The New York Times sues OpenAI and Microsoft for copyright infringement

By <u>Clare Duffy</u> and <u>David Goldman</u>, CNN

New York Times newspaper office building is seen in Manhattan, New York, United States, on October 26, 2022.
Beata Zawrzel/NurPhoto/Getty Images

New York (CNN) — The New York Times has sued OpenAI and Microsoft for copyright infringement, alleging that the companies' artificial intelligence technology illegally copied millions of Times articles to train ChatGPT and other services to provide people with instant access to information — technology that now competes with the Times.

The complaint is the <u>latest in a string of lawsuits</u> that seek to limit the use of alleged scraping of wide swaths of content from across the internet — without compensation — to train so-called large language artificial intelligence models. Actors, <u>writers, journalists and other creative types</u> who post their works on the internet fear that AI will learn from their material and provide competitive chatbots and other sources of information without proper compensation.

But the Times' suit is the first among major news publishers to take on OpenAI and Microsoft, the most recognizable AI brands.

Microsoft (MSFT) has a seat on OpenAI's board and a multi-billion-dollar investment in the company.

In a complaint filed Wednesday, the Times said that it has a duty to inform its subscribers, but Microsoft and OpenAI's "unlawful use of The Times's work to create artificial intelligence products that compete with it threatens The Times's ability to provide that service." The paper noted that OpenAI and Microsoft used other sources in its "widescale copying," but "they gave Times content particular emphasis" seeking "to free-ride on The Times's massive investment in its journalism by using it to build substitutive products without permission or payment."

"We respect the rights of content creators and owners and are committed to working with them to ensure they benefit from AI technology and new revenue models," OpenAI said in a statement from spokesperson Lindsey Held. "Our ongoing conversations with the New York Times have been productive and moving forward constructively, so we are surprised and disappointed with this development. We're hopeful that we will find a mutually beneficial way to work together, as we are doing with many other publishers."

Microsoft did not respond to a request for comment on the lawsuit.

Disney, The New York Times and CNN are among a dozen major media companies blocking access to ChatGPT as they wage a cold war on A.I.

The Times, in its complaint, said that it objected when it discovered months ago that its work had been used to train the companies' large language models. Starting in April, the Times said it began negotiating with OpenAI and Microsoft to receive fair compensation and set terms of an agreement.

But the Times alleges it has been unable to reach a resolution with the companies. Microsoft and OpenAI claim that the Times' works are considered "fair use," which gives them the ability to use copyrighted material for a "transformative purpose," the complaint states.

The Times strongly objected to that claim, saying ChatGPT and Microsoft's Bing chatbot (also known as "copilot") can provide a similar service as the New York Times.

"There is nothing 'transformative' about using The Times's content without payment to create products that substitute for The Times and steal audiences away from it," the Times said in its complaint. "Because the outputs of Defendants' GenAI models compete with and closely mimic the inputs used to train them, copying Times works for that purpose is not fair use."

Pushing back against AI

The Times is among a number of leading newsrooms, also including CNN, who earlier this year added code to their websites that blocks OpenAI's web crawler, GPTBot, from scanning their platforms for content.

In separate but related lawsuits earlier this year, comedian Sarah Silverman and two authors sued Meta and OpenAI in July, alleging the companies' AI language models were trained on copyrighted materials from their books without their knowledge or consent. Neither company has commented on the lawsuit. A judge in November dismissed most of the lawsuit's claims.

And a group of famous fiction writers joined the Authors Guild in filing a separate class action suit against OpenAI in September, alleging the company's technology is illegally using their copyrighted work.

In its lawsuit, The Times alleges that the datasets used to train the most recent OpenAI large language models, which power its AI tools, "likely used millions of Times-owned works." In a 2019 English-language snapshot of one of those datasets — called Common Crawl and known as a "copy of the internet" — the New York Times website is the third most highly represented source of information, behind Wikipedia and a database of US patent documents, according to the complaint.

The Times claims that because the AI tools have been trained on its content, they can "generate output that recites Times content verbatim, closely summarizes it, and mimics its expressive style, as demonstrated by scores of examples … These tools also wrongly attribute false information to The Times," the complaint states.

In one instance cited in the complaint, ChatGPT provided a user with the first three paragraphs of the 2012 Pulitzer Prize-winning article "Snow Fall: The Avalanche at Tunnel Creek," after the user

complained in the chat of having hit the Times' paywall and being unable to read it.

The news outlet also alleges that Microsoft's Bing search engine, which was upgraded earlier this year with OpenAI's technology, "copies and categorizes" Times content to produce longer and more detailed responses than traditional search engines.

"By providing Times content without The Times's permission or authorization, Defendants' tools undermine and damage The Times's relationship with its readers and deprive The Times of subscription, licensing, advertising, and affiliate revenue," the complaint states.

Embracing AI … with limits

But fighting AI is like sticking a finger in a dike. It's coming, and publishers like the New York Times recognize they'll have to embrace the future. They just want to ensure it's a future in which they're fairly compensated, the New York Times said.

The New York Times Executive Vice President and General Counsel Diane Brayton told the outlet's staffers in a memo Wednesday morning that, "We recognize the potential of [generative AI] for the public and for journalism."

"But at the same time, we believe that the success of GenAI and the companies developing it need not come at the expense of journalistic institutions," according to the memo, which was obtained by CNN. "The use of our work to create GenAI tools must come with permission and an agreement that reflects the fair value of that work, as the law provides."

With its lawsuit, the Times is claiming billions of dollars in damages, but did not specify the compensation it demands for the alleged infringement of its copyrighted materials. It also seeks a permanent injunction that would prevent Microsoft and OpenAI from continuing the alleged infringement. The Times is also seeking the "destruction" of GPT and any other AI models or training datasets that incorporate its journalism.

The Times lawsuit could ultimately set a precedent for the wider industry, because the question of whether using copyrighted material to train AI models violates the law is an unsettled legal matter, according to Dina Blikshteyn, partner in the artificial

intelligence and deep learning practice group at law firm Haynes Boone.

"I think there are going to be a lot of these types of suits that are popping up, and and I think eventually [the issue will] make it up to the Supreme Court, at which point we'll have some definite case law," Blikshteyn said, adding that, right now, "there is nothing specific to large language models and AI just because it's so new."

This story has been updated with additional developments and context.