Before the

U.S. Copyright Office
101 Independence Ave. S.E.
Washington, D.C. 20559

In the matter of Docket No. 2023-6	**Artificial Intelligence Study: Notice and Request for Public Comment**

Comments of The New York Times Company

The New York Times Company ("The Times"), publisher of *The New York Times*, makes this submission in response to the U.S. Copyright Office's notice of inquiry and request for comments. We are a member of the News/Media Alliance (N/MA) and of Digital Content Next ("DCN"), and their submissions, including the N/MA White Paper,[1] summarize the position of those of us in the publishing industry. The Times also writes separately to emphasize the importance of maintaining a healthy ecosystem of trusted, quality journalism, and to highlight the risks posed to that ecosystem by Generative AI ("GAI") tools.

The Times sees great potential in GAI, a technology we have been at the forefront of covering and understanding. More broadly, we believe in the ability of technology — when developed and implemented responsibly — to unlock innovation and productivity for news organizations and beyond. At the moment, however, some GAI companies have engaged in deeply problematic practices: they have copied The Times's millions of copyrighted news articles, opinion pieces, images, recipes, artwork, and more, and used them for products that siphon away the very readers who make our work possible. These practices undermine the continued production of quality journalism, and copyright law does not protect them.

I. The Value of Our Journalism

By keeping the public informed about a complex world, the news industry plays a vital role in our democracy. The Times has been providing groundbreaking, frontline coverage and analysis of the world's events for more than 170 years, relying on a network that today encompasses 32 international bureaus and 5,800 full-time employees, including our journalists who investigate and report the news to our readers in 200 countries. Our journalism has impacted the way people experience the world, informed law and foreign policy, exposed corruption, and shifted the social discourse not only among our millions of readers, but far beyond. This type of newsgathering requires tremendous effort and resources.

[1] The White Paper is titled: "How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement and Not a Fair Use."

To do this work, news organizations need to pay salaries, fund research, cover travel, provide security, deliver content, and more. For The Times, this is made possible by the print and digital subscribers who pay for access to our content, the advertisers who pay to appear next to it, and the licensees who pay to copy and disseminate it.

Our content is valuable and we work hard to protect it from being misappropriated by others. The Times protects its intellectual property by, among other things, erecting a metered paywall, publishing Terms of Service that prohibit commercial use of its content (including for training GAI models), investigating and enforcing against infringement, requiring licenses for use, and registering its copyrights with the U.S. Copyright Office each day. These steps have been effective because, as described below, they are supported by the strong legal protections that copyright law gives to content creators. Without these critical protections and the business models they enable, we and others like us would not be able to create the independent journalism that is essential to democratic society.

II. Our Copyrighted Content Is Being Infringed

As this Office knows, the Founders recognized that the financial incentive of exclusive ownership of original works was necessary to "the Progress of Science and useful Arts." U.S. Const., Art. I, § 8, cl. 8. The Copyright Act implements this principle and "encourages creativity by granting to the author of an original work 'a bundle of exclusive rights,'" which "includes the rights to reproduce the copyrighted work" and "prepare derivative works." *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258, 1273 (2023) (citation omitted). For those in the news industry — whose principal assets are the works created by journalists — the ability to monetize those exclusive rights sustains our businesses. The incentive of exclusive ownership has allowed the press to thrive in this country for centuries because those who gather and report news are able to enjoy (and protect) the fruits of their labor and investment.

The Times's content, including its expressive, original news coverage, is protected by copyright law.[2] In fact, The Times has published millions of articles online that have registered copyrights. And there is no "open web" in which the rules of copyright do not apply. Nevertheless, many GAI companies have infringed our exclusive rights by copying a massive amount of content from our website and unauthorized third-party platforms to create tools that mimic, closely paraphrase, and copy our work. As the N/MA explains in its submission and White Paper, many GAI companies simply copy, without permission, vast quantities of protected content to develop their models, including by: (1) crawling and scraping news publishers' websites and unauthorized third-party platforms to build training datasets; (2) training the Large Language Models ("LLMs") that are used to power GAI tools by passing the training datasets through the models for one or more training cycles; and (3) generating output containing copies or derivatives of publisher content. The Times's content plays a critical role in each of these steps.

A substantial portion of The Times's content produced during its 170-year history was copied and used to train LLMs without permission from, or compensation to, The Times. While GAI developers have often been opaque about the sources of content used to build their LLMs, what they have revealed

[2] While our content includes facts (in addition to images, opinions, art, fiction, poetry, games, and more), those facts are reported in carefully written and edited expressive, creative language, which is clearly protected by copyright.

shows the extensive presence of The Times's content. For example, one of the most commonly used datasets, Common Crawl, until recently contained millions of URLs linking to our content.[3] Similarly, a recreated version of WebText, the dataset used to train OpenAI's ChatGPT-2, shows that a stunning 1.2% of the dataset is The Times's content. And The Times was reported to be the fourth-largest source for Google's C4 dataset,[4] which powers GAI products like Google's Bard and its Generative Search Experience.

Once powered with our content, GAI tools can do a number of things with it, including reciting it verbatim, summarizing it, drafting new content with a similar style of expression, and using it to generate misinformation attributed to The Times that appears to be fact.[5] Increasingly, some GAI products go so far as to retrieve and copy our most recent and relevant content in order to "ground" generative AI output, through a process known as "retrieval augmented search." GAI products like Google's Bard, Microsoft's Bing Chatbot, and ChatGPT's "Browse with Bing" combine an LLM with a search index so that they can generate lengthy, detailed summaries of our unique coverage. While traditional search engines encouraged users to click on links to publishers' websites in order to find substantive content, which in turn resulted in engagement that publishers could monetize, GAI products are designed to keep readers on the companies' own tools and websites by providing expressive, satisfying summaries in response to queries that obviate the need for users to travel to publishers' platforms. And these entities (some of whose products require subscriptions) are commercial, with lucrative consumer and business-to-business revenue models. Copyright law does not permit the kind of systematic and competitive infringement that has occurred here.[6]

[3] At The Times's request, Common Crawl Foundation has begun the process of removing our content from its existing datasets and has agreed to not scrape more content.

[4] Kevin Schaul, Szu Yu Chen and Nitasha Tiku, *Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart*, Washington Post (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

[5] This last category of output is separately problematic because certain tools falsely attribute misinformation to The Times, thereby tarnishing our brand and harming our reputation.

[6] In addition to violating our exclusive rights by inputting our content into their models and using it to develop their tools, GAI companies have separately violated our rights by producing output that contains closely detailed summaries, verbatim excerpts, and/or similar structure and expression to the original work. *See Castle Rock Entm't, Inc. v. Carol Publ'g Grp., Inc.*, 150 F.3d 132, 141-46 (2d. Cir. 1998) (affirming finding that "Seinfeld Aptitude Test" was an infringing derivative work that did not constitute fair use); *Wainwright Sec. Inc. v. Wall St. Transcript Corp.*, 558 F.2d 91, 95-96 (2d Cir. 1977) (affirming finding of infringement where summaries of Wall Street Journal articles appropriated "the manner of expression, the author's analysis or interpretation of events, the way he structures his material and marshals facts, his choice of words, and the emphasis he gives to particular developments"); *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 561 (S.D.N.Y. 2013) (finding excerpting of AP news articles to be infringing and not fair use); *Warner Bros. Entm't Inc. v. RDR Books*, 575 F. Supp. 2d 513, 554 (S.D.N.Y. 2008) (finding "Lexicon" of facts, summaries, and supplemental material drawn from the *Harry Potter* series to be infringing and not fair use).

III. GAI Companies' Use of Our Content Falls Outside the Scope of Fair Use

The Times and most other news publishers rely on fair use often, and it is a critically important tool in our coverage. But the doctrine is not a blanket invitation to take copyrighted content; instead, each use must be assessed on a case-by-case basis. News organizations engage in this careful, detailed analysis every day when deciding whether to publish, for example, a newsworthy video that has been shared on social media but whose owner cannot be found. We also invest in license agreements with content providers, such as Getty Images, the Associated Press, and others in order to use their protected content.

In contrast, GAI companies concede that they have engaged in widespread copying of our content, during ingestion, in their output, and likely during the stages in between, but claim that their conduct is protected as fair use because it is "transformative." But taking expressive, protected content and using it to power tools that output close, sometimes verbatim, summaries of that very content is not transformative for purposes of copyright law. If there was any ambiguity about that before the Supreme Court's recent decision in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258 (2023), which addressed the first factor of fair use, there should be none now. The Court's ruling addressed the first factor of fair use ("the purpose and character of the use") in detail, and stressed the substitutive and commercial nature of the use of Warhol's artwork, noting: "the first factor [of the fair use analysis] relates to the problem of substitution—copyright's bête noire. The use of an original work to achieve a purpose that is the same as, or highly similar to, that of the original work is more likely to substitute for, or 'supplan[t],' the work." *Id.* at 1274 (citation omitted). Accordingly, even though there was some transformation of the underlying work, the fair use defense failed primarily because Warhol's use was a commercial one that competed with the copyright owner. That is precisely the case here: GAI products use our content for purposes that are clearly commercial and harm The Times by creating output that is substitutive of our content. Moreover, although GAI companies focus on the purported "transformativeness" of their uses of copyrighted content, this is one component of one factor of a four-factor test. The other three factors, like the first, weigh decisively against a finding of fair use.[7]

GAI tools may be new, but the fair use arguments that proponents advance have been routinely rejected in a variety of parallel contexts where creators of a new digital product seek to use others' content and then compete against them. That is a copyright violation, not a transformative use. *See, e.g.*, *Hachette Book Grp., Inc. v. Internet Archive*, 2023 U.S. Dist. LEXIS 50749, *18-26 (S.D.N.Y. Mar. 24, 2023) (holding Internet Archive's electronic copying and unauthorized lending of 3.6 million books

[7] GAI companies cannot prevail on the second factor ("the nature of the copyrighted work"), the third factor ("the amount and substantiality of the portion used") or the fourth factor ("the effect of the use upon the potential market for or value of the copyrighted work"). 17 U.S.C. § 107. This is because they are copying massive quantities of entire, expressive, copyright-protected works, and using them in a manner that directly harms the copyright owners' core revenue streams. GAI companies then use those copied works for profit in commercial, enterprise businesses, including by licensing their AI models and GAI tools to third parties. And not only do the companies' infringing uses of our content harm our subscription and advertising businesses, they also reduce the value of, and harm, our existing licensing business, which grants licenses for a wide variety of uses, including data mining and media monitoring. Therefore, any fair use defense is further undermined because even if "'intermediate' copies [are] not ultimately used in any end product, they threaten to reduce the value of the right to copy the [copyrighted works] and undermine [the copyright holder's] relationships with licensees who pay for that right." *Fox Broad. Co. Inc. v. Dish Network, L.C.C.*, 905 F. Supp. 2d 1088, 1106 (C.D. Cal. 2012).

protected by valid copyrights is not a transformative fair use); *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 561 (S.D.N.Y. 2013) (holding that crawling of various websites for Associated Press's stories and scraping "snippets" of those stories for use in notifying and informing Meltwater's own customers of certain stories was not a transformative use and directly competed with the Associated Press such that Meltwater's copying would deprive the Associated Press of a stream of income to which it was entitled).[8] Nor does case law create an exception to the Copyright Act for all innovative technology, particularly when that technology risks putting copyright holders out of business.[9]

IV. Conclusion

Our ability to serve as an essential source of journalism for readers around the world depends on our ability to receive fair compensation for content that we have developed at extraordinarily high costs. We are also confident that the success of GAI and the companies developing it need not come at the expense of journalistic institutions. News organizations are willing and well equipped to enter into negotiated, fair agreements that permit and govern the use of our content. We ask the Copyright Office to help protect journalism and shape the development of this technology through guidance and legislative recommendations, including the following:

- The Times would see benefit in legislation (such as through an amendment to the fair use provision in the Copyright Act, 17 U.S.C. § 107) and/or guidance from the Copyright Office that would clarify that using copyrighted content to train and develop GAI products is not per se transformative or fair use. As the Office knows, Congress added similar "thumb on the scale" language to § 107 in 1992, in the wake of a judicial decision that gave substantial weight to the

[8] Critically, even decisions finding fair use are very clear that the principle will not extend to cover works that replace the work of the copyright owner, inevitably harming them in the marketplace. For example, in *Authors Guild v. Google, Inc*., 804 F.3d 202, 208-12, 221-25 (2d Cir. 2015), and *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 90, 99-101 (2d Cir. 2014), the issue revolved around the creation of a database of millions of scanned books supporting a comprehensive search engine. The Second Circuit found that the unauthorized reproduction of copyrighted works was a transformative fair use largely *because* the services were not substitutive of the market for the original books. Similarly, in *Sony Corp. of Am. v. Universal City Studios, Inc*., 464 U.S. 417, 456 (1984), the Supreme Court permitted the sale of VCRs because merely recording a broadcast for later viewing did not harm the market for the copyrighted works. *See also Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1165-66 (9th Cir. 2007) (holding that image thumbnails were fair use because they merely served as pointers to direct users to the original content); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 821-22 (9th Cir. 2003) (finding that small, poor quality thumbnail images served an entirely different function than the original images, and thus caused no market harm); *Field v. Google Inc.*, 412 F. Supp. 2d 1106, 1121-22 (D. Nev. 2006) (finding no market harm from copying web pages that were created to manufacture copyright claims arising from Google's search function). There are also key differences between GAI output and traditional search results that weigh decisively against developers in a fair use analysis, including: (i) the substitutive, satisfying, and expressive nature of GAI, which makes it less likely to send users to sites owned by the content creators; (ii) the risk of societal harm from GAI and the harm to the public interest by, among other things, disincentivizing independent content creation; and (iii) the commercial nature of the tools.

[9] *See Hachette Book Grp., Inc.*, 2023 U.S. Dist. LEXIS 50749, at *24 ("Unlike Sony, which only sold the machines, IA scans a massive number of copies of books and makes them available to patrons"); *see also A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1019 (9th Cir. 2001) (finding *Sony* to be "inapposite" because its time shifting did not "involve distribution of the copyrighted material to the general public").

unpublished nature of a work to find infringing the reproduction of unpublished letters in a biography. Though the conduct of GAI companies is unlawful under existing law, it is clear that they are operating under a misapprehension of the fair use doctrine, and a clarification of this type may help rectify that.

- To help news organizations enforce their rights, provide for a registration process for bulk online content and dynamic websites that reflect these new digital publication channels.
- To address the secretive nature of AI training, endorse detailed transparency requirements for GAI companies regarding the source, contents, and use of training datasets.

We rely on the submissions of DCN and N/MA, as well as the N/MA White Paper, to address the remaining questions posed in the notice of inquiry, and thank you for your time and consideration.