UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

THE INTERCEPT MEDIA, INC., Plaintiff, -v- OPENAI, INC., OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, OPENAI HOLDINGS, LLC, and MICROSOFT CORPORATION, Defendants.

24-cv-1515 (JSR)

OPINION AND ORDER

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

On November 21, 2024, the Court issued a bottom-line order in the above-captioned case. See ECF No. 122. That order granted the motion of defendant Microsoft Corporation to dismiss in full and with prejudice the claims against it brought by the plaintiff, The Intercept Media, Inc. ("The Intercept"), and granted in part the motion of defendants OpenAI[1] to dismiss, dismissing with prejudice plaintiff's claim against OpenAI under 17 U.S.C. § 1202(b)(3), but allowing plaintiff's claim under 17 U.S.C. § 1202(b)(1) to proceed against OpenAI past the motion-to-dismiss stage. This Opinion

[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 Opinion treats them collectively as "OpenAI."

explains the reasons for the Court's decision and orders counsel to call Chambers to set a revised case-management schedule.

I. Factual Background

The relevant factual background of this case depends heavily on the statute that supplies the causes of action stated in The Intercept's complaint. So before turning to the specific factual allegations, some background on that statute is necessary. The Intercept alleges that OpenAI and Microsoft violated provisions of the Digital Millenium Copyright Act ("DMCA") -- specifically, 17 U.S.C. §§ 1202(b)(1) and 1202(b)(3). Those provisions concern so-called "copyright management information" (or "CMI"), which includes, among other material, a work's title, author, and copyright notice. See 17 U.S.C. § 1202(c).

A DMCA claim is different from a traditional copyright claim. Rather than allege, for example, that defendants unlawfully reproduced its copyrighted material, The Intercept claims that OpenAI and Microsoft intentionally removed CMI from its articles and distributed its articles without CMI. Correspondingly, § 1202(b)(1) provides that "[n]o person shall . . . intentionally remove or alter any [CMI]," and § 1202(b)(3) provides that "[n]o person shall . . . distribute . . . works . . . , knowing that [CMI] has been removed or altered."

Section 1202(b) imposes an important additional element. Liability attaches only to those who removed CMI or distributed

works without CMI while "knowing, or, . . . having reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement of any right under this title." 17 U.S.C. § 1202(b). When this element is paired with the requirement of an "intentional[] remov[al]" of CMI under § 1202(b)(1) and "know[ledge]" of CMI's removal under § 1202(b)(3), the statute is said to impose a double scienter requirement. See Mango v. BuzzFeed, Inc., 970 F.3d 167, 171 (2d Cir. 2020).

The facts of this case test the application of these provisions to cutting-edge technology.[2] At base, The Intercept's claims relate to OpenAI's training of its generative artificial intelligence ("AI").[3] OpenAI's generative AI chatbot -- called ChatGPT[4] -- is powered by a "large language model" ("LLM"), which is a deep-learning algorithm that can generate human-language text. Am. Compl. ¶ 34. An LLM's capacity to produce an output stems from "training on works written by humans," often collected in

---

[2] For purposes of outlining the factual background, the Court accepts The Intercept's allegations as true, as it must at the motion-to-dismiss stage. See infra at p. 19.

[3] Generative AI is a type of artificial intelligence that employs models to produce text, images, videos, or other kinds of data, often in response to specific prompts.

[4] At times, The Intercept references Microsoft's AI product, Copilot, too. But the basis for these allegations is generally unsupported, see infra at pp. 25–26; instead, the complaint pleads facts specifically related to OpenAI and ChatGPT. For that reason, the factual background presented above focuses on OpenAI and ChatGPT.

"training sets." Id. ¶ 35. More technically, chosen works in a training set are "encode[d] . . . in computer memory as numbers called 'parameters.'" Id. ¶ 36. The Intercept alleges that thousands of its own works were included in the training sets used to train ChatGPT, see Ex. 2 to Am. Compl., but that OpenAI intentionally omitted CMI -- in particular, author and title information -- from the articles included in those training sets.

Although the content of the training sets is critical to The Intercept's claims, defendants, according to The Intercept, have been "fully secret" about the sets used in the latest version of ChatGPT, GPT-4.[5] Am. Compl. ¶ 37. To overcome this hurdle, The Intercept bases its allegations both on information OpenAI disclosed about training sets used for versions of ChatGPT prior to GPT-4 and on "consultations with a data scientist." Id. Three training sets are described in the complaint: WebText, WebText2, and "sets derived from Common Crawl." Id. ¶ 39.

WebText and WebText2 are comprised of all outbound links on Reddit that have received at least three "karma" (a measure of the amount of engagement on a Reddit post). Id. ¶ 40. In a list published by OpenAI of the top 1,000 web domains in the WebText training set, "6,484 distinct URLs from [The Intercept's] web

[5] OpenAI has issued different versions of ChatGPT, with the trailing number (e.g., the "4" in "GPT-4") identifying the version. A higher number indicates a more recent version. See Am. Compl. ¶ 39.

domain were included." Id. ¶ 41. "[A]n approximation of the WebText dataset" -- an open-source recreation called "OpenWebText" -- yields a similar number: 5,026 distinct URLs. Id. ¶ 43.

More than just suggest that its articles were included in the WebText training sets, The Intercept explicitly alleges that OpenAI and Microsoft removed CMI from its articles that were in those sets. This allegation relies on the process by which an article is converted into a format digestible for the AI. Specifically, OpenAI used "Dragnet and Newspaper" algorithms "to extract text from websites." Id. ¶ 45. Dragnet "'separate[s] the main article content' from other parts of the website, including 'footers' and 'copyright notices,' and allow[s] the extractor to make further copies only of the 'main article content.'" Id. ¶ 46. In other words, Dragnet cannot extract author and title information. And while the Newspaper algorithm allows a user to capture author and title information, The Intercept alleges that OpenAI "chose not to extract [this] information because [it] desired consistency with the Dragnet extractions."[6] Id. ¶ 47.

Emphasizing that these CMI-removal features are common knowledge, The Intercept insists that -- by employing Dragnet and Newspaper -- OpenAI intentionally and knowingly removed CMI when

[6] The complaint further alleges that "Newspaper algorithms are incapable of extracting copyright notices" -- another kind of protected CMI. Id. ¶ 47; see 17 U.S.C. § 1202(c)(3).

creating the WebText training sets. Id. ¶ 50. Further, it alleges, on information and belief, that OpenAI has "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." Id. ¶ 54. The Intercept has attached exhibits that illustrate the format of an article when Dragnet and Newspaper were applied to three URLs listed in OpenWebText. See Exs. 3 & 4 to Am. Compl.

Like the WebText training sets, ChatGPT's Common Crawl sets purportedly include The Intercept's articles without CMI. Common Crawl, a dataset created by a nonprofit, is based on a scrape of most of the Internet. Am. Compl. ¶ 55. The Intercept alleges that OpenAI used a Common Crawl-derived training set similar to one used by Google, which is called "C4," to train its own generative AI models. Id. ¶ 57. A recreation of C4, based on Google's instructions, is published online. Id. The Intercept's data scientist found 2,753 distinct URLs from its web domain in that set. Id. ¶ 58. None of those articles includes copyright notice or terms-of-use information; indeed, the vast majority lack both author and title information. Id. The Intercept has attached an exhibit collecting examples of those articles' appearance in C4. See Ex. 5 to Am. Compl.

The remainder of the complaint aims to show that OpenAI "kn[ew], or, . . . [had] reasonable grounds to know, that [the CMI removal] [would] induce, enable, facilitate, or conceal an

infringement of" The Intercept's copyrights. See § 1202(b). The Intercept has two theories of potential copyright infringement. To start, it alleges that the removal of CMI "enable[s], facilitate[s], and conceal[s]" OpenAI's own copyright infringement in the training process. Am. Compl. ¶ 81. The Intercept argues that by downloading its articles without permission, OpenAI "infringes [its] copyright," namely "the right to control reproductions of copyright-protected works." Id. ¶ 61. In support, The Intercept notes that OpenAI has acknowledged that it needs a license to use copyright-protected works to train ChatGPT, including by entering into agreements with large copyright owners. Id. ¶ 77. Although The Intercept admits that an OpenAI employee explained that "these deals focus on 'the display of news content and use of the tools and tech,' and are thus 'largely not' about training," it interprets the latter part of this quote to "confir[m] that these deals involve training, at least in part." Id. ¶ 79. The Intercept also points to OpenAI's creation of "tools in late 2023 to allow copyright owners to block their work from being incorporated into training sets." Id. ¶ 80.

The second theory of infringement shifts focus from the training inputs to ChatGPT's outputs. The Intercept alleges that "[a]t 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 CMI. Id.

¶ 64. And even when the AI models do not "regurgitate verbatim," they still "mimic significant amounts of material from copyright-protected works of journalism." Id. ¶ 65. Despite OpenAI's adjustment of ChatGPT's settings to "reduce regurgitations for copyright reasons," The Intercept's data scientist still generated three regurgitations of parts of its articles. Id. ¶¶ 66–67; see also Ex. 6 to Am. Compl. The Intercept insists that defendants knew ChatGPT would regurgitate articles without CMI because they "were aware that ChatGPT responses are the product of its training sets and that ChatGPT generally would not know any [CMI] that was not included in training sets." Am. Compl. ¶ 82.

The Intercept points to the CMI itself to try to establish defendants' knowledge of these potential infringements. And so, according to The Intercept, defendants knew that users would be less likely to distribute infringing responses if outputs included CMI because "at least some likely users . . . respect the copyrights of others or fear liability for copyright infringement." Id. ¶ 84. OpenAI's profit motive provides further evidence: ChatGPT might generate less revenue if it added CMI because some users would not subscribe to a service that could result in copyright liability. Id. ¶ 85. Finally, The Intercept references a policy under which, "[i]f a commercial" ChatGPT or Copilot "user . . . is sued for copyright infringement, [d]efendants 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 . . . only if the user uses the product as advertised." Id. ¶ 86. To The Intercept, this policy indicates that "[d]efendants 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 [d]efendants." Id.

Although most of the complaint focuses on OpenAI, The Intercept targets Microsoft too. Central to the allegations against Microsoft is its "partnership with OpenAI": Microsoft has invested "billions of dollars," "will own a 49% stake in the company," and "provides the data center and bespoke supercomputing infrastructure used to train ChatGPT." Id. ¶¶ 20–22. To connect Microsoft to ChatGPT's training sets, The Intercept repeatedly references a quote from Microsoft's CEO: "'If OpenAI disappeared tomorrow,' Microsoft could still 'continue the innovation' alone because, among other reasons, 'we have the data, we have everything.'" Id. ¶ 23. The Intercept adds (on information and belief) that "Microsoft hosts ChatGPT training sets and provides access to those training sets to one or more of the OpenAI [d]efendants, and some of those training sets were created by the OpenAI [d]efendants and provided to Microsoft." Id. ¶ 24. Microsoft, in turn, has "shared copies of [The Intercept's] works" without CMI "with the OpenAI [d]efendants." Id. ¶ 75.

II. Procedural Background

On February 28, 2024, The Intercept sued OpenAI and Microsoft. See ECF No. 1. Around two months later, on April 15, 2024, both OpenAI and Microsoft moved to dismiss The Intercept's complaint, arguing both that The Intercept failed to allege a concrete injury sufficient to establish standing under Article III of the Constitution and that it failed to state a claim under § 1202(b). See ECF Nos. 49, 52; see also ECF Nos. 50, 53. After holding oral argument on defendants' motions on June 3, 2024, the Court issued an order several days later that "determined that [The Intercept] should be granted leave to amend its complaint to attempt to rectify some of the seeming lack of specificity in its current complaint." ECF No. 81 at 1. The Court set a schedule for The Intercept to file an amended complaint and for the parties to submit supplemental briefing related to defendants' motions to dismiss. See id. at 1–2.

On June 21, 2024, The Intercept filed an amended complaint. See ECF No. 87. Microsoft and OpenAI submitted supplemental memoranda of law on July 8, 2024, see ECF Nos. 88, 89, and The Intercept submitted a supplemental response the following week, see ECF No. 90. In late August, OpenAI and The Intercept identified supplemental authorities for the Court in support of their respective positions. See ECF Nos. 99, 100. Around two months later, the Court advised the parties that "there remain several

difficult issues presented by the instant motions that would benefit from oral argument," and set a hearing for November 1, 2024. Ct. Email of Oct. 17, 2024. Ahead of the hearing, the Court identified three issues that the parties should be prepared to address. Ct. Email of Oct. 29, 2024. On November 21, 2024, the Court issued a bottom-line order granting Microsoft's motion to dismiss and granting in part and denying in part OpenAI's motion to dismiss. See ECF No. 122.

III. Analysis

A. Standing

OpenAI and Microsoft argue that The Intercept has failed to demonstrate a concrete injury and thus lacks Article III standing. Their argument is rooted in TransUnion LLC v. Ramirez, 594 U.S. 413, 424 (2021), in which the Supreme Court explained that, "with respect to the concrete-harm requirement," "courts should assess whether the alleged injury to the plaintiff has a close relationship to a harm traditionally recognized as providing a basis for a lawsuit in American courts."[7] The Supreme Court also addressed what types of "intangible harms" are "concrete." Id. at 425. Again, it emphasized that "[c]hief among them are injuries with a close relationship to harms traditionally recognized as

[7] Unless otherwise indicated, case quotations omit all internal quotation marks, alterations, footnotes, and citations.

providing a basis for lawsuits in American courts." Id. And, it added, Congress's word is not enough: when assessing the injury-in-fact requirement, courts must look beyond the fact that "a statute grants a person a statutory right and purports to authorize that person to sue to vindicate that right." Id. at 426.

Contrary to defendants' position, the Intercept has pleaded a concrete injury of a kind long protected by American courts. Copyright claims predate the Constitution's ratification, see The Federalist No. 43 (James Madison) ("The copyright of authors has been solemnly adjudged, in Great Britain, to be a right of common law."), and the Constitution itself includes copyright among Congress's list of enumerated powers in Article I, see U.S. Const. art. I, § 8, cl. 8. Indeed, soon after ratification, Congress enacted the Copyright Act of 1790. See Copyright Act of 1790 § 2 (recognizing a claim for infringement). Congress has repeatedly updated the copyright laws over the past two centuries. See Copyright Act of 1790; Copyright Act of 1831; Copyright Act of 1870; Copyright Act of 1909; Copyright Act of 1976.

Defendants counter that a DMCA claim differs from a traditional copyright claim. True, copyright claims protect against infringements of the exclusive rights over works granted by the Copyright Act, see 17 U.S.C. § 106; by contrast, DMCA claims seek to protect those works' CMI. But, as noted by the Court in TransUnion, Article III "does not require an exact duplicate in

American history and tradition." TransUnion, 594 U.S. at 424. In slightly different terms, the Court does not evaluate whether there has always been a legal basis for a copyright holder to sue for the removal of CMI on its works. Cf. Saba Cap. Cef Opportunities 1, Ltd. v. Nuveen Floating Rate Income Fund, 88 F.4th 103, 116 (2d Cir. 2023) ("The question is not, as [defendant] frames it, whether there was always a common-law basis to sue for a lack of equal voting rights for every share -- or in other words, an injury in law."). Instead, "[t]he correct question is whether there is a 'close historical or common-law analogue' for [The Intercept's] 'asserted injury' -- its 'injury in fact.'" Id. at 116–17 (quoting TransUnion, 594 U.S. at 424–27).

Here, The Intercept's injury under the DMCA is similar to the harm traditionally actionable in copyright. The individual harm forming the basis of Founding Era copyright suits was grounded in notions of "property rights." See Oren Bracha, The Ideology of Authorship Revisited: Authors, Markets, and Liberal Values in Early American Copyright, 118 Yale L.J. 186, 199 (2008); see also id. at 224 ("Within the late eighteenth-century conception of authorship, authors were envisioned as having property rights in their intellectual creations. Copyright was thus reimagined as ownership -- that is to say, total control -- over an intangible object of property."); see also Dowling v. United States, 473 U.S. 207, 216–17 (1985) (distinguishing "the property rights of a

copyright holder" from other possessory interests); Comm'r v. Wodehose, 337 U.S. 369, 401 (1949) (Frankfurter, J., dissenting) ("In the exercise of its power 'To promote the Progress of Science and useful Arts,' Congress, by granting copyrights, has created valuable property rights.").

Those property rights, in turn, are designed to encourage creative production. See Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 450 (1984) ("The purpose of copyright is to create incentives for creative effort."). Indeed, this "incentive" justification for copyright finds support in the Constitution, U.S. Const. art. I, § 8, cl. 8 ("The Congress shall have Power . . . To promote the Progress of . . . useful Arts, by securing for limited Times to Authors . . . the exclusive Right to their respective Writings . . . ."") (emphases added), and, according to some scholars, even predates ratification, see Jane C. Ginsburg, Creation and Commercial Value: Copyright Protection of Works of Information, 90 Colum. L. Rev. 1865, 1866 n.3 (1990) (noting, after describing the economic justification for copyright in the Constitution, that "[a] similar policy underlay the English Statute of Anne of 1710, titled 'An Act for the Encouragement of Learning, by Vesting the Copies of Printed Books in the Authors or Purchasers of such Copies'").

The Intercept's claims in this case implicate the same kind of property-based harms traditionally actionable in copyright. The

DMCA adds another stick to the bundle of property rights already guaranteed to an author in her work under traditional copyright law. Cf. Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 546 (1985) ("Section 106 of the Copyright Act confers a bundle of exclusive rights to the owner of the copyright."). The fact that the specific right at issue here is not expressly rooted in that overall history misses the point; the exact contours of the property rights given to a copyright holder are not frozen in time by the Copyright Act of 1790. To the contrary, Congress has, responding to technological change, supplemented and changed those rights regularly over time. See, e.g., Act of Feb. 3, 1831, ch. 16, 4 Stat. 436 (extending protection of copyright law to musical compositions); Act of July 8, 1870, ch. 230, § 86, 16 Stat. 198, 212 (giving authors the right to create their own derivative works). This practice aligns with the Second Circuit's observation that "Congress frequently elevates to the status of legally cognizable injuries concrete, de facto injuries that were previously inadequate in law." Saba, 88 F.4th at 116. Put differently, the constant in copyright is an author's property right in his original work of authorship; by contrast, the author's specific interests in his work evolve over time. Other courts have described the DMCA as this kind of evolution. See Murphy v. Millenium Radio Grp. LLC, 650 F.3d 295, 303 (3d Cir. 2011) ("[I]t is undisputed that the DMCA was intended to expand -- in some cases

. . . significantly -- the rights of copyright owners."); Mango, 970 F.3d at 172 n.2 (explaining that "the purpose of the DMCA . . . is to provide broad protections to copyright owners").

Likewise, the harm faced by The Intercept -- in the form of defendants' alleged interference with its property right -- implicates the same incentives to create that justify traditional copyright. This is the key contribution of the requirement that a defendant "kno[w], or, . . . hav[e] reasonable grounds to know, that [the CMI removal] will induce, enable, facilitate, or conceal an infringement of" The Intercept's copyright. This element ensures that any violation of the DMCA is tied to concerns of downstream infringement. The increased possibility of infringement makes it more likely that The Intercept (or some other publication) will no longer find it worthwhile to create new articles. To be sure, this specific "harm" is not always felt directly by The Intercept. But the Founders recognized that copyright, in addition to protecting the "claims of individuals," serves the "public good." The Federalist No. 43. And, though the Court is concerned only with The Intercept's standing in this case, the close relationship of its property-based harm to the policy concern that has long animated copyright law confirms the close relationship of the harm it has allegedly suffered to a traditional copyright injury.

To state the conclusion clearly, even though the specific right created by the DMCA may be comparatively new, the injury experienced by The Intercept because of the violation of that right sounds in the same kind of harm long recognized in copyright suits.

Defendants' strongest rebuttal posits that The Intercept's injury derives not from a property-based harm but from harms related to non-attribution. They explain that "American law traditionally rejected droit morale -- the moral rights theories that would recognize injury based on harm to attribution or expressive integrity." Microsoft Suppl. Mem. in Supp. of Mot. to Dismiss at 4-5. But defendants' characterization of The Intercept's injury in a different manner and identification of a different historical analogue does not undermine the "close relationship" that The Intercept's injury enjoys with the property-based harms traditionally associated with copyright law. Moreover, the Court's inquiry "does not require an exact duplicate in American history and tradition." TransUnion, 594 U.S. at 424. So although its injuries might bear some resemblance to attribution-related harms, it is enough for its injuries to relate to the property-based harms of copyright law.

Defendants' other arguments are less persuasive. They insist that The Intercept has failed to identify a public "dissemination" necessary to support a concrete injury. See, e.g., OpenAI Suppl. Mem. in Supp. of Mot. to Dismiss at 11, 13-14. They quote

TransUnion for support: "The mere presence of an inaccuracy in an internal . . . file, if not disclosed to a third party, causes no concrete harm." Id. at 11 (quoting TransUnion, 594 U.S. at 434). TransUnion, however, dealt with an analogy to the common-law tort of defamation. See TransUnion, 594 U.S. at 432. A copyright injury -- or, more generally, an injury to a property right -- does not require publication to a third party.

Finally, defendants conflate the concreteness inquiry with their arguments on the merits. For example, OpenAI contends that The Intercept "cannot plausibly allege that any ChatGPT user has ever used the kind of prompt The Intercept used -- or any other similar prompt -- to generate a similar [regurgitated] output." OpenAI Suppl. Mem. in Supp. of Mot. to Dismiss at 14. And even if these prompts were "illustrative of a typical user's interactions with ChatGPT," OpenAI contends that the "claimed injury is purely imaginary and fabricated." Id. at 15. But "in reviewing the standing question, the court must be careful not to decide the questions on the merits for or against the plaintiff, and must therefore assume that on the merits the plaintiffs would be successful in their claims." City of Waukesha v. EPA, 320 F.3d 228, 235 (D.C. Cir. 2003) (per curiam). Here, The Intercept specifically alleges that defendants removed CMI from its articles reproduced in the training sets, which concealed their own systematic practice of copyright infringement and facilitated

infringement by ChatGPT users. Although the Court must consider whether The Intercept has sufficiently supported its claims to survive defendants' challenge on the merits, it should not allow its evaluation of the merits to influence the standing inquiry.

B. Failure to State a Claim

OpenAI and Microsoft both move to dismiss The Intercept's claims under Rule 12(b)(6) of the Federal Rules of Civil Procedure. "To survive a motion to dismiss, a complaint must contain sufficient factual matter, accepted as true, to state a claim to relief that is plausible on its face." Ashcroft v. Iqbal, 556 U.S. 662, 678 (2009). A complaint must offer more than "a formulaic recitation of the elements of a cause of action," or "naked assertion[s]" devoid of "further factual enhancement." See Bell Atl. Corp. v. Twombly, 550 U.S. 544, 555, 557 (2007). If the plaintiffs have "not nudged their claims across the line from conceivable to plausible, their complaint must be dismissed." Id. at 570. However, the Court must "construe[] the complaint liberally, accepting all factual allegations in the complaint as true, and drawing all reasonable inferences in the plaintiff's favor." Goldstein v. Pataki, 516 F.3d 50, 56 (2d Cir. 2008).

The Court starts with the § 1202(b)(1) claims and then turns to the § 1202(b)(3) claims.

i. Section 1202(b)(1): OpenAI

To state a claim under § 1202(b)(1), a plaintiff must allege four elements: "(1) the existence of CMI on the allegedly infringed work, (2) the removal or alteration of that information, . . . (3) that the removal was intentional," and (4) that defendant knew or had "reasonable grounds to know" that the CMI removal "[would] induce, enable, facilitate, or conceal" copyright infringement. See Fischer v. Forrest, 968 F.3d 216, 223 (2d Cir. 2020); § 1202(b)(1).

The Intercept has plausibly alleged that OpenAI intentionally removed CMI from its articles. In support, it identifies the specific training sets that it claims OpenAI uses to train ChatGPT and lists the specific URLs from its web domain that its data scientist has mined from approximations of these datasets. Moreover, The Intercept has described the algorithms that OpenAI employs to build the training sets; it explains that Dragnet can only capture an article's main text and suggests that OpenAI likely omitted CMI from its extractions using the Newspaper algorithm, to ensure uniformity with Dragnet. Of course, these allegations rely in part, even if implicitly, on information and belief. But the Court does not expect more at this early stage of the litigation, particularly because of OpenAI's secrecy over the contents of the sets used to train the latest versions of ChatGPT. Indeed, the Second Circuit has recognized that "[t]he Twombly plausibility

standard . . . does not prevent a plaintiff from pleading facts alleged upon information and belief where the facts are peculiarly within the possession and control of the defendant." Arista Recs., LLC v. Doe 3, 604 F.3d 110, 120 (2d Cir. 2010).

The more difficult question is whether The Intercept has plausibly alleged that OpenAI knew or had "reasonable grounds to know" that the alleged CMI removal "[would] induce, enable, facilitate, or conceal" copyright infringement. The Intercept emphasizes "this Circuit's 'lenient' pleading rules" that often allow scienter elements to survive motions to dismiss. Pl. Suppl. Opp'n at 2 (citing Aaberg v. Francesca's Collections, Inc., No. 17-cv-115, 2018 WL 1583037, at *9 (S.D.N.Y. Mar. 27, 2018) (Nathan, J.) (observing, in the context of a § 1202(b) claim, that "[t]he Second Circuit has stated that courts should be lenient in allowing scienter issues to survive motions to dismiss").

Still, a more lenient pleading standard does not give The Intercept a free pass. Other courts to study this particular element have emphasized its purpose in making out a § 1202(b) claim. For example, the Ninth Circuit has explained that "the plaintiff must provide evidence from which one can infer that future infringement is likely, albeit not certain, to occur as a result of the removal or alteration of CMI." Stevens v. Corelogic, Inc., 899 F.3d 666, 675 (9th Cir. 2018). In other words, "the mental state requirement in Section 1202(b) must have a more

specific application than the universal possibility of encouraging infringement." Id. at 674. To summarize in slightly different terms, the knowledge requirement isolates the kind of conduct that is the concern of copyright law -- knowing interference with a copyright-protected work that may risk the legal protections afforded that work.[8]

The Intercept cites factual matter that it believes plausibly supports OpenAI's knowledge of two categories of likely infringement: (1) its concealment of its own infringement in reproducing The Intercept's articles in training sets; and (2) its facilitation of ChatGPT users' downstream infringement of regurgitations of its articles produced in ChatGPT outputs. The Intercept does not clearly explain the first category. Most directly, it is difficult to understand how the removal of CMI from articles collected in a non-public database "conceals" infringement. The Intercept appears to express concern that the removal might conceal infringement from a ChatGPT user, but that argument largely repackages The Intercept's second theory --

[8] Compare, for example, a reader who removes CMI to increase the readability of an article he has downloaded and a copywriting service that, instead of generating new content, reuses other publication's articles and shares those articles, passed off as its own work, with its clients with critical CMI removed. The latter risks downstream copyright infringement much more than the former. Accordingly, the knowledge requirement likely saves the casual reader and singles out the unscrupulous company for CMI-removal liability.

addressed in greater detail below -- that OpenAI's alleged removal of CMI facilitates downstream infringement by ChatGPT users. To the extent The Intercept suggests that OpenAI concealed its copyright infringement in the training sets from Microsoft, The Intercept has not plausibly alleged that OpenAI and Microsoft exchanged training sets, as explained in greater detail below. See infra section III(B)(iii).

The Court is, however, persuaded by the second theory, which concerns downstream infringement. The Intercept explains that OpenAI "possess[es] a repository of every regurgitation of [its] works" by ChatGPT. Am. Compl. ¶ 63. And even though, according to The Intercept, OpenAI has more recently adjusted ChatGPT's settings to limit regurgitation, The Intercept still alleges that ChatGPT "regurgitate[s] verbatim . . . copyright-protected works of journalism" without CMI, "[a]t least some of the time." Id. ¶ 64. In fact, The Intercept's data scientist was able to produce three regurgitations from ChatGPT in response to detailed prompts. See Ex. 6 to Am. Compl. The Intercept further alleges that CMI removal in regurgitations can facilitate infringement, because ChatGPT was promoted "as a tool that can be used by a user to generate content for a further audience." Am. Compl. ¶ 83. OpenAI's indemnification policy for "commercial" users sued for copyright infringement provides additional support for its allegation that

the company knew some users would likely infringe copyright. See id. ¶ 86.

OpenAI tries to locate factual deficiencies in the complaint, but these arguments do not fairly construe The Intercept's allegations. For example, OpenAI argues that the evidence cited by The Intercept demonstrates, at most, that it had knowledge of the regurgitation problem only after it had allegedly removed CMI from articles in the training set. Contrary to OpenAI's cramped interpretation of the complaint, the allegations suggest that OpenAI continues to use the same methods to build training sets.

Likewise, OpenAI emphasizes that the regurgitations produced by The Intercept's data scientist are de minimis and otherwise depend on highly contrived prompts. But the complaint explicitly alleges that OpenAI has altered its regurgitation settings; The Intercept does not allege that its data scientist's ChatGPT interaction is similar to all such regurgitations, just that the chatbot still has some capacity to regurgitate. Contrary to OpenAI's position, The Intercept appears to cite whatever evidence is publicly available to support its position that OpenAI knew at the time it removed CMI that the removal could facilitate copyright infringement.

For these reasons, the Court denies OpenAI's motion to dismiss the § 1202(b)(1) claim.

ii. Section 1202(b)(1): Microsoft

The Intercept's § 1202(b)(1) claim against Microsoft depends on several discrete pieces of evidence: (1) its "partnership with OpenAI," including its "billions of dollars" of investment and its future "49% stake in the company"; (2) its alleged provision of "the data center and bespoke supercomputing infrastructure used to train ChatGPT"; and (3) its CEO's statement that "we have the data, we have everything." Id. ¶¶ 20–23. Otherwise, without including any factual detail about its training process, the complaint adds Microsoft's own AI product, Copilot, to some of its allegations about ChatGPT. See, e.g., id. ¶ 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 CMI.) (emphasis added).

Without the factual specificity supporting The Intercept's § 1202(b)(1) claim against OpenAI (e.g., the description of the training sets and the algorithms used to scrape articles, the examples of regurgitations), the allegations against Microsoft do not plausibly allege liability for CMI removal. Critically, none of the specific items of evidence related to Microsoft's involvement with ChatGPT or its development of Copilot bears any relationship to alleged CMI removal. All the specific factual matter in the complaint related to CMI removal connects only to

OpenAI. For these reasons, the Court dismisses the § 1202(b)(1) claim against Microsoft.

iii. Section 1202(b)(3) Claims: OpenAI & Microsoft

To state a claim under § 1202(b)(3), a plaintiff must allege: "(1) the existence of CMI in connection with a copyrighted work; and (2) that a defendant distributed works or copies of works; (3) while knowing that CMI has been removed or altered without authority of the copyright owner or the law; and (4) while knowing, or having reasonable grounds to know that such distribution will induce, enable, facilitate, or conceal an infringement." Mango, 970 F.3d at 171.

The Intercept includes no factual support for its allegation that Microsoft and OpenAI distributed its articles. The complaint asserts that Microsoft has "shared copies of [The Intercept's] works" without CMI "with the OpenAI [d]efendants," Am. Compl. ¶ 75, and, similarly, that OpenAI has shared its training-set data, which allegedly includes The Intercept's articles, with Microsoft, see id. ¶ 24. The factual support for these allegations, however, derives exclusively from general observations about the business relationship between OpenAI and Microsoft. But the mere fact that Microsoft has a partnership with OpenAI and that it provides a "data center and bespoke supercomputing infrastructure" indicates neither that the companies share training-set data for their

competing AI products nor that they shared The Intercept's articles.

The Intercept's repeated invocation of the Microsoft CEO's statement that "we have the data" omits key context from the cited interview (which the Court can consider even on a motion to dismiss[9]). The CEO was expressing his confidence in Microsoft's own AI capabilities -- separate and apart from its investment in OpenAI -- and cited, in support of his position, Microsoft's ownership of "IP rights" and possession of "the people," "the data," and "everything." Intelligencer Staff, Satya Nadella on Hiring the Most Powerful Man in AI, The Intelligencer (Nov. 21, 2023), https://nymag.com/intelligencer/2023/11/on-with-kara-swisher-satya-nadella-on-hiring-sam-altman.html. More plausibly, the CEO referenced Microsoft's rights under its partnership agreement with OpenAI, as well as its own capabilities with Copilot. It is a much more remote inference -- without more evidence -- that the CEO intended to suggest that Microsoft had

[9] When considering a motion to dismiss, a district court "may consider . . . documents incorporated by reference in the complaint" and a document that is "integral to the complaint" because the complaint "relies heavily upon its terms and effect." DiFolco v. MSNBC Cable L.L.C., 622 F.3d 104, 111 (2d Cir. 2010). Here, the web address of the full article is cited in the complaint, see Am. Compl. ¶ 23 n.2, and the terms of the article are critical to The Intercept's allegations against Microsoft. For these reasons, the Court will consider the full article.

access at that moment to the specific training sets (including The Intercept's articles) that had been used to train ChatGPT.

The Court therefore dismisses the § 1202(b)(3) claims against both OpenAI and Microsoft.

* * *

Finally, OpenAI briefly raises a statute-of-limitations defense. It cites the Copyright Act's three-year statute of limitations and asserts that the amended complaint admits that OpenAI constructed the training sets at least five years ago. That argument, however, mischaracterizes The Intercept's complaint, which cites the only publicly available information on OpenAI's training sets to support its allegations that OpenAI has continued to use the same process to train ChatGPT. See, e.g., Am. Compl. ¶ 54.

Further, "[t]he lapse of a limitations period is an affirmative defense that a defendant must plead and prove." Staehr v. Hartford Fin. Servs. Grp., 547 F.3d 406, 425 (2d Cir. 2008). "[A] defendant may raise an affirmative defense in a pre-answer Rule 12(b)(6) motion if the defense appears on the face of the complaint." Id. Despite OpenAI's arguments to the contrary, its statute-of-limitations defense in this case is not clearly presented on the face of the complaint and "thus is inappropriate to resolve on a motion to dismiss." See Kelley-Brown v. Winfrey, 717 F.3d 295, 308 (2d Cir. 2013) (discussing a court's

consideration of an affirmative defense at the motion-to-dismiss stage in the context of a fair-use defense).

IV. Conclusion

For the foregoing reasons, the Court denies OpenAI's motion to dismiss as to the § 1202(b)(1) claim but grants its motion to dismiss as to the § 1202(b)(3) claim. The Court grants Microsoft's motion to dismiss in full. Finally, the Court directs counsel to jointly call Chambers by no later than two business days after the date of this Opinion, in order to schedule a revised case-management plan.

SO ORDERED.

New York, NY
February 20, 2025

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