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.

Case No. 1:24-cv-01515-JSR

DEFENDANT MICROSOFT CORPORATION'S REPLY MEMORANDUM
IN SUPPORT OF MOTION TO DISMISS THE COMPLAINT

May 16, 2024

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Page(s)

PRELIMINARY STATEMENT

The Intercept's Opposition confirms that its sparsely pleaded § 1202(b) claims are as threadbare as they appear.

As to Article III standing, The Intercept concedes that it alleges no injury based on public dissemination of its works without CMI. Opp. 8. Its asserted harm is nothing more than alleged removal of attribution information during the entirely internal training and development of an LLM. And The Intercept also points to no common law recognition that such non-public non-attribution is a cognizable injury. That ends the inquiry—The Intercept lacks standing. Its only attempted workaround is highly dubious: It equates the non-attribution in private with copyright infringement, a claim The Intercept does not bring. The analogy does not work. CMI is merely information *about* a copyright-protected work, not copyright-protected material itself. The Intercept's claims based on private removal of such information thus could only allege a bare technical violation of § 1202(b), with no concrete injury required to support standing. *Infra* § I.

The Intercept also fails to state a claim. Tasked with defending a Complaint light on factual allegations and heavy on conclusions, The Intercept largely retreats to platitudes about the plausibility standard. Most of the allegations The Intercept points to are not factual allegations at all; they are conclusory assertions that cannot support a plausible claim. When those are cast aside, all that remains of the Complaint are theories, hypotheticals, and rumors strung together to suggest that (a) because some unidentified Intercept works are allegedly in the training set, they *must* have had CMI removed; (b) because Microsoft and OpenAI have a "close relationship," Microsoft *must* have removed CMI from The Intercept's works; and (c) because there is some bare possibility that GPT-based products can emit output that matches some text in the training set, it *must* be likely that they will do so with The Intercept's works, and that end-users will then infringe them further. For The Intercept's claims to survive, all of these premises

must be plausibly alleged. None actually are. The Complaint should be dismissed in its entirety. *Infra* § II.

<div align="center">**ARGUMENT**</div>

I.   **THE INTERCEPT LACKS STANDING BECAUSE IT FAILS TO ALLEGE AN ACTUAL OR THREATENED INJURY.**

Microsoft's motion explained that the mere removal of CMI in a non-public setting, without public dissemination of the CMI-stripped copy, causes no concrete injury, and therefore cannot support standing. Mot. 8-13. In response, The Intercept does not dispute that the common law recognized no injury based on non-public non-attribution with respect to intellectual property. Instead, it claims that "the plaintiff gets to pick" its common-law analogy, Opp. 9 n.3, and selects "copyright infringement," Opp. 6—the *actual invasion* of intellectual property rights themselves. The Intercept thinks its selection solves its non-dissemination problem because "the common law … recognizes interference with property, without more, as a concrete injury." Opp. 7. But the analogy is inapt because CMI is simply not intellectual property, and The Intercept otherwise alleges no actual or imminent injury to any actual intellectual property right.

A plaintiff does not get to "pick" whatever common-law analogue seems advantageous. Opp. 9 n.3. Its "asserted harm" must have a "'close relationship' to a harm traditionally recognized as providing a basis for a lawsuit in American courts." *TransUnion LLC v. Ramirez*, 594 U.S. 413, 417 (2021) (quoting *Spokeo, Inc. v. Robins*, 578 U. S. 330, 340-341 (2016)). Merely invoking some legally adjacent or similar-sounding claim does not suffice—the injuries themselves must be analogous in nature. Nor, of course, does it suffice to claim that the "asserted harm" is a prelude to some other, future harm that may have a common-law analogue. *See id*. at 435-36. Thus, the plaintiffs in *TransUnion* could not prevail by invoking defamation

<div align="center">2</div>

on the theory that the false OFAC labels in the defendant's files might cause future injury to reputation if disseminated. *Id.* at 435; Mot. 8-9.

The Intercept's attempted analogy to copyright infringement fails these basic tenets. The Intercept asserts that copyright infringement is analogous to CMI removal because the CMI removal provisions are housed in "title 17," i.e., the Copyright Act, and that violations carry "similar remedies" as copyright infringement. Opp. 6. These observations are irrelevant because they say nothing about whether the asserted *harm*—private non-attribution—is analogous to the invasion of exclusive intellectual property rights.

The Intercept next claims that the CMI-based rights in § 1202 are similar to the "exclusive rights" in 17 U.S.C. § 106 because "both grant the copyright owner the sole prerogative to decide how future iterations of the work may differ from the version the owner published." Opp. 6-7. This peculiar contrivance manages to obscure the relevant injury for both traditional copyright infringement and CMI. The harm from copyright infringement is the invasion of intangible intellectual *property* rights; the Copyright Act grants "protection" in "original works of authorship" in defined "categories," 17 U.S.C. § 102, then confers to the copyright owner certain "exclusive rights," *id.* § 106. Section 1202 in no way expands exclusive rights in original works of authorship, and many such works have no associated CMI. CMI, by contrast, is not a copyright-protected "work," but "information conveyed in connection with copies … of a work." *Id.* § 1202. It is attribution information that "inform[s] the *public*" of copyright status. *See Roberts v. BroadwayHD LLC*, 518 F. Supp. 3d 719, 737 (S.D.N.Y. 2021). In short, copyright infringement is a property-based injury, and CMI removal is not.

Once this inapt analogy to copyright infringement is dispatched, The Intercept has nothing left to stand on. Its "asserted harm" is what Microsoft—and The Intercept itself—said it

3

is: an allegation that Defendants "strip away" CMI and therefore do "not … provide attribution" to "human journalists," Compl. ¶ 6. The Intercept offers no explanation for how this would constitute a cognizable injury under the common law given that it is alleged to be occurring out of public view. *See* Mot. 10 (explaining that the closest common law analogues—like unfair competition—all depend on public dissemination). And it expressly disavows "a dissemination-based injury," Opp. 8-10, effectively conceding that it cannot bring a suit for damages because it cannot allege that a GPT-based product has ever disseminated its works without CMI. The Intercept therefore lacks standing to pursue damages in this case.

To attempt to salvage some right to remain in court, The Intercept says that "even if standing did require dissemination, … The Intercept has plausibly alleged facts to support standing for an injunction." Opp. 10. The Intercept's argument depends on the inconceivable suggestion that a plaintiff need not "plead dissemination of their *own* works to have standing to pursue an injunction," Opp. 10, and that *Doe 1 v. GitHub, Inc.,* 672 F. Supp. 3d 837, 850 (N.D. Cal. 2023), supports that view. To the contrary, *Doe 1* made clear that "the party who invokes the court's authority [must] 'show that he *personally* has suffered some actual or threatened injury.'" *Id.* (quoting *Valley Forge Christian Coll. v. Ams. United for Separation of Church and State, Inc.*, 454 U.S. 464, 472 (1982)). It then found, in connection with an LLM that exclusively suggests *software code completions*, that the plaintiffs had alleged a "substantial risk" that the LLM "will reproduce *Plaintiffs' licensed code*." *Id.* at 851 (emphasis added). The Intercept offers no basis upon which to conclude that the GPT-based products are imminently likely to reproduce The Intercept's works without CMI, and indeed offers no response to the Motion's argument (at 12) that the Complaint "contains no allegations concerning how any AI

4

product even works" or the "user behavior" that would plausibly elicit a reproduction of The Intercept's works. The Intercept therefore lacks standing to pursue injunctive relief as well.

II.   THE INTERCEPT FAILS TO STATE A DMCA § 1202 CLAIM.

The Intercept's opposition also confirms that its § 1202(b) claims are bare conjecture based on virtually no plausible factual allegations that Microsoft has done anything to any Intercept work. Mot. 13-23. The Complaint identifies no works thought to be in the training set nor offers a basis for concluding that CMI was removed from them. *Infra* § A. It provides no plausible basis for the contention that *Microsoft* created training sets or removed CMI from works in the training sets (much less from *The Intercept's* works). *Infra* § B. And it proffers no plausible account of how any alleged removal of CMI from The Intercept's works will likely facilitate or conceal piracy, as required for a § 1202(b) claim. *Infra* § C. Faced with these shortcomings, The Intercept largely falls back on standards for notice pleading and plausibility on a motion to dismiss. But those standards are not nearly as forgiving as The Intercept would need to survive dismissal. The Motion should be granted.

A. The Complaint Does Not Plausibly Allege "Removal" Of CMI From Copies Of Works.

As Microsoft's Motion (at 15-18) explained, The Intercept's suggestion that CMI has been removed from its works during training is based purely on unsupported generalities about how GPT-based products work, without a single example or supporting allegation explaining why those generalities make its claims plausible. The Intercept does not dispute its failure to name a copyrighted work it thinks is in a training set, identify any specific CMI removed from it, or otherwise substantiate its theory that any work contained in the training set must have had CMI removed. Instead, it asserts that it does not need to do so. This is wrong.

In response to Microsoft's argument that The Intercept fails to identify any works from which CMI was removed, Mot. 17-18, The Intercept concedes that it "cannot name all its works contained in [the] training sets," but says that "[r]equiring a plaintiff to name all its works" would create perverse incentives to "conceal" CMI violations. Opp. 14. This is distraction, because The Intercept has not identified *a single work* from which CMI was allegedly removed—much less "identical copies" of its works, as is required for a CMI claim, *Doe 1 v. GitHub, Inc.*, No. 22-cv-06823-JST, 2024 WL 235217, at *8-9 (N.D. Cal. Jan. 22, 2024). The only allegation in the Complaint is based on some anonymous promise that "thousands" of The Intercept's works were included in training sets, without identifying a single one. That is not enough to survive a motion to dismiss.

The Intercept not only fails to identify the specific works from which CMI was allegedly removed but also fails to identify the specific CMI that was allegedly removed from those works. Mot. 17. The Intercept's response is that Microsoft "ignores the Complaint, which alleges that Defendants removed author, title, copyright notice, and terms of use from its works," which "are all types of CMI." Opp. 13. Again, this is bare generality. Any plaintiff can recite types of CMI—that does not mean they have plausibly identified *their* CMI that was removed from *their* works, as both *Andersen v. Stability AI Ltd.*, No. 23-cv-00201-WHO, 2023 WL 7132064, at *11 (N.D. Cal. Oct. 30, 2023) and *Free Speech Sys., LLC v. Menzel*, 390 F. Supp. 3d 1162, 1175 (N.D. Cal. 2019), confirm.

Most importantly, The Intercept does not meaningfully support the inference that "*everything* ever included in the training set … must have been stripped of CMI." Mot. 15-16. That is fatal—because The Intercept offers no example of any one of its own works being stripped of CMI, the plausibility of its claim depends entirely on it establishing that any work

contained in the training set will generally be stripped of CMI. Yet The Intercept offers only the single-sentence assertion that "it is plausible that an LLM doesn't output CMI precisely because it wasn't trained on CMI." Opp. 17. But as Microsoft explained, Mot. 16, the *premises* of this theory are themselves unsupported, so the conclusion is not an available inference. It is just the sort of conclusory hypothesis the Second Circuit rejected in *Melendez v. Sirius XM Radio, Inc.*, 50 F.4th 294, 307 (2d Cir. 2022), a case The Intercept does not contend with.

B. The Complaint Does Not (And Cannot) Plausibly Allege That *Microsoft* Removed CMI Or Distributed Works Lacking CMI.

The Intercept also fails to explain why Microsoft's work with OpenAI is sufficient to plausibly allege that Microsoft removed CMI or distributed allegedly CMI-less works to OpenAI. The Intercept readily concedes that the entire basis for its claim that Microsoft removed The Intercept's CMI is the "close working relationship" between Microsoft and OpenAI. Opp. 17; 23. But it offers no authority for the notion that such a relationship is enough to plausibly allege that anything one entity did, the other must have done also. That is because "[t]he fact that two companies … do work together" is insufficient "to attribute their acts to one another." *Como v. Commerce Oil Co.*, 607 F. Supp. 335, 340 (S.D.N.Y. 1985). Indeed, the mere fact of a "working relationship" is insufficient to support an inference that Microsoft was even involved in training OpenAI's GPT-based products, let alone that it *removed CMI from The Intercept's works* when doing so or shared copies of any such works with OpenAI. Mot. 17-19.

The Intercept's final attempt to plausibly allege that Microsoft removed CMI from its works rests on an unsupported (and unsupportable) assertion that Microsoft "ha[s] created Common Crawl datasets." Compl. ¶ 31; *see* Mot. 18. The Intercept clarifies that its theory is that Microsoft "took data from Common Crawl" and "created training sets from that data," Opp. 17 n.11, and that it must have removed CMI from Intercept works in that process. Compl. ¶ 38.

7

The Opposition confirms that this is all mere conjecture, offering no response to the Motion's direct assertion that The Intercept offers "no basis for [its] information and belief, nor details of any inquiry it conducted to form such a belief," Mot. 18.

Not only does the Opposition point to nothing in the Complaint supporting the notion that Microsoft was involved in training, it points to material from *outside* the Complaint that *undermines* The Intercept's Common-Crawl theory. Perhaps appreciating the deficiencies in its Complaint, The Intercept tries to backfill by claiming that "OpenAI has 'published a list of the top 1,000 domains present in WebText and their frequency,'" which list has "6,484 (unidentified URLs from The Intercept's web domain." Opp. 14 & nn.7, 9. Leaving aside the impropriety of adding new allegations in an opposition, *Ariel (UK) Ltd. v. Reuters Grp. PLC*, No. 05-cv-9646-JFK, 2006 WL 3161467, at *7 (S.D.N.Y. Oct. 31, 2006), these new allegations suggest that if The Intercept's works are in a training set, it is a *WebText* training set, not a Common Crawl training set. And the Complaint does not allege, even on a weak information-and-belief basis, that Microsoft has any connection to these training sets. *See* Compl. ¶ 31 (distinguishing between the WebText training sets, which were not created by Microsoft, and the Common Crawl set, which The Intercept tries to allege Microsoft helped create). The Intercept's misfire further confirms that The Intercept cannot tie Microsoft to CMI removal and that its CMI claims against Microsoft should be dismissed.

C. The Complaint Does Not Plausibly Allege A Likelihood That Removal Of CMI During Training Of An AI Tool Will Induce, Enable, Facilitate, Or Conceal Infringement.

Finally, The Intercept fails to allege a likelihood of future infringement, as necessary to establish Microsoft's awareness that removal of CMI from a work in a training set will somehow "induce, enable, facilitate, or conceal" such infringement under 17 U.S.C. § 1202(b). Mot. 20-23. The Opposition does not dispute that this element of the claim, though "ultimately a scienter

requirement," requires allegations that "infringement is objectively 'likely.'" Mot. 21 (quoting *Stevens v. Corelogic, Inc.*, 899 F.3d 666, 675 (9th Cir. 2018)). But it points to nothing in the Complaint plausibly establishing that GPT users, in the real world, are likely to infringe The Intercept's works as a result of the absence of CMI—another fatal failing in light of The Intercept's inability to point to a single example of its works being output at all.

The Intercept begins by stressing that scienter allegations may be "sparse" and that courts in the Second Circuit are "lenient" when it comes to pleading them. Opp. 18. But even in the scienter case The Intercept likes best, *Hirsch v. CBS Broadcasting Inc.*, the plaintiff identified the specific work from which he claimed CMI had been removed (a particular photo); described the features of the CMI that was removed (a gutter credit); and explained how the work, stripped of CMI, had been disseminated to the public (via a television program). *See* No. 17-cv-1860-PAE, 2017 WL 3393845, at *8 (S.D.N.Y. Aug. 4, 2017). Only with those allegations describing the actual dissemination of a photo lacking a gutter credit to the public did the court conclude that the "sparse" scienter allegations sufficed to show awareness of likely future infringement. *Id.* As explained (at 5-7), The Intercept offers no such allegations about its works and the CMI that was allegedly removed; nor does it offer factual matter demonstrating that—or how—its works would be reproduced to the public without CMI, making it impossible to infer from context that further infringement resulting from the absence of CMI is likely.

This same problem undermines The Intercept's reliance on the general allegation that reproducing works without CMI would "misinform the public about the source of the content Defendants' products provide to users." Opp. at 22. Again, as Microsoft explained in its Motion (at 15), The Intercept provides not a single example of any such work ever being reproduced, nor even describes with any detail what such a reproduction would look like in the context of the

generative AI tools at issue. Would it be a paraphrase? A few sentences in a larger response?

What user prompt might elicit it? But even leaving that defect aside, The Intercept would still

need an account of why (a) an end-user would be inclined to make a further use of an

undescribed, unattributed natural-language output; (b) that use would be infringing—versus a

fair use, 17 U.S.C. § 107; and (c) why a user so inclined would *not* make such a use if it

observed CMI. That is precisely the teaching of *Stevens v. Corelogic*, as well as the several cases

that have applied it in dismissing § 1202 claims at the pleading stage. Mot. 23 (collecting cases).

The Intercept has no answer. It acknowledges that the only allegations in the Complaint

are generalizations suggesting that users might "morally oppose stealing others' works or fear

liability." Opp. 23. None of this is grounded in actual, observed behavior of real people using

GPT-based products. And so The Intercept falls back on the generic statement that "one function

of CMI is to 'discourage piracy,'" and argues that because "Congress concluded that CMI will

discourage infringement, that is surely enough to survive a motion to dismiss." Opp. 23. This

circular reasoning fails—the whole point of requiring a showing of objectively likely

infringement is that the removal of CMI may be so incidental or so unlikely to facilitate

infringement that § 1202 liability would *not* serve the aim of thwarting piracy. The need for such

a limitation is most pressing in cases like this one, involving not purposeful distribution of

pirated copies, but multi-purpose technology that end-users put to countless non-infringing uses.

Because The Intercept has not plausibly alleged that these end-users are likely to infringe The

Intercept's works as a result of absent CMI, the § 1202(b) claims should be dismissed.

CONCLUSION

The Court should grant the motion to dismiss in its entirety.

10

Dated: May 16, 2024   Orrick, Herrington & Sutcliffe LLP

San Francisco, CA

By: */s/ Annette L. Hurst*
       Annette L. Hurst (admitted *pro hac vice*)
       The Orrick Building
       405 Howard Street
       San Francisco, CA 94105-2669
       Telephone: (415) 773-5700
       ahurst@orrick.com

       Lisa T. Simpson
       51 West 52nd Street
       New York, NY 10019
       Telephone: (212) 506-3767
       lsimpson@orrick.com

       Christopher J. Cariello
       51 West 52nd Street
       New York, NY 10019
       Telephone: (212) 506-3778
       ccariello@orrick.com

       Attorneys for Defendant
       Microsoft Corporation

11