLC v. Doe 3*, 604 F.3d 110, 118 (2d Cir. 2010) (holding that "contributory infringement liability is imposed on persons who know or have reason to know of the direct infringement") (citation omitted); *Arista Recs., Inc. v. Mp3Board, Inc*., 2002 WL 1997918, at *6 (S.D.N.Y. Aug. 29, 2002) ("A defendant must possess either actual or constructive knowledge of the infringing activity to be found contributorily liable."). Further, in this Circuit, the defendant need not have knowledge of specific acts of infringement. *See Usenet*, 633 F. Supp. 2d at 154 ("[K]nowledge of specific infringements is not required to support a finding of contributory infringement"). Defendants' out-of-circuit appellate cases are unpersuasive as to the knowledge required for contributory infringement, since they both disavow the constructive knowledge prong the Second Circuit expressly adopted in *Doe 3*. *See BMG Rts. Mgmt. (US) LLC v. Cox Commc'ns, Inc*., 881 F.3d 293, 310 (4th Cir. 2018) (holding that contributory infringement does not lie when defendant "ha[s] reason to know of direct infringement"); *Luvdarts, LLC v. AT & T Mobility, LLC*, 710 F.3d 1068, 1072-73 (9th Cir. 2013) (requiring "actual knowledge" or "willful blindness").

Whatever the standard, CIR satisfies it. For one, OpenAI has "admitted that [its] products regurgitate material in response to user prompts." FAC ¶ 144.[11] Viewed in the light most favorable to CIR, OpenAI plausibly made this admission based on actual knowledge of specific instances of infringing regurgitations by users—how else would it have learned about the regurgitations? In fact, ChatGPT itself has said that ***specific*** regurgitations are infringing: in its

user would generate such output prompting in the dark." Microsoft Br. at 24. But viewed in the light most favorable to CIR, and as CIR plausibly alleges, the opposite is true: Defendants recently modified their products to emit less copyrighted information. See FAC ¶ 83. This suggests that users were more likely to generate infringing outputs earlier in the limitations period.

[11] *See also* OpenAI and journalism, *OpenAI* (Jan. 8, 2024), https://openai.com/index/openai-and-journalism/.

iteration operative at the time, it refused to produce regurgitations prompted by the *Daily News* plaintiffs because the original articles were "copyrighted content." *Id.* ¶ 83. More, Defendants plausibly know that their users infringe given that they have committed to defend and indemnify their users for that very conduct. *Id.* ¶ 118. These allegations raise a plausible inference of Defendants' knowledge, which is best resolved at summary judgment given that evidence of their knowledge is in their possession. *See Snail Games USA Inc. v. Tencent Cloud LLC*, No. 22-cv-02009, 2022 WL 3575425, at *5 (C.D. Cal. June 6, 2022) (holding knowledge of contributory infringement is best resolved at summary judgment). The Court should deny Defendants' motions as to contributory infringement.

C. The DMCA claims should go forward.

1. CIR has Article III standing.

CIR alleges that Defendants violated the DMCA by removing CMI from its copyright-protected works and distributing those works to each other without CMI. As discussed below, CIR advances two independently sufficient theories of standing, only one of which Defendants address. First, Defendants interfered with CIR's copyright-protected property, which is without more an Article III harm. And second, CIR has standing because Defendants unlawfully profited from removing CIR's CMI. Defendants' theory—that Article III requires public dissemination of CMI-less works—is therefore misguided. But assuming *arguendo* that it is correct, CIR has satisfied it, or is at least entitled to jurisdictional discovery.

i. CIR has standing because Defendants' interference with its copyright-protected works is an Article III injury.

Article III standing requires the plaintiff to have been injured in a way that (1) is "concrete, particularized, and actual or imminent," (2) was caused by the defendant, and (3) can be redressed by a court. *TransUnion LLC v. Ramirez*, 594 U.S. 413, 423 (2021). CIR first has standing because by removing CMI from its articles and distributing them without permission, Defendants interfered with CIR's copyright-protected property, which is an Article III injury. FAC ¶¶ 145-69. Defendants dispute this theory only on concreteness grounds.[12]

A plaintiff satisfies Article III's concreteness requirement if it has suffered either a historical or common-law injury, or an "analogue" to such an injury, which need not be an "exact duplicate." *TransUnion*, 594 U.S. at 424. Here, that analogue is copyright infringement. For the following reasons, though not an exact duplicate of the DMCA injuries CIR suffered, copyright infringement is close enough under Second Circuit and Supreme Court precedent to satisfy Article III. That is because CIR's DMCA injuries likely ***are*** copyright infringement, or at least very close.

Start with the Section 1202(b)(1) claim for unlawful removal of CMI. Defendants removed CIR's CMI through a technical process that involved downloading CIR's copyright-protected works and applying computer algorithms to create further copies with CMI deliberately removed. FAC ¶¶ 59-68. Their removing CMI in this way—through creating unauthorized copies of CIR's articles—violates CIR's exclusive right "to reproduce the copyrighted work." 17 U.S.C. § 106(a); *see also A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1014 (9th Cir. 2001) (holding that downloading files violates the reproduction right). Defendants cannot reasonably dispute that such copying amounts to *prima facie* copyright infringement. They instead maintain that they will prevail at a later stage of this proceeding based on the affirmative defense of fair use.[13] So because

[12] The other elements are plainly satisfied. The injury is particularized because Defendants manipulated CIR's works. FAC ¶¶ 145-169. The injury is actual because Defendants have already done so. *Id.* Defendants caused the injury by manipulating the articles themselves. *Id.* And CIR's injury is redressable by an award of statutory damages and/or an injunction. See 17 U.S.C. § 1203(b)(1); 17 U.S.C. § 1202(c)(3)(B).

[13] OpenAI's motion to dismiss the claim as to abridgements is irrelevant. The relevant analogue to the DMCA violations is infringement based on training, which OpenAI has not moved to dismiss.

CIR's Section 1202(b)(1) injuries are ***actual*** *prima facie* copyright infringement, these injuries are at least ***analogous*** to copyright infringement under Article III.[14]

Take next the Section 1202(b)(3) claims for distributing works without CMI. Violations of that section will often infringe the right to "distribute" copies of the work. 17 U.S.C. § 106(3). By definition, both require the distribution. While the distribution between OpenAI and Microsoft might not qualify as distribution "to the public," *see id.*, Article III standing does not require an "exact duplicate" of this historical injury. *TransUnion*, 594 U.S. at 424.[15]

These analogies are far closer than that between (1) encumbering voting shares through a fund's amendment to its bylaws and (2) trespass to chattels, which the Second Circuit found sufficient in *Saba Capital Cef Opportunities 1, Ltd. v. Nuveen Floating Rate Income Fund*, 88 F.4th 103 (2d Cir. 2023). Unlike here, the injury in *Nuveen* could not even ***theoretically*** satisfy the common law tort. Trespass to chattels requires the defendant to assume "physical control" of the chattel, which does not occur when a fund amends its bylaws. *Gary Friedrich Enters., LLC v. Marvel Enters., Inc.*, 713 F. Supp. 2d 215, 231 (S.D.N.Y. 2010) (quoting *Dowling v. U.S.*, 473 U.S. 207, 217 (1985)). It instead sufficed that the bylaw amendment caused something "analogous to a property-based injury." *Nuveen*, 88 F.4th at 116. Here, CIR's DMCA injuries likely are and arise from copyright infringement. The copyright infringement analogy is close.

[14] It is irrelevant to standing that the *prima facie* infringement might be fair use. Fair use is an affirmative defense. *See Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994) It does not deprive a court of jurisdiction.

[15] Though the relevant analogy is copyright infringement rather than common law defamation, *TransUnion*—which involved an analogy to defamation—if anything supports the proposition that distribution to just one other person ***can*** satisfy Article III. There, the Court divided the plaintiffs into two groups based on whether they met defamation's "publication" element. *TransUnion*, 594 U.S. at 432. Those who met it had standing; those who failed it did not. And defamation's publication element is satisfied by communication "to a single individual other than the one defamed." Restatement (Second) of Torts, ¶ 577 cmt. b (1977). OpenAI points out that publication does not include statements within a company or to a vendor. *See* OpenAI Br. at 16 n.17. But OpenAI and Microsoft are distinct entities that work together, and OpenAI cites no case rejecting the possibility of defamation between colleagues. Indeed, under the right analogy—copyright infringement—a violation occurs when a corporation uses copies for its own internal purposes. *See, e.g.*, *Am. Geophysical Union v. Texaco*, 60 F.3d 913, 914 (2d Cir. 1994) (holding that company infringed copyright in journal articles by coping them for researchers' internal use).

Having analogized its injuries to copyright infringement, CIR need do no more. That is because copyright infringement alone has always been actionable, including when its results are otherwise "uninjurious and unprofitable," *F. W. Woolworth Co. v. Contemp. Arts*, 344 U.S. 228, 233 (1952), and even when, other than the infringement itself, there is "no showing as to actual loss," *Jewell-La Salle Realty Co. v. Buck*, 283 U.S. 202, 208 (1931). In this respect copyright infringement is like quintessential common-law injuries, such as trespass to land and trespass to chattels, that recognize a distinct and concrete harm from interference with property. *See* Restatement (Second) of Torts § 163 (1977) (providing that trespass to land is actionable without any additional showing of harm); *Grace v. Apple, Inc.*, 328 F.R.D. 320, 346 (N.D. Cal. 2018) (listing states where, "interference is enough" for liability for trespass to chattels, even without a further "showing of harm"). CIR suffered similar harm when Defendants interfered with its copyright-protected works by removing CMI and distributing them to each other without authority.

Defendants resist this conclusion for three reasons. First, they argue that the copyright-based injury CIR asserts cannot be concrete because it "has no real-world effect." OpenAI Br. at 17. Not so. Under *TransUnion*, courts ***determine*** concreteness by considering whether the asserted injury is analogous to one "traditionally recognized as providing a basis for a lawsuit in American courts." *TransUnion*, 594 U.S. at 425 (cleaned up). Such injuries can be "intangible harms," including those "specified by the Constitution itself." *Id*. at 425. The Constitution, of course, specifies copyright. U.S. Const. art. I § 8 cl. 8. And the first Copyright Act, passed by the first Congress in 1790, mirrors the current one by authorizing statutory damages—50 cents per infringing page—without any further showing. *See* Act of May 31, 1790, ch. 15, § 2. Like all Article III injuries, this historical tradition is what ***makes*** copyright infringement concrete. CIR's DMCA injuries, as analogous to copyright infringement, are concrete for the same reason.

Further, Defendants' approach of second-guessing historical tradition would upend standing law and contradict *TransUnion*'s central holding. If interference with property is no longer concrete, then it hard to see how perhaps ***the*** quintessential common-law tort—pure trespass to land—could even be heard in federal court, despite its pre-Founding provenance. *See Entick v. Carrington*, 2 Wils. K. B. 275, 291, 95 Eng. Rep. 807, 817 (K. B. 1765).[16] Such a result is irreconcilable with *TransUnion*, under which the concreteness requirement derives its content from just these historical claims. This argument is meritless.

Second, according to Microsoft, (1) any DMCA injury must be analogous to attribution, and (2) attribution lacks historical pedigree. *See* Microsoft Br. at 11-12. But the first premise is flawed. It is contradicted by the very authority on which Microsoft relies, which correctly notes that, unlike any genuine attribution statute, "nothing in section 1202 affirmatively requires that an author be credited." U.S. Copyright Office, *Authors, Attribution, and Integrity: Examining Moral Rights in the United States*, 90 (April 2019).[17] Section 1202 prohibits removing CMI. But it does not require adding CMI. So one can comply with it without attributing anything to anyone.

This argument is also inconsistent with the DMCA's text. If attribution were the purpose of Section 1202(b), a statute prohibiting distribution of CMI-less works—Section 1202(b)(3)—would have been entirely sufficient, and a removal-based statute—Section 1202(b)(1)—would have been superfluous. So attribution is not the DMCA's purpose. The purpose is "to provide broad protections to copyright owners." *Mango v. BuzzFeed, Inc.*, 970 F.3d 167, 172 n.2 (2d Cir. 2020). And Congress knows how to grant actual attribution rights when it wants to. It did that

[16] Though such cases would not arise under federal question or diversity jurisdiction (for failure to satisfy the amount-in-controversy requirement), they could be heard in federal court if, for example, the United States is a party, *see, e.g.*, *U.S. v. Spear*, No. 2:22-cv-00439, 2023 WL 6463039, *6-7 (D. Idaho Oct. 4, 2023) (trespass action brought as United States), or through the Court's exercise of supplemental jurisdiction, *see, e.g.*, *Feld v. Feld*, 783 F. Supp. 2d 76, 77 (D.D.C. 2011) (trespass action brought as counterclaim).

[17] Indeed, the absence of any attribution right in 17 U.S.C. § 1202 moved the Copyright Office to urge Congress to enact a new statute to serve the currently unprotected purpose of protecting authors' attribution right. *Id.* at 90-98.

when it enacted the Visual Artists Rights Act ("VARA") just eight years before the DMCA. *See* Pub. L. No. 101-650, 104 Stat. 5128 (1990) (enacting VARA); Pub. L. No. 105-304, 122 Stat. 2860 (1998) (enacting DMCA). VARA expressly grants certain visual artists "rights of attribution and integrity," including the right "to claim authorship of that work." 17 U.S.C. § 106A(a). The DMCA contains nothing similar. So it is irrelevant that attribution might lack sufficient historical pedigree. It is simply not analogous to the DMCA claims at issue here.

Third, Defendants analogize to *TransUnion* to argue that DMCA injuries require dissemination. But they make that argument only by misreading the case. As Defendants would have it, *TransUnion* stands for the broad proposition that some class of injuries—which Defendants do not define, but which apparently includes DMCA claims—requires "dissemination" to be concrete. *See* Microsoft Br. at 9-10; OpenAI Br. at 16-17.

This argument ignores *TransUnion*'s basic approach to standing and ***why*** it required dissemination for the injuries there. According to *TransUnion*, the plaintiff must identify a historical analogue, and the court must then decide if that analogy is close enough to satisfy Article III. *See TransUnion*, 594 U.S. at 424 (holding that concreteness inquiry "asks whether plaintiffs have identified a close historical or common-law analogue for their asserted injury"). There, the plaintiffs argued that the injury they suffered—maintenance of false information in a credit file—was analogous to defamation. *See id*. at 432 ("The plaintiffs contend that this injury bears a 'close relationship' to a harm traditionally recognized as providing a basis for a lawsuit in American courts—namely, the reputational harm associated with the tort of defamation."). Because the plaintiffs analogized to defamation, the Court considered whether the analogy was close enough. In doing so, it observed that a core element of defamation is "publication," *i.e.* dissemination *See* *id*. at 434 ("Publication is essential to liability in a suit for defamation.") (cleaned up). So it held that an Article III injury analogous to defamation requires publication too. *See id*.

If CIR were analogizing its DMCA claims to defamation, *TransUnion* would require Defendants to have publicly disseminated CIR's works. But CIR is not analogizing to defamation. It is analogizing to copyright infringement, which, as shown above, does not require dissemination. And *TransUnion* is clear that the ***plaintiff***, not the defendant, chooses the analogy the court must consider. *See id*. at 424.[18] Nor do Defendants offer any reason to analogize DMCA injuries to defamation, which involves an unprivileged false and defamatory statement of and concerning a person, made with at least negligence, that causes reputational harm. *See* Restatement (Second) of Torts, § 558. That is not remotely analogous to Section 1202(b) injuries. Why not, for that matter, analogize them to battery, intentional infliction of emotional distress, or breach of contract? Copyright infringement is the best analogy. And under that analogy, CIR has standing.

ii. CIR has standing because Defendants unlawfully profited from removing CMI.

CIR also has standing for an independent reason: Defendants realized profits from removing CMI from its articles. Had Defendants trained ChatGPT and Copilot on copies of articles that included CMI—such as copyright notices and terms of use—the chatbots would have falsely learned that ordinary English writers convey CMI in situations when they do not. FAC ¶ 119. That would have been incompatible with Defendants' purpose in training ChatGPT and Copilot: to replicate how people normally use English. *Id*. The only alternative would be for Defendants to retrain ChatGPT and Copilot not to produce CMI in unconventional situations. *Id*.

[18] Microsoft's appellate case, *Dinerstein v. Google, LLC*, 73 F.4th 502 (7th Cir. 2023), confirms this approach. There the court spent considerable effort determining what analogy the plaintiff was attempting to draw—including asking which analogies he abandoned at what stage of the proceeding—before rejecting the analogy between a state law privacy harm and the tort of publicity given to private life. *See id*. at 512-14.

But that alternative would require more computing resources and thus be more expensive. *Id.* Thus, removing CIR's CMI helped Defendants reap a financial windfall. *Id.*

The DMCA allows plaintiffs to recover profits defendants obtain by violating the statute. *See* 17 U.S.C. § 1203(c)(1)(A). And as the Second Circuit has held—and reaffirmed after *TransUnion*—a defendant's unlawful profits, which are the plaintiff's by law, constitute an Article III injury whether or not they correspond to an economic loss. *See Donoghue v. Bulldog Invs. Gen. P'ship*, 696 F.3d 170, 175-80 (2d Cir. 2012); *Packer on behalf of 1-800-Flowers.Com, Inc. v. Raging Cap. Mgmt., LLC*, 105 F.4th 46, 51-56 (2d Cir. 2024) (reaffirming *Donoghue*).[19] That result applies here given the analogy between DMCA injuries and copyright infringement and the fact that, for over a century, profits have been an available remedy for copyright infringement absent any further showing. *See* 17 U.S.C. § 504(b) (providing for recovery of profits); Copyright Act of 1909, Pub. L. 60-349, ch. 320, 35 Stat. 1075, 17 U.S.C. §101(b) (1909) (same); *see also Packer*, 105 F.4th at 54 (upholding lost profit as Article III harm based on historical analogy). So CIR has standing because it is entitled to Defendants' unlawful profits.

iii. CIR has alleged a dissemination-based injury, or at least deserves jurisdictional discovery.

As argued above, CIR's DMCA injuries do not require public dissemination of CMI-less works. But if they do, CIR has plausibly alleged it, or at least deserves jurisdictional discovery. As CIR alleges, by removing CMI during training, Defendants caused ChatGPT and Copilot to omit CMI from responses that include CIR's copyright-protected articles. This is because if the

[19] As *Donoghue* noted, courts have required economic loss for disgorgement actions brought under ERISA. *Donoghue* distinguished the ERISA cases on the ground that the statute at issue there, unlike ERISA, "is not general, but confers a specific right on" the plaintiff. *Donoghue*, 696 F.3d at 178. DMCA violations are analogous to the statute at issue in *Donoghue*, and unlike ERISA, since the DMCA confers rights on specific individuals. *See* 17 U.S.C. § 1202(b); 17 U.S.C. § 1203. Plus, even in ERISA cases, profits can ground standing to seek an injunction, which CIR seeks here. *See Faber v. Metro. Life Ins. Co.*, 648 F.3d 98, 102 (2d Cir. 2011).

models were trained on CMI, they would have learned to include CMI in their outputs. FAC ¶ 101. The omission of CMI thus fails to inform users of the facts the CMI is necessary to convey: the article's title, who wrote it, who owns the copyright, and the terms of use. Defendants do not appear to dispute that depriving users of information conveyed by CMI would work an Article III injury. OpenAI instead resists dissemination-based standing on three grounds.

First, OpenAI argues that the injury is not concrete. According to OpenAI, CIR's examples supposedly show that "[a]ny user who elicited such an output would already have access to the original article—and thus would already be well aware of its provenance." OpenAI Br. at 18. That is false. Consider Copilot Example 4 from Exhibit 11. There the prompt firsts asks Copilot to list all the "articles in the current issue of Mother Jones," and then asks Copilot to summarize one of them. FAC Ex. 11 at 13-14. Such a user obviously need not possess the article.

Further, CMI includes not just the article's "provenance," but other information, including who currently owns the copyright to it, and the terms of use—information omitted from all the examples in FAC Exhibits 10 and 11. OpenAI claims that the absence of this information "is entirely without real-world consequence." OpenAI Br. at 18. But that is both false and unsurprising in light of OpenAI's disregard for others' copyright. Only a copyright notice—not the author or title—could inform the user that the work is copyrighted, who owns the copyright, and thus whose permission the user needs to make further use of the work. Likewise, the terms of use tell users what they are allowed to do with the article. Without them, the user cannot know whether CIR has prohibited further copying, allowed it for certain purposes, or released the work to the public domain. This is surely the reason Congress elected to include this information as CMI and mandate its preservation. Omitting this information works a concrete harm.

If the Court concludes that CIR's examples do not show a concrete harm, it should at least allow jurisdictional discovery into how ChatGPT and Copilot users have prompted Defendants' products. Discovery is often an appropriate device for assessing subject matter jurisdiction. *See Amidax Trading Grp. v. S.W.I.F.T. SCRL*, 671 F.3d 140, 149 (2d Cir. 2011) (holding that "a court should take care to give the plaintiff ample opportunity to secure and present evidence relevant to the existence of jurisdiction"). It is especially appropriate "where the facts are peculiarly within the knowledge of the opposing party." *Kamen v. Am. Tel. & Tel. Co.*, 791 F.2d 1006, 1011 (2d Cir. 1986). *Compare, e.g.*, *Amidax*, 671 F.3d at 149 (affirming denial of jurisdictional discovery because the plaintiff, "not the defendants, is in control of the relevant jurisdictional evidence") *with Mercer v. Jericho Hotels, LLC*, No. 19-cv-5604, 2019 WL 6117317, at *3 (S.D.N.Y. Nov. 18, 2019) (granting jurisdictional discovery on voluntary cessation exception to mootness because only the defendant knew its "substantive plans" for the future). It is warranted here, since how users have prompted ChatGPT and Copilot are entirely within Defendants' possession.

Second, OpenAI disputes that CIR alleges a causal connection between its removal of CMI and the dissemination of CMI-less outputs. It argues that ChatGPT did not produce CMI because CIR's prompts "forbade" it from doing so and thus that the absence of CMI in the outputs was caused by the nature of CIR's prompts. *See* OpenAI Br. at 19. But this argument ignores the allegations that OpenAI deprived ChatGPT of CMI in the first place by removing CMI from articles in the training sets. *See* FAC ¶¶ 59-68 101. Clearly, depriving ChatGPT of CMI caused ChatGPT not to produce CMI. Plus, CIR's prompts did not forbid ChatGPT from producing CMI. OpenAI points out that some prompts asked for "continuations" of articles. But copyright notices and terms of use are at the ***end*** and thus were not forbidden by a request for continuations.

Third, OpenAI argues that CIR's injury is not actual or imminent because its prompts are not realistic enough to have been employed by a user. *See* OpenAI Br. at 19-20. But what prompts users have employed is a classic fact question requiring jurisdictional discovery, since the prompts and responses are solely within OpenAI's possession. FAC ¶ 80.

Further, OpenAI admits in a footnote that one of CIR's prompts ***is*** sufficiently realistic: one that asked ChatGPT to provide a "summary [of the article] first followed by the actual text," in response to which ChatGPT reproduced nine consecutive paragraphs verbatim (minus one sentence). FAC Ex. 11 at 7. As to that prompt, OpenAI argues that it demonstrates no concrete injury because ChatGPT provided a link to the Mother Jones website, which would supposedly leave "no doubt about the provenance of the information included therein." OpenAI Br. at 20 n.19. But this argument fails on its own terms, and exemplifies the confusion that can result from removing CMI, as OpenAI did here. After all, while the ChatGPT response does link to the Mother Jones website, it also links to a different website called "Snippings," which CIR does not own, and which contains a summary of the original article that omits copyright notice, terms of use, and author information.[20] In fact "Snippings" ***does*** provide the name of ***an*** author—just not the author who wrote the CIR article.[21] Since ChatGPT has linked to two different webpages with two different sets of CMI, it confuses the user as to which CMI is accurate. So the article's provenance is far from clear. Plus, not all ChatGPT responses provide links to the original article at all. *See, e.g.*, FAC Ex. 11 at 5 (ChatGPT Example 3). Viewed in the light most favorable to CIR, these prompts alone show an injury that is actual, imminent, and concrete.

[20] *See* FAC Ex. 11 at 7 (linking to David Birch, The Dark Side of the $100 Bill – Mother Jones, *Snippings* (Feb. 7, 2024), https://snippings.home.blog/2024/02/07/the-dark-side-of-the-100-bill-mother-jones/).

[21] *Compare* Oliver Bullough, The Dark Side of the $100 Bill, *Mother Jones* (Jan-Feb. 2024), https://www.motherjones.com/politics/2024/01/100-bill-crime-corruption-treasury-tax-evasion/ *with* David Birch, The Dark Side of the $100 Bill – Mother Jones, *Snippings* (Feb. 7, 2024), https://snippings.home.blog/2024/02/07/the-dark-side-of-the-100-bill-mother-jones/.

In sum, CIR satisfies Article III by alleging Defendants' interference with CIR's copyright-protected property in a way analogous to copyright infringement and profits from their unlawful removal of CMI. In the alternative, it has shown injury from the dissemination of its articles without CMI. Any questions that remain on CIR's standing are best answered through discovery.

2. CIR has statutory standing.

OpenAI argues that CIR lacks statutory standing because it has not satisfied the DMCA's supposed "distinct injury requirement." OpenAI Br. at 20. But this "distinct requirement" is simply a section of the DMCA that creates a right of action for violations of the statute's substantive provisions. *See* 17 U.S.C. § 1203(a) ("Any person injured by a violation of section 1201 or 1202 may bring a civil action"). This is clear from the DMCA Senate Report, which states that this section "sets forth the general proposition that civil remedies are available for violations of sections 1201 and 1202." S. Rep. 105-190, at 38 (1998).

Neither the text of Section 1203(a) nor OpenAI's cited cases even suggests that "injury" means anything different than it does under Article III. In *Steele v. Bongiovi*, 784 F. Supp. 2d 94, 98 (D. Mass. 2011), the plaintiff argued the defendant's DMCA violation injured him by causing him to lose a separate copyright case. The court rejected that argument on its own terms, holding that the plaintiff would have lost the case regardless. *See id.* Likewise, the plaintiff in *Alan Ross Mach. Corp. v. Machinio Corp.*, No. 17-cv-3569, 2019 WL 1317664, *4 (N.D. Ill. Mar. 22, 2019) argued that the DMCA violation injured him by causing market confusion, and the court held that no confusion resulted. There is no difference between "injury" under Article III and "injury" under Section 1203(a). Because CIR has Article III standing, it also has statutory standing.

3. CIR states 1202(b)(1) claims.

A violation of section 1202(b)(1) requires "(1) the existence of CMI on the allegedly infringed work, (2) the removal or alteration of that information and (3) that the removal was intentional." *Fischer v. Forrest*, 968 F.3d 216, 223 (2d Cir. 2020). It also requires defendant to know, or have "reasonable grounds to know," that removal "will induce, enable, facilitate, or conceal" infringement. 17 U.S.C. § 1202(b)(1).

Apart from issues concerning Microsoft's involvement, discussed in Section IV.C.5, *infra*, Defendants rightly do not contest the adequacy of CIR's allegations as to the first three elements. For the first, CIR has alleged that its works are conveyed with author, title, copyright notice and terms of use information, all of which are forms of CMI. FAC ¶¶ 62, 65-66, 102; *see also* 17 U.S.C. § 1202(c) (defining CMI). For the second and third, CIR has alleged, among other things, that Defendants copied its articles using two specific computer algorithms—Dragnet and Newspaper—that are deliberately designed to omit these types of CMI. FAC ¶¶ 59-69; 120-25.

The fourth element is the second prong of the DMCA's "double-scienter" requirement. *Mango*, 970 F.3d at 171. (The first prong is the uncontested intent element.) The "scienter" label is significant because "[t]he Second Circuit has stated that courts should be lenient in allowing scienter issues to survive motions to dismiss." *Aaberg v. Francesca's Collections, Inc.*, No. 17-cv-115, 2018 WL 1583037, at *9 (S.D.N.Y. Mar. 27, 2018) (citing *In re DDAVP Direct Purchaser Antitrust Litig.*, 585 F.3d 677, 693 (2d Cir. 2009)). Thus, courts in this District have allowed DMCA claims to proceed based even on "sparse" allegations of scienter. *Hirsch v. CBS Broad. Inc.*, No. 17-cv-1860, 2017 WL 3393845, at *8 (S.D.N.Y. Aug. 4, 2017); *see also Devocean Jewelry LLC v. Associated Newspapers Ltd.*, No. 16-cv-2150, 2016 WL 6135662, at *2 (S.D.N.Y. Oct. 19, 2016) ("relatively sparse"). This reflects the commonsense proposition that a defendant's knowledge is typically in its own possession and a plaintiff cannot learn it without discovery.

Defendants argue that *Stevens v. Corelogic, Inc.*, 899 F.3d 666, 675 (9th Cir. 2018) sets a higher standard by requiring an objective likelihood of infringement. Microsoft Br. at 17-18. But

Stevens was decided at summary judgment and so does not address pleading standards. Thus, courts in the Ninth Circuit have found it "inapposite" at the dismissal stage, as resolution of scienter "is more suited to summary judgment." *Izmo, Inc. v. Roadster, Inc.*, No. 18-cv-06092, 2019 WL 13210561, at *4 (N.D. Cal. Mar. 26, 2019); *see also Doe 1 v. Github*, 672 F. Supp. 3d 837, 858 (N.D. Cal. 2023) (holding that *Stevens*, as a summary judgment case, does not alter the rule that, at the pleading stage, "mental conditions generally need not be alleged with specificity").

CIR's pleadings satisfy the proper standard by identifying four reasons for which Defendants knew, or had reason to know, that removing CMI would induce, enable, facilitate, or conceal infringement.[22]

Concealing Defendants' training-based infringement from their users. Defendants had reason to know that removing CMI during training would conceal their training-based infringement from users. A ChatGPT or Copilot response that ***contained*** CIR's CMI would inform the user that the response blossomed out of an infringing copy in the training set. After all, "responses are the product of its training sets." FAC ¶ 113. Thus, a response that ***omits*** CMI conceals that training-based infringement. And, critically, ChatGPT and Copilot omit CMI precisely because Defendants removed the CMI during training: if they were trained on CMI, they would have produced CMI. FAC ¶ 101.

Defendants respond primarily by reprising a standing argument: that if a ChatGPT response lacks CMI, it must be because the prompt did not ask for CMI, or forbade it from producing CMI by asking for a "continuation." OpenAI Br. at 23-24. This response fails for the same reasons. If

[22] OpenAI suggests, but does not directly argue, that the second scienter element requires the defendant to remove CMI with the ***purpose*** of abetting or concealing infringement. *See* OpenAI Br. at 23 (disputing CIR's purported allegation that OpenAI "removed CMI ***in order to*** 'conceal' the alleged infringement) (emphasis added). But there is no purpose requirement. The text is clear that the second scienter element involves only actual or constructive ***knowledge***. *See* 17 U.S.C. § 1202(b)(1).

ChatGPT never ***had*** the CMI—because Defendants removed it from the start—then it could not ***produce*** the CMI. Defendants plausibly knew as much, since they deliberately used text extraction algorithms designed not to copy CMI into the training data. FAC ¶¶ 59-68. Plus, none of CIR's prompts actually forbade ChatGPT from producing CMI. The prompts that asked for continuations did not forbid copyright notices and terms of use. And some CMI-less responses stemmed from prompts that did not ask for a continuation at all. *See, e.g.*, FAC Ex. 11 at 7, 14.

OpenAI also says that CIR has not alleged that "including CMI in training datasets would cause ChatGPT to include that CMI in outputs." OpenAI Br. at 24. That is false. CIR alleges that "[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." FAC ¶ 101. It also alleges that had ChatGPT and Copilot been trained on CMI, "they would have falsely learned that ordinary English speakers convey copyright management information in situations when they do not," and thus produced that CMI in unconventional situations. *Id.* ¶ 119.[23]

Concealing Defendants' output-based infringement from their users. CIR next alleges that removing CMI during training concealed the infringement Defendants committed by producing unauthorized regurgitations and abridgements. Had Defendants included the CMI in their training data, ChatGPT and Copilot would have produced that CMI in their regurgitations and abridgements. FAC ¶ 101. This in turn conceals from users that ChatGPT or Copilot output infringes CIR's copyright.

[23] Contrary to OpenAI's suggestion, *see* OpenAI Br. at 22, CIR has not alleged that OpenAI violated Section 1202(b)(1) merely by removing CMI from unpublished training data, without more. OpenAI's citation to *Tremblay v. OpenAI, Inc*., 716 F. Supp. 3d 772 (N.D. Cal. 2024), which rejected such a theory, is therefore inapposite.

Apart from the same causal arguments just rejected, *see* OpenAI Br. at 23-24, Defendants' only response is to doubt that these outputs "[h]appen in the real world" or that they "constitute copyright infringement at all." Microsoft Br. at 19. But no defendant has moved to dismiss CIR's output-based infringement claims in full, and Microsoft has not moved to dismiss them at all. Since CIR has pleaded Defendants' output infringement as a standalone claim, FAC ¶ 130-32, 138-40, it has necessarily pleaded likely output infringement as an element of its DMCA claims.

Inducing, enabling, or facilitating users to infringe. The FAC next alleges that Defendants knew, or had reason know, that removing CMI would induce, enable, or facilitate users to infringe CIR's copyright by distributing infringing responses to a future audience. Defendants know users distribute ChatGPT and Copilot responses this way—they even advertise ChatGPT for that purpose. FAC ¶ 115. Defendants know their users are capable of infringing, including when the users deploy their products as advertised—they have even committed to indemnifying their commercial clients in just these circumstances. *Id.* ¶ 118. They also at least have reason to know that removing CMI from training data—causing it not to be included in outputs—would induce users to infringe. That is because some people respect copyright or fear liability. *Id.* ¶¶ 116-17.

Defendants respond on two bases. First, they falsely claim that CIR's prompts led to responses too "trivial" for users to distribute, which are supposedly limited to "a couple-dozen words … or a bullet-point summary." Microsoft Br. at 19-20. This ignores both that a couple dozen words can qualify for protection from infringement, *see Enterprise Mgmt. Ltd. v. Warrick*, 717 F.3d 1112, 1115, 1119 (10th Cir. 2013), and CIR's allegations of fuller responses, including two that reproduced many paragraphs of text verbatim or near verbatim, FAC Ex. 11 at 7, 14; Match Decl. Exs. 1, 2.

Second, Defendants doubt that CMI would cause a user with a work's title to behave any differently. Microsoft Br. at 20. This argument also fails for reasons discussed above. A work's title does not convey who owns the copyright or what its permitted use are—information contained only in copyright notice and terms of use. And a user who knows that a work is copyrighted, and the owner has not allowed them to copy the work, will plausibly behave differently than a user without that information. Nothing more is required at the dismissal stage.

Facilitating Defendants' training-based infringement. CIR last alleges that by removing CMI, Defendants facilitate their own training-based infringement. Through training, Defendants aim for ChatGPT and Copilot to mimic the expression contained in the works on which they were trained. Except when referring to copyrighted material, ordinary English writers do not include CMI—especially copyright notices and terms of use. But because of the prevalence of CMI in the content on which ChatGPT and Copilot were trained, including CMI in the training data would have caused them to falsely learn otherwise. Removing CMI at the outset, then, prevented ChatGPT and Copilot from learning false things about how people use English. FAC ¶ 119.

Defendants respond that facilitating training does not count as "facilitat[ion of] … an infringement" under section 1202(b)(1). According to them, the infringement is not the training, but copying narrowly described, which Defendants conducted as part of the training process. And because CIR has not alleged that CMI removal facilitates copying narrowly described, CIR has not adequately alleged facilitation. *See* Microsoft Br. at 18; OpenAI Br. at 22-23.

This argument defines the infringing act at too narrow a level of generality. This case and others like it are fundamentally about whether Defendants infringe when they copy news articles and use those copies to train LLMs. Defendants have described the purportedly infringing acts in this way. *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"); Defendant Microsoft Corporation's Memorandum of Law in Support of Partial Motion to Dismiss the Complaint, at 2, *The New York Times Company v. Microsoft Corp.*, No. 23-cv-11195 (S.D.N.Y. Mar. 3, 2024), ECF No. 65 (describing infringement claims as being about "building the GPT-based models"). Understood at the proper level of generality, the infringing act is LLM training, whose facilitation CIR has alleged.

If, *arguendo*, facilitation must aid copying narrowly described, CIR respectfully requests leave to replead additional facts. CIR would allege, at a minimum, that removing CMI facilitates Defendants' copying by freeing up computer storage for more unlawful copies. This is especially plausible considering the sheer size of the training data. Since GPT-4 was trained on 1.8 trillion parameters, FAC ¶ 122, removing CMI from every copy would save substantial storage space and allow that space to be occupied by more infringing copies. Leave to replead shall be "freely given when justice so requires," *Cortec Indus., Inc. v. Sum Holding L.P.*, 949 F.2d 42, 48 (2d Cir. 1991), as it would here given the novelty and complexity of the issues. Of course, it will be unnecessary if the Court holds, as it should, that CIR has adequately alleged scienter.

4. CIR states 1202(b)(3) claims.

A Section 1202(b)(3) claim has four elements: (1) the existence of CMI in connection with a copyrighted work; (2) that the defendant distributed works or copies of the work; (3) that the defendant knew CMI had been removed; and (4) while knowing or having reason to know that such distribution will induce, enable, facilitate, or conceal an infringement. *Mango*, 970 F.3d at 171. Here, CIR alleges that Defendants violated Section 1202(b)(3) by distributing training data, from which CMI had been removed, with each other. FAC ¶¶ 156-57; 168-69. Defendants do not dispute the first and third elements, which CIR has alleged for reasons already discussed.[24] They only dispute the second and fourth.

As to the second element, CIR has plausibly alleged distribution of its CMI-less works between OpenAI and Microsoft. Microsoft's CEO, Satya Nadella, said, "we have the data," giving rise to the plausible inference that Microsoft does, in fact have the data, especially in conjunction with Microsoft's close partnership with OpenAI, its massive investment in the company, and its provision of the data center and bespoke supercomputing infrastructure that powers ChatGPT. FAC ¶¶ 26-29. And if Microsoft has the data, which OpenAI played a part in developing, then it plausibly got the data from OpenAI. So OpenAI and Microsoft exchanged the data.

Microsoft reads Mr. Nadella's quote differently as saying that Microsoft has the "rights" to the data, not the data itself. Microsoft Br. at 15. True, Mr. Nadella mentions rights elsewhere in the interview. But in the same ***sentence***, he also says, "we have the people, we have the compute." 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. This suggests Mr. Nadella was talking about what Microsoft physically has, not what it rightfully has. At worst, Mr. Nadella's statement is ambiguous, and ambiguities at the dismissal stage must be resolved in CIR's favor. *See Shultz v. Congregation Shearith Israel of City of New York*, 867 F.3d 298, 302 (2d Cir. 2017). Regardless, distribution would be plausible even without this interview given the close working relationship between the two companies, Microsoft's 49 percent stake in OpenAI, and its provision of the data centers and bespoke supercomputing infrastructure used to train ChatGPT. FAC ¶¶ 26-28. *Contra* Microsoft, the allegation is not that Microsoft is a passive receptacle of the data, but an active

[24] On the third element, since Defendants removed the CMI, they knew CMI was removed.

participant in the training process. *See* Microsoft Br. at 16. And while OpenAI notes that the FAC does not include certain details about the "when, why, or how" of the distribution, such details are not required at the pleading stage. *See Pilla v. Gilat*, No. 19-cv-2255, 2020 WL 1309086, at *12 (S.D.N.Y. Mar. 19, 2020) ("Although Plaintiff does not allege how, when, or where this removal occurred, such details are not necessary at the pleading stage for a claim under the DMCA.").

OpenAI also argues that CIR fails to plead that CMI-less copies it and Microsoft exchanged are completely identical to the originals. *See* OpenAI Br. at 24-25 That is false. While ***some*** copies contain minute differences, such as the omission of a description associated with an embedded photo, other copies are "completely identical to the original." FAC ¶¶ 65-66, 73-74. The Court can verify this by comparing original articles with certain exhibits to the First Amended Complaint, which CIR created by applying the same CMI-removing algorithms OpenAI admitted to using. *See* FAC ¶¶ 59, 65-66, 73-74. *Compare, e.g.*, FAC Ex. 8 at 10 (Mother Jones Example 4) *with* Andy Kroll, Tales from the Debt Collection Crypt, *Mother Jones* (Jan. 3, 2011), https://www.motherjones.com/criminal-justice/2011/01/tales-debt-collection-crypt/; *compare* FAC Ex. 8 at 28-29 (Reveal Example 4) *with* Aaron Glanz, Trump administration seeks to legalize payments to VA officials by for-profit schools, *Reveal* (Sept. 14, 2017), https://revealnews.org/blog/trump-administration-seeks-to-legalize-payoffs-to-va-officials-by-for-profit-schools/. And in any case there is no basis to require identicality. *See ADR Int'l Ltd. v. Inst. for Supply Mgmt. Inc.*, 667 F. Supp. 3d 411, 424-31 (S.D. Tex. 2023) (analyzing the issue and concluding that the DMCA has no identicality requirement); *We Protesters, Inc. v. Sinyangwe*, No. 22-cv-9565, 2024 WL 1195417, at *9 (S.D.N.Y. Mar. 20, 2024) (holding that "close to identical" is enough).[25] This argument fails. CIR has alleged distribution.

[25] Despite concluding otherwise, the Northern District of California recently certified the issue for interlocutory appeal because it recognized "substantial ground for difference of opinion on the question." Order Granting Motion

CIR has also alleged scienter in light of the lenient pleading standards for DMCA claims: that OpenAI knew or had reason to know that its distribution of CMI-less works would induce, enable, or facilitate Microsoft's infringement, and vice-versa. By distributing CMI-less works to each other, Defendants enabled each other to use those works to infringe CIR's copyright by training their models and producing infringing regurgitations and excerpts to their users.

OpenAI responds by misreading the statute. According to it, the "absence of CMI" must conceal, induce, enable, or facilitate infringement, and CIR supposedly has not alleged that the absence of CMI on works it distributed to Microsoft would conceal or abet Microsoft's infringement. OpenAI Br. at 25. But that is not what the statute says. Instead, what must abet or conceal is the ***distribution*** of CMI-less works. *See* 17 U.S.C. § 1202(b)(3) (prohibiting the "distribut[ion]" of CMI-less works "knowing [or] having reasonable grounds to know" that "it will" contribute to infringement). And clearly, providing Microsoft with new CMI-less works would abet Microsoft's infringement of those works.

CIR also alleges scienter on OpenAI's interpretation. For reasons discussed above, OpenAI had reason to know that removing CMI—and thus the absence of CMI—would abet and conceal infringement by both itself and its users. *See* Section IV.C.3, *supra*. Since OpenAI had reason to know this for its ***own*** use of CMI-less training data, it also had reason to know this for ***Microsoft's*** use of CMI-less training data. So because CIR alleges scienter for Section 1202(b)(1), it also alleges scienter for section 1202(b)(3). CIR states a claim.

to Certify Order for Interlocutory Appeal and Motion to Stay Pending Appeal, at 2, *Doe 1 v. Github, Inc.*, No. 22-cv-06823 (N.D. Cal. Sept. 27, 2024), ECF No. 282. This Court need not resolve the issue at this juncture since CIR has alleged complete identicality for at least some distributions.

5. CIR alleges Microsoft's involvement in LLM training.

Microsoft objects that the DMCA claims should be dismissed because CIR has not plausibly alleged CMI removal and distribution by Microsoft, as opposed to OpenAI. *See* Microsoft Br. at 14. But it neglects CIR's allegations that Copilot outputs lack CMI, which for reasons explained above plausibly results from the removal of CMI during training. FAC ¶¶ 82, 119; FAC Ex. 11 at 14; *see also* Section IV.C.3, *supra*. Of course, some of CIR's allegations derive from admissions made by OpenAI, including those about the algorithms it used to assemble its training sets. But the First Amended Complaint identifies no contrary admissions by Microsoft, and given the companies' close working relationship with respect to LLM training, it is at least plausible that Microsoft used similar methods to OpenAI.

Microsoft does rightly note that CIR's allegations are fuller as against OpenAI. *See* Microsoft Br. at 14. CIR culled those details largely from public admissions by a younger, formerly nonprofit company with "Open" in its name and different past objectives—much to its present chagrin. But Microsoft has always had the incentive and the business sense to keep similar details secret. So for its claims against Microsoft CIR relies on OpenAI's public statements, Mr. Nadella's admission to possessing the data, Microsoft's massive investment in OpenAI, its provision of infrastructure, and the companies' close relationship. *See* Section IV.C.4, *supra*.

Last, it bears noting that while Microsoft has moved to dismiss the DMCA claims against it, it has not sought dismissal of CIR's direct infringement claim, including as to LLM training. This is significant because these claims are based on many of the same allegations, such as those concerning how Defendants assembled their training sets. It is of course possible that Microsoft trained Copilot with removing CMI. Microsoft is at full liberty to prove that. If so, it may prevail on summary judgment. But CIR has alleged enough to warrant discovery into Microsoft's training practices. And that requires denying Microsoft's motion.

V. CONCLUSION

The Court should deny Defendants' motions to dismiss.

RESPECTFULLY SUBMITTED,

/s/ Stephen Stich Match

Jon Loevy (*pro hac vice*)
Michael Kanovitz (*pro hac vice*)
Lauren Carbajal (*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 (p)
312-243-5902 (f)
jon@loevy.com
mike@loevy.com
carbajal@loevy.com
match@loevy.com
matt@loevy.com
kayes@loevy.com
steve@loevy.com
wallenberg@loevy.com

November 5, 2024