STATE BAR LITIGATION SECTION REPORT

THE ADVOCATE

Artificial Intelligence

VOLUME 109

WINTER

2024

POISONING THE WELL(M): PIRATED DATA, LARGE LANGUAGE MODELS, AND COPYRIGHT

BY JUSTIN A. NELSON, ROHIT D. NATH & J. CRAIG SMYSER

A. The Napster Problem

It's the year 2000 and you're 16 years old, exploring this new thing called the "internet." You open up Napster so you can download copies of 'N Sync's *No Strings Attached*, the best-selling album that year. Maybe you also get yourself a free copy of *Marshall Mathers LP*. You want to listen to it to enjoy it, and you're also looking for inspiration for your own garage boy band.

Did you commit copyright infringement? We know now the answer is yes. After years of litigation, the Ninth Circuit held that Napster was liable for enabling the infringing acts of its users. A necessary predicate of that decision was that Napster's users—people downloading free copies of music to listen to at home—were copyright infringers. *See A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001), as amended (Apr. 3, 2001).

Fast forward to today. While the Napster model is no longer mainstream, illegal enclaves of pirated literature, music, and art still persist across the internet. Today's patrons of pirated material are not teenagers who want to avoid paying $16 for a compact disc. Instead, they include multibillion-dollar companies developing artificial intelligence models, like OpenAI or Anthropic, which have an insatiable appetite for high-quality, copyrighted material. These companies say they need this material to "teach" their models how to write well, make quality music, or other art. Like the 16-year-olds of yesteryear, these AI companies procured copies of copyrighted material for free from pirated websites to avoiding paying the cost of a legal copy of it from their local book or record store.

While the Napster model is no longer mainstream, illegal enclaves of pirated literature, music, and art still persist across the internet.

Since the breakout success of ChatGPT in November 2022, copyright questions have loomed large. For academics, commentators, the tech community, and others, debates over AI's copyright compliance are normally expressed in terms of whether large language models ("LLMs") (and their training) represent "fair uses" of the training data. These debates have typically focused on whether reproducing copyrighted material to "train" an LLM is fair use, or whether an LLM itself constitutes a thinly-veiled copy of the training data.

These questions also provoke metaphysical questions about whether computer programs can truly "learn," or what differentiates "human learning" from computer algorithms. As attorneys for putative classes authors in copyright infringement lawsuits against OpenAI and their major competitor Anthropic, the reader will be unsurprised to learn that we think none of these "uses" is "fair."

But put those topics aside. What commentators miss in analyzing these higher-order questions is a much more basic act—one that need not invoke any soul-searching or require any stance on the ontological status of LLMs. That act is the initial acquisition and copying of the data in question.

Unlike public research programs regarding training and model-interpretability, some AI companies have been opaque about (1) what data they use to train their models and (2) how they obtained that data. That silence is in part a recognition that data-quality is the single biggest driver of model quality—and so keeping datasets secret is a key competitive edge. It also may reflect, however, a recognition of the tremendous copyright liability arising out of the *way* they acquired training data—namely, by taking it without permission from pirated sources.

B. Initial Acquisition as the Crux of Copyright Liability

Consider the following scenario: a remix-artist purchases a song. The remix-artist then produces a remix of the song and distributes it for sale. The copyright holder of the original song sues the remix-artist and challenges (1) the remixing and (2) the distribution of the remix. This was basically the situation which the Supreme Court considered in *Campbell*. While individual cases present unique issues, the "fair use" analysis is a well-worn application of the four statutory fac-

tors: (1) nature of the use, i.e. was the remixing sufficiently transformative, (2) the nature of the copyrighted work, (3) amount of the work used, and (4) the effect on the potential market for or value of the copyrighted work.

Now consider that same fact pattern, but with the following wrinkle: instead of purchasing the original song, the remix-artist illegally pirated the song. The copyright holder of the original song again sues. This time it challenges the *initial act* of piracy by which the remix-artist acquired the copy in question, in addition to the remixing and distribution of the remix. While the fair use question for the acts of remixing and distribution remains the same as in the previous example, no substantial fair use defense exists for the initial act of piracy. Nor would a court finding that the remix and/or its distribution was "fair use" constitute a defense to the initial act of piracy. The commonsense answer is that the initial acquisitive act itself is not a "use" amenable to a fair use defense.

And it's not just common sense. This is the unanimous conclusion of the federal courts. After the rise of file-sharing business models like Napster, the music industry sought to hold individual users accountable for copyright infringement. While the majority these such suits settled, music labels won two cases that went to trial, and those judgments were unanimously affirmed on appeal without any issue as to "fair use." *See, e.g.*, *Sony BMG Music Ent. v. Tenenbaum*, 660 F.3d 487, 500 (1st Cir. 2011); *Capitol Recs., Inc. v. Thomas-Rasset*, 692 F.3d 899, 906 (8th Cir. 2012). So too for companies that download massive quantities of copyrighted material from the same type of illegal websites. Whatever the arguments for or against fair use of the training data, that initial acquisition of known pirated material does not and should not have legal protection under the guise of fair use.

C. A Taxonomy of LLM Training Data and Acquisitive Acts

In general, training data for these models comes from a variety of sources, including:

1. Collections of works in the public domain for which any copyright protection has expired, like Bram Stroker's *Dracula*.

2. Data on the "open web" which is "scraped" by the AI Company or non-profits like Common Crawl;

3. Collections of works sourced from pirated repositories. For example, Books3, described in further detail below.

Category 1 poses no issue from an "initial-acquisition" perspective, as works in the public domain may be reproduced and exploited freely.

The second category, however is problematic. Particularly given the large volume of pirated material available on the open web, simply trawling without care or concern as to pirated data "bycatch" may demonstrate that the initial act of copying this data constituted a copyright violation.

Regardless of any defense with respect to web scraping (and we think none exists especially where a company knows its scrape includes pirated material), the final category—where an AI company acquires large repositories of pirated data—has no fair use defense. Take, for example, the most popular "open source" compilation of AI training data available today: a dataset known as "The Pile," a dataset used by the company Anthropic as alleged in the Complaint we filed on behalf of authors against the company. The Pile's authors noted that its goal was to replicate the data set which OpenAI used to train GPT-3. One of the many datasets in the original version of The Pile is a dataset called Books3. Books3 is a compilation of nearly 200,000 books, all sourced from a notorious pirated book collection called Bibliotik.

Why books? Here's what the firm that created The Pile, EleutherAI, said: "We included Bibiotik because books are invaluable for long-range context modeling research and storytelling. (See https://arxiv.org/pdf/2101.00027 (accessed Oct. 18, 2024)). In the AI world, "you are what you eat" is particularly salient. High quality, lengthy, coherent text as training material means a large language model will be better able to process longer and more complicated text inputs and generate longer text output that is coherent. But to get this high-value training material, companies like Anthropic or OpenAI, the allegations go, didn't approach authors or publishers or bookstores; they took them from corners of the web that few readers would ever even think to look.

With Books3, we have a dataset that (1) clearly consisted of pirated material, (2) which the AI companies, like Anthropic, are alleged to have downloaded. Whatever those companies did with that data after that point, the initial act of copying from a known pirated website was unlawful. In this way, the AI company is no different from the teenager using Napster, except that AI companies have downloaded illegal copies for a commercial purpose versus personal use—a factor that weighs only further *against* a finding of fair use. *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 511 (2023)

A number of suits are percolating through the federal courts advancing the theory that the acquisition of pirated datasets constitutes infringement. *See, e.g.*, *Authors Guild v. OpenAI Inc.*, No. 1:23-cv-8292-SHS (S.D.N.Y.); *Bartz v. Anthropic PBC*, No. 3:24-cv-5417-WHA (N.D. Cal.). To be sure, these suits also challenge the use of the data, however acquired, in training, but it is a mistake to think of these suits as rising or falling solely on the fairness of those uses. Focusing on the initial acquisition clarifies the liability and exposure of companies like OpenAI and Anthropic who are alleged to have knowingly downloaded pirated material.

* * *

The current slate of lawsuits against AI companies raise a variety of questions, and some of these questions are more difficult and more hotly contested than others. But one of those questions is straightforward under the law: Is it copyright infringement to make unlicensed copies of works by obtaining them from pirated sources? The answer to that question is yes. This issue was resolved long ago in the Napster era. The answer today should be no different. If a pimply 16-year-old is liable for an illegal download, so is the AI company worth $160 billion.

Justin A. Nelson, Rohit D. Nath, and J. Craig Smyser are attorneys of Susman Godfrey L.L.P. ★