PUBLICATION

Artificial Intelligence and Copyright Law: *The NYT v. OpenAI* – Fair Use Implications of Generative AI

Authors: Edward D. Lanquist, Jr, Jeremy Dale Ray

February 05, 2024

The legal implications of artificial intelligence (AI), specifically generative AI, quickly became a topic of casual conversation following the launch of ChatGPT in November of 2022. Generative AI is a type of AI with the ability to create many forms of unique content (*e.g.*, images, video content, text, poems, stories, musical compositions, sound recordings, and even deepfakes). Generative AI platforms necessarily rely upon large pools of data, oftentimes including unlicensed third-party content, as input to "train" their platforms to create generative AI outputs.

Certain content creators have entered into licensing agreements with AI companies allowing for the use of their works. OpenAI and others have reportedly been in talks with dozens of publishers to license third-party content for their AI platforms. However, when AI companies use unlicensed material to train their LLMs, copyright law comes into play.

The lawsuit recently filed by The New York Times (The NYT) against OpenAI in the Southern District of New York illustrates the significant tension between AI companies and the entities that own or control the materials and content AI companies use to train their large language models (LLMs). The creators and content owners understandably want to be compensated and given proper attribution for the use of their works while the AI companies need access to significant amounts of content to effectively train their LLMs (preferably, under terms that are not cost-prohibitive).

The lawsuit claims that OpenAI's "commercial success is built in large part on OpenAI's large-scale copyright infringement." The NYT alleges that: (1) OpenAI's platform is powered by LLMs containing copies of The NYT's content; and (2) OpenAI's platform generates output that recites The NYT's content verbatim, closely summarizes it, mimics its expressive style, and even wrongly attributes false information to The NYT. Thus, the alleged misuses relate to both training the LLMs *and* the generative AI output based upon the underlying input. The NYT claims that, prior to the litigation being filed, it and OpenAI were in conversations to work out a potential license agreement. However, in the lawsuit, The NYT implies that OpenAI's insistence that their conduct will be protected as "fair use" under the Copyright Act may have interfered with such negotiations.

Fair use is a legal doctrine under the Copyright Act that promotes freedom of expression by permitting the unlicensed use of copyrighted works in certain circumstances. The statutory factors for fair use are as follows:

1. The purpose and character of the use, including whether such use is of a commercial nature or if it is for non-profit educational purposes;
2. The nature of the copyrighted work;
3. The amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
4. The effect of the use upon the potential market for the value of the copyrighted work.

The alleged infringer has the burden of proving their use was a fair use. We find it highly likely that fair use will be central to OpenAI's defense.

The court in *The NYT v. OpenAI* matter will likely bifurcate its analysis between alleged misuses related to training the LLMs (which, based on the case law set forth below, are likely to be found transformative and fair use) and specific generative AI outputs (with a focus on whether such outputs are substantially similar to specific inputs). Addressing the latter example first, let us assume that an AI company has access to a copyrighted work without a license, uses such work to train their LLM, and then creates a generative AI output that is substantially similar to the copyrighted work. Unless the output is determined to be transformative or meets another requirement of fair use, then such output is likely to be found infringing under the current copyright framework. After all, to prove copyright infringement, the copyright holder merely needs to prove that the alleged infringer has access to the copyrighted work and creates a substantially similar work.

An AI company's use of unlicensed content to train their LLM *without* creating an output that is substantially similar to the underlying input presents a more nuanced analysis of intermediate copying and whether such copying amounts to copyright infringement. One argument against infringement is that an intermediate copy is not fixed in a tangible medium of expression and, therefore, is not a copy. However, in the context of training LLMs, it is likely that some type of copying, at least in a digital sense, is made during the training. As a result, the analysis will likely move to the second step, which will focus on whether the use of copyrighted material in training the LLMs is subject to a fair use exception.

Relevant case law from the Second Circuit, Ninth Circuit, and United States Supreme Court helps guide the fair use analysis. The below-referenced opinions support a likely finding of fair use as it relates to using unlicensed content to simply train LLMs.

In 2015, the Second Circuit found that Google's unauthorized digitizing of copyright-protected works, creation of a search functionality, and display of snippets of those works were non-infringing fair uses. *Authors Guild v. Google, Inc.*, 804 F.3d 202, 229 (2d. Cir. 2015). The Second Circuit reasoned that the purpose of the copying was highly transformative, the public display of the text was limited, and the revelations did not provide a significant market substitute for the protective aspects of the originals. *Id.* The Second Circuit found that the fact that Google's use was of a commercial nature and had a profit motivation, did not justify denial of fair use. *Id.* In *Sega Enterprises, Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992), the Ninth Circuit found copying software code to permit a video game to run on a console to be fair use. Likewise, in *Perfect 10, Inc. v. Google, Inc.*, 508 F.3d 1146 (9th Cir. 2007), the Ninth Circuit found a search engine's collection and display of thumbnail images to be a fair use.

In *Google LLC v. Oracle America, Inc.*, 141 S. Ct. 1183 (2021), the Supreme Court found that Google's copying of Oracle's Java SE API was a fair use of such material. The Supreme Court focused its fair use analysis on whether the use was transformative (*i.e.*, whether it adds something new, with a further purpose or different character). *Id.* at 1202-1203. The opinion noted that Google copied the API only insofar as needed to include tasks that would be useful in smartphone programs and only insofar as needed to allow programmers to call upon those tasks without discarding a portion of a familiar programming language and learning a new one, which supported a finding of fair use. *Id.* at 1203. These facts supported the fair use finding.

The Supreme Court's recent *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508 (2023), opinion similarly focused on whether the uses at issue were transformative. In *Goldsmith*, the Supreme Court found that, while the use of a copyrighted work may be fair if the use has a purpose and character that is sufficiently distinct from the original, the uses at issue before the Court were not transformative because they shared substantially the same commercial purpose (*i.e.*, to illustrate a magazine about Prince with a portrait of Prince). *Id.* at 541-550.

We expect AI companies to rely upon the fact that their uses of copyrighted works in training their LLMs have a further purpose or different character than that of the underlying content. At least one court in the Northern

District of California has rejected the argument that, because the plaintiffs' books were used to train the defendant's LLM, the LLM itself was an infringing derivative work. *See Kadrey v. Meta Platforms*, Case No. 23-cv-03417, Doc. 56 (N.D. Cal. 2023). The *Kadrey* court referred to this argument as "nonsensical" because there is no way to understand an LLM as a recasting or adaptation of the plaintiffs' books. *Id.* The *Kadrey* court also rejected the plaintiffs' argument that every output of the LLM was an infringing derivative work (without any showing by the plaintiffs that specific outputs, or portion of outputs, were substantially similar to specific inputs). *Id.*

While we expect significant clarity from courts over the coming year concerning the application of fair use to generative AI, at the end of the day well-funded industry leaders such as OpenAI will likely win regardless of the outcome. On the one hand, should OpenAI prevail across the board, it will owe nothing for the content used to train its LLM. On the other hand, if OpenAI is forced to license the content used to train its LLM, then such a finding will likely create an economy where only the most well-funded companies will be able to afford the licenses necessary to effectively train their LLMs.

If you have any questions about this topic, please contact Edward D. Lanquist, Jeremy D. Ray, or any member of Baker Donelson's Intellectual Property Group.