LOS ANGELES & SAN FRANCISCO

Daily Journal

www.dailyjournal.com

MONDAY, JUNE 16, 2025

GUEST COLUMN

Why AI's Use of Shadow Libraries Should Alarm Us All

For $100,000 in crypto, Anna's Archive is offering AI companies high-speed access to 140 million pirated books and articles--fueling a digital gold rush where copyright law is ignored, piracy is rebranded as fair use, and the stakes for authors, courts, and the creative economy have never been higher.

By Margaux Poueymirou and Maxwell V. Pritt

For $100,000–payable in Bitcoin or other cryptocurrency –notorious "shadow library" Anna's Archive will sell you premier access to over 140 million pirated books and articles. At least 30 companies have already taken up Anna's Archive on its offer of "high-speed access" to copy-pirated works. While their identities are secret, "most are LLM companies" that have spent the last few years collecting and creating vast private digital libraries of pirated works for commercial use with their Large Language Models (LLMs). (https://annas-archive.org/blog/ai-copyright.html). Whether OpenAI, Anthropic, or Meta (to name a few) are among these companies remains unknown. What is now known is that each of these companies used illegal shadow libraries to source and copy high-quality AI training data: copyrighted books and other literature–or, more accurately, their pirated market substitutes.

The cost-benefit analysis was criminally simple, particularly for companies caught flat-footed by OpenAI's release of ChatGPT: expend time and resources to source and buy or license tens of millions of books and papers; or take them all for free now and either pay damages later–or not at all if they can convince courts to excuse commercial piracy as fair use. The rapid rise of AI technology has thus ushered in a new era of digital piracy on a scale never before seen. Indeed, Anna's Archive boasts that AI saved illegal "shadow libraries:"

This art was created with the assistance of Shutterstock AI tools

"Not too long ago, 'shadow-libraries' were dying. Sci-Hub, the massive illegal archive of academic papers, had stopped taking in new works, due to lawsuits. 'Z-Library,' the largest illegal library of books, saw its alleged creators arrested on criminal copyright charges. They incredibly managed to escape their arrest, but their library is no less under threat.

[. . .]

Then came AI. Virtually all major companies building LLMs contacted us to train on our data. Most (but not all!) US-based companies reconsidered once they realized the illegal nature of our work. By contrast, Chinese firms have enthusiastically embraced our collection, apparently untroubled by its legality. This is notable given China's role as a signatory to nearly all major international copyright treaties." (https://annas-archive.org/blog/ai-copyright.html)

The United States is also a signatory to these international copyright treaties, including the 1996 World Intellectual Property Organization (WIPO) treaties, which confronted the novel issues facing copyright enforcement in the digital age. (U.S. Copyright Office. "International Issues," https://www.copyright.gov/international-issues/). Shadow libraries not only contravene the U.S. Copyright Act but also exemplify the problems Congress contemplated when enacting the Digital Millennial Copyright Act (DMCA), which sought to implement the WIPO treaties and combat digital piracy.

Every year, millions of individuals copy and distribute pirated books from shadow libraries like Library Genesis (LibGen) or Z-Library (Z-lib) through various methods, including decentralized file-sharing systems like BitTorrent and the Inter-Planetary File System (IPFS), as well as through direct downloading. Whenever an individual uses a shadow library in lieu of a bookstore to acquire a book, that book's author and publisher are harmed by the loss of a sale in an otherwise functioning and well-established market for books. More-over, as the documented in its survey of "Notorious Markets for Counterfeiting and Piracy," which included LibGen in two recent editions, the harms of piracy are undulating, impacting economies reliant on legitimate markets–in the case of books, artists and graphic de-signers, bookstores, publishers,

printing presses, and copyeditors, and others working in creative economies. "USTR Releases 2024 Review of Notorious Markets for Counterfeiting and Privacy." https://ustr.gov/about/policy-offices/press-office/ustr-archives/2007-2024-press-releases/ustr-releases-2024-review-notorious-markets-counterfeiting-and-piracy.

How can illegal online databases that traffic in piracy and are permanently enjoined from operating simultaneously exist as legitimate sources of AI training data? And what are the ramifications of permitting major companies to use illegal online databases to source AI training data, while the FBI and the Department of Homeland Security actively attempt to shutter these websites and their domains? Several district courts are currently grappling with these surprising questions–including in two highly watched cases pending in the Northern District of California against Meta and Anthropic. While everyone expected these historic AI copyright cases to test the boundaries of the fair use defense, no one anticipated that billion-dollar companies would be arguing–and courts would be contemplating–that online *piracy* could also fall within the ambit of the fair use defense. But as these cases progressed, it became clear that companies had abandoned efforts to source books as AI training data through legitimate means and turned to piracy, sometimes using peer-to-peer file sharing to acquire the books and, in the case of Meta, even distributing tens of millions of pirated books to fellow pirates in order to scale the company's acquisition of books. If any of the companies prevail in these cases, it will be the first time in history that piracy and trafficking in stolen goods are given a pass under the fair use doctrine.

But if it seems challenging to square the fair use defense with rampant piracy, that's because it is. The U.S. Supreme Court recently reiterated that fair use is an equitable doctrine that aims to ensure that the Copyright Act is not applied so rigidly that it undermines its goal of "promoting broad public availability of literature, music, and the other arts." *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 526 (2023). Courts thus consider four non-exhaustive factors in evaluating the fair use defense, though two tend to be most dispositive. The first focuses on whether a defendant's copying of protected expression is necessary to achieve a "distinct" and "transformative purpose"–not to be confused with mere "transformation"–and this analysis is balanced against the commercial nature of the use, as well as a consideration of "bad faith." Transformativeness is thus one part of a three-part inquiry under the first factor and serves as a shorthand for when there is a compelling justification for targeting the protected work at issue. *See, e.g.*, Shyamkrishna Balganesh & Peter S. Menell, Going "Beyond" Mere Transformation, 47 Colum. J.L. & Arts 413, 417-19 (2024). Without a showing of transformativeness, however, a fair use defense will almost assuredly fail–and it is this showing that AI companies are banking on across all pending copyright litigation. But "the single most important element" is market harm or "the effect of the use upon the potential market for or value of the copyrighted work." *Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 566 (1985). This is why the use of illegal pirate websites to pilfer copyrighted works should doom defendants in these cases.

Had AI companies licensed millions of books from libraries (like Google did over two decades ago when it partnered with major research libraries to digitize books and build its books search engine), or even bought millions of books and then scanned them to use for AI training data, the initial acquisition would not have harmed any book sales market. While exceeding the terms of a license in one case could amount to fair use, it may not in another. *See Am. Geophysical Union v. Texaco, Inc.*, 60 F.3d 913 (2d Cir. 1995) (rejecting a fair use defense where Texaco exceeded its subscription terms by copying scientific journal articles for research aimed at developing new products and technologies). But by using illegal websites to source the books, plaintiffs in these cases can advance two separate theories of market harm: one based on the well-established books sales market, and the other on the burgeoning licensing market for AI training data. This alone belies defendants' claims that training on a legally acquired copy of a book versus a pirated copy of a book is legally insignificant. There is also nothing transformative about pirating a book: It is the same as the original, and the specific use of piracy is to get for free what you otherwise have to buy or license–a use that is squarely incompatible with the "principles of good faith and fair dealing" that courts consider when evaluating "bad faith" under factor one. *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1164 n.8 (9th Cir. 2007) (citing *Harper & Row*, 471 U.S. at 562-63). Indeed, while the U.S. Supreme Court has expressed skepticism about the role of bad faith in the fair use analysis, that skepticism has focused exclusively on whether evidence of mal intent matters. But evidence of piracy is not simply evidence of questionable intent; it is objective "bad faith" conduct that undermines the very goals of the Copyright Act.

While courts may ultimately conclude that one or more of the many other copies and uses AI companies make of copyrighted books are transformative, or that there is no existing or potential market for licensing copyrighted books as AI training data, it would upend well-established public policy and decades of case-law criminalizing online piracy to permit massive technology companies to claim fair use for brazen piracy simply because it is AI-related. As U.S. District Judge William Alsup stated recently during oral argument in *Anthropic*:

"If somebody downloads from Napster a copy of a song, you could go to prison for that. That is a copyright violation. And now if you download from a pirate site a copy of somebody's book, standing alone, you could go to prison for that too. So ... I have a hard time seeing that you can commit what is ordinarily a crime and yet get exonerated because you wind up using it for transformative use."

But if federal courts buy defendants' ends-justifies-the-means argument, authors, publishers and the public will be facing a brave new world of unbridled copyright infringement–ironically committed by some of the world's largest corporations building highly commercial and profitable products. Such a holding would sanction online piracy, permitting shadow libraries to commercialize and monetize pirated books obtained through theft–just as Anna's Archive is doing today. And why stop with books?

This article is the first in a series on AI copyright issues.

Margaux Poueymirou *and* **Maxwell V. Pritt** *are partners at Boies Schiller Flexner LLP.*