LAW360

Portfolio Media. Inc. | 111 West 19th Street, 5th Floor | New York, NY 10011 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

# AI Models' Link To Nonprofit Data Raises Fair Use Question

By **Ivan Moreno**

*Law360 (January 18, 2024, 10:29 PM EST)* -- To develop ChatGPT and text-to-art program Stable Diffusion, artificial intelligence companies did not have to look far for the material that helped their programs wow the world.

All OpenAI and London-based Stability AI had to do to acquire content to train their programs was what the rest of us do when we want to learn something — go online.

The companies did not have to pay for or even request vast datasets used for AI training because, in many cases, they were already there — assembled by nonprofits whose stated purpose is to archive portions of the internet and provide the material for free.

Besides the cost and convenience, nonprofits have another advantage that makes using their material more appealing: Because the content they collect is for academic research and provided for free, it's fair use, AI companies have argued in lawsuits and comments to the U.S. Copyright Office.

But now the datasets AI companies have used from nonprofits such as Common Crawl and the Large-scale Artificial Intelligence Open Network, or LAION, are under scrutiny in lawsuits from authors, artists, entertainers and other content creators who allege the generative AI models are infringing copyrights.

Common Crawl, LAION and others are mentioned in many of the pending complaints against AI companies, although not as defendants. That doesn't mean the nonprofits can ultimately escape liability, however, according to interviews with intellectual property attorneys, law professors and AI researchers.

"It's an interesting dynamic for these cases to deal with because the company that's being [accused of] copyright infringement is not necessarily the company that created the dataset that the AI is using," said Darius Gambino, partner and chair of Saul Ewing LLP's sports and entertainment practice.

In LAION's case, artists who have filed a proposed class action accuse the nonprofit of having engineers from Stability AI and Google while receiving funding from Stability AI CEO Emad Mostaque, who called himself "the biggest backer of LAION" in 2022. The artists also claim Mostaque funded the nonprofit's dataset used for Stable Diffusion. The complaint said Mostaque walked back his comment after the artists sued.

LAION, Stability AI and Google did not respond to requests for comment for this story.

The debate over nonprofit datasets for AI training evokes memories of intellectual property rights after the proliferation of free music downloads in the early 2000s and "intermediate copying" to create video games compatible with certain consoles in the 1990s, attorneys said, and it's not yet clear where courts are going to land in the pending disputes.

"Does using [the datasets] in a commercial way then remove that fair use? And does it remove the fair use for the institution who originally created it, or does it remove it for that end user? Both of them have to consider that, and quite frankly, I think courts could go different ways on this," said Meaghan Kent, a partner at Morgan Lewis & Bockius LLP.

**'Not Some Cabal of Investors'**

AI companies are counting on courts sharing their interpretation of what's fair use.

"The American AI industry is built in part on the understanding that the Copyright Act does not proscribe the use of copyrighted material to train generative AI models," Meta Inc. said in its comments to the U.S. Copyright Office, which asked for public input last year on IP questions raised by the groundbreaking technology.

Teaching generative AI models how to provide responses to a seemingly endless number of prompts or to create art requires billions of lines of text and images. Algorithms break that content into trillions of component parts called "tokens."

"In practice, there is no other way to amass a training corpus with the scale and diversity necessary to train a complex [large language model] with a broad understanding of human language and the world in general," Anthropic PBC, founded by former members of OpenAI, said in a brief this week opposing a preliminary injunction request in a lawsuit from music publishers led by Universal Music Group.

Music publishers allege Anthropic used hundreds of copyrighted songs to train Claude, its large language model. Anthropic said although "it would of course be possible" to enter into a business relationship with the music publishers, it still "would not be possible to amass sufficient content to train an LLM like Claude in arm's length licensing transactions, at any price."

Anthropic said it uses data "broadly assembled from the publicly available Internet," including datasets from nonprofits. The 500 songs the music publishers allege were ripped off are a "miniscule fraction" of the material used to train Claude, Anthropic said.

OpenAI has made a similar argument in response to a December complaint from The New York Times, which also names Microsoft Corp. as a defendant, saying in a blog post last week that material from the newspaper "is a tiny slice of overall training data."

Microsoft, a large financial backer of OpenAI, declined to comment, and OpenAI did not respond to a request for comment.

Multiple lawsuits, including the one from the Times, refer to Common Crawl as a nonprofit backed by wealthy investors. Common Crawl Executive Director Rich Skrenta chuckled at the description.

"There's just been a single patron, which is our original founder and chairman, Gil Elbaz. So there's not some cabal of investors," he told Law360.

Common Crawl has archived 240 billion web pages and has been cited thousands of times in academic research papers since it was founded in 2007 by Elbaz, according to the nonprofit. Elbaz also co-founded Applied Semantics, one of Google's first acquisitions.

In response to a call for comments on AI issues by the Copyright Office, Microsoft cited Common Crawl specifically as one of the nonprofits that has made it possible to obtain the datasets required to train ChatGPT.

In Common Crawl's own comment submission to the Copyright Office, the nonprofit said it is the primary training dataset for every large language model. Even sets assembled by other nonprofits, such as LAION, build upon Common Crawl datasets.

Skrenta said the nonprofit has always given away its datasets for free, and the organization has no plans to change that. But whether nonprofits receive money from AI companies for their datasets may not be the most important question in a fair use analysis, said Evan Gourvitz of Ropes & Gray LLP.

"It seems to me that whether company X pays company Y for material company Y has is less important than how company X is using the material," Gourvitz said. "Is the way company X uses that material for profit? Is it commercial? Is it undercutting a market for the licensing of the work?"

That is what the Times argues is happening, saying ChatGPT reproduces parts of the newspaper when prompted, diverting readers and potential subscribers. The Times also alleged its product review publication Wirecutter is losing revenue from merchants when a reader clicks on a product link because ChatGPT is reproducing those reviews.

Skrenta said the nonprofits' crawler, called CC Bot, which archives material that has "been intentionally put on the web," doesn't go behind paywalls and obeys the Robots Exclusion Protocol, or robots.txt, telling it not to crawl a particular page.

"It goes to great lengths to crawl politely," Skrenta said. "It's not crawling under cover of darkness. It identifies itself. It says, 'Hi, I'm a bot.' It looks at robots.txt, and it only takes one page every five seconds, and it does its best to be a very gentle, polite citizen of the web."

The Times, which alleged in its complaint that ChatGPT gives users articles behind the paper's paywall when requested, asked last year that Common Crawl stop archiving material from its site, and the nonprofit has complied, according to Skrenta.

On its website, Common Crawl lists various examples "of illegal stuff" users agree not to do when downloading its datasets, including violating people's IP rights. How those terms are enforced is unclear. Anyone can download Common Crawl's datasets, "so they don't necessarily have to do a click to agree," Skrenta said.

While nonprofit datasets like Common Crawl's have played a pivotal role in the development of generative AI, that doesn't mean it will always be the case, according to Tinglong Dai, a professor at the Johns Hopkins Carey Business School who has studied how people interact with AI technologies.

"In the short term, they are very crucial. But I don't think they will be very valuable in the long run, just because of the fact that everyone has the same datasets," he said. "So if you're developing another generative AI model, and you use the same datasets that competitors are using, I don't see how you could actually beat your competitors."

**'A Fight About the Cost of the Inputs'**

While Common Crawl archives portions of the internet, LAION data sets consist of URLs. LAION is based in Hamburg, Germany.

AI developers argue their models do not memorize or store training content and that what they're doing is "intermediate copying," which the Ninth Circuit held to be fair use in Sega Enterprises Ltd. v. Accolade Inc. and Sony Computer Entertainment Inc. v. Connectix Corp.

In each case, decided in 1992 and 2000, respectively, the appeals court concluded the defendants made intermediate copies of computer code to reverse-engineer video games that could be played on Sega and Sony's PlayStation consoles.

In its comments to the Copyright Office last year, OpenAI likened ChatGPT's learning process to the way people read books.

"Much like a person who has read a book and sets it down, our models do not have access to training information after they have learned from it," OpenAI said.

The Times' complaint, the suit from music publishers against Anthropic and the class action from artists against Stability AI seek to poke holes in the intermediate copying arguments by including examples of articles, lyrics and images being replicated by AI models after prompting.

AI developers have responded by accusing plaintiffs of inducing their models to reproduce copyrighted material, which the companies say is against their terms of use.

Courts are more likely to find intermediate copying to be fair use if the "copying is purely internal to a technological system," according to James Grimmelmann, a professor at Cornell University Law School and Cornell Tech.

"I think that argument works to the extent that the model then doesn't emit things that are similar to the work that was trained on," he said. "If the model is actually capable of producing very close replicas of The New York Times article or very close screencaps of a movie, I don't think the argument that, 'Well, we're just making a temporary transient intermediate copy, and it's now been deleted,' is even true on the facts."

Grimmelmann co-wrote a paper in September with two doctoral students titled, "Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain," which concluded there are no easy answers to whether AI training material is fair use, saying "copyright law has too many open decision points to provide clear answers."

The arguments that datasets with publicly available content from the internet should be treated as fair use to train is reminiscent of the debate over peer-to-peer music sharing platforms in the early 2000s, said Gambino of Saul Ewing.

"I think one of the interesting parallels here is to the music industry in the 2000s when you had a lot of people out there saying, 'Well, music should be free. I should be able to make copies of my music and put it on Napster,'" Gambino said. "I think a lot of similar arguments are being made now — not by the everyday person who was the one making that argument in the music industry case, but by big companies saying, 'Well, this stuff is out there on the internet. We should be able to use it even though it's owned by and was created by various people.'"

Peer-to-peer music sharing showed it's possible to find a licensing solution, attorneys said, pointing out how Apple Inc. began charging $1 for downloads on its iTunes platform.

"This is really just a fight about the cost of the inputs," said Justin Hughes, a professor at Loyola Law School, Los Angeles and a visiting law professor at University of Oxford. "And if any tech executive says this is about whether or not we will have AI or whether or not generative AI will exist, that's malarkey."

--Editing by Philip Shea and Lakshna Mehta.