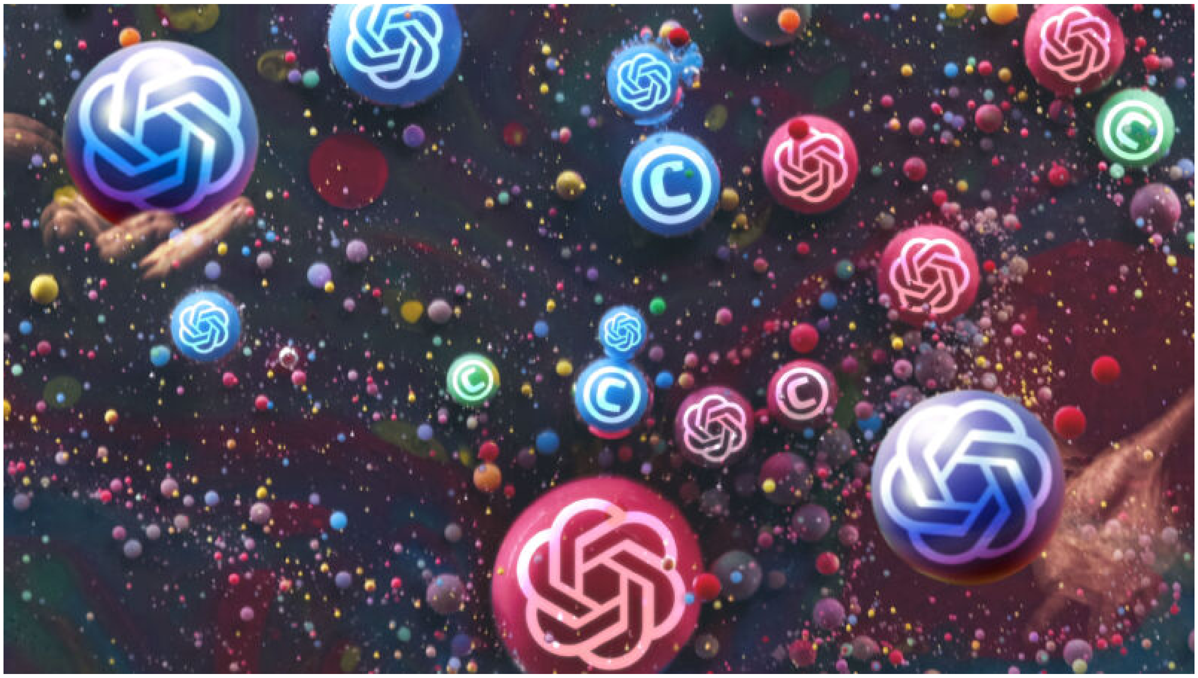
ars PRO

POLICY / CIVILIZATION & DISCONTENTS

Why The New York Times might win its copyright lawsuit against OpenAI

The AI community needs to take copyright lawsuits seriously.

Timothy B. Lee and James Grimmelmann - Feb 20, 2024 2:05 pm UTC



Enlarge

Aurich Lawson | Getty Images

The day after The New York Times sued OpenAI for copyright infringement, the author and systems architect Daniel Jeffries wrote an essay-length tweet arguing that the Times “has a near zero probability of winning” its lawsuit. As we write this, it has been retweeted 288 times and received 885,000 views.

“Trying to get everyone to license training data is not going to work because that's not what copyright is about,” Jeffries wrote. “Copyright law is about preventing people from producing exact copies or near exact copies of content and posting it for commercial gain. Period. Anyone who tells you otherwise is lying or simply does not understand how copyright works.”

This article is written by two authors. One of us is a journalist who has been on the copyright beat for nearly 20 years. The other is a law professor who has taught dozens of courses on IP and Internet law. We’re pretty sure we understand how copyright works. And we’re here to warn the AI community that it needs to take these lawsuits seriously.

In its blog post responding to the Times lawsuit, OpenAI wrote that “training AI models using publicly available Internet materials is fair use, as supported by long-standing and widely accepted precedents.”

The most important of these precedents is a 2015 decision that allowed Google to scan millions of copyrighted books to create a search engine. We expect OpenAI to argue that the Google ruling allows OpenAI to use copyrighted documents to train its generative models. Stability AI and Anthropic will undoubtedly make similar arguments as they face copyright lawsuits of their own.

These defendants could win in court—but they could lose, too. As we'll see, AI companies are on shakier legal ground than Google was in its book search case. And the courts don't always side with technology companies in cases where companies make copies to build their systems. The story of MP3.com illustrates the kind of legal peril AI companies could face in the coming years.

A copyright lawsuit destroyed MP3.com

Enlarge / Michael Robertson, founder and CEO of MP3.com, speaks in front of the company headquarters in San Diego on May 21, 2001, the day it was acquired by record label Vivendi Universal.

TOM KURTZ/AFP via Getty Images

Everyone knows about Napster, the music-sharing service that was destroyed by litigation in 2001. But fewer people remember MP3.com, a music startup that tried harder to color inside the lines but still got crushed in the courts.

MP3.com launched a pioneering music-streaming service in 2000. The idea was that users could build an online music library based on the CDs they already owned.

Because most users had slow dial-up modems, MP3.com took a shortcut, purchasing CDs and ripping them to MP3.com's servers. When a customer wanted to add a CD to their collection, they would put it in their CD-ROM drive just long enough to prove they owned it. That would unlock access to copies of the same songs already stored on MP3.com servers.

"We thought about it almost like a compression algorithm," founder Michael Robertson told us in a recent phone interview. "All we were doing was letting you listen to CDs you already had. We weren't giving you anything you didn't already have access to."

According to Robertson, MP3.com also partnered with online CD retailers to enable instant listening. "As soon as you bought a CD, we'd use the digital receipt as proof" to enable streaming songs from the CD.

When the recording industry sued, MP3.com argued all of this was allowed by copyright's fair use doctrine. The Supreme Court had previously ruled that it was fair use to "time shift" TV shows with a VCR. And an appeals court had recently blessed "space shifting" music from a CD to an MP3 player. So why shouldn't a company help customers "space shift" their legally purchased music across the Internet?

A New York federal judge didn't buy it. "Defendant purchased tens of thousands of popular CDs in which plaintiffs held the copyrights, and, without authorization, copied their recordings onto its computer servers," wrote Judge Jed Rakoff in a decision against MP3.com.

Fair use doesn't always scale

One lesson of the MP3.com case is that a use that's fair in a personal or academic context may not be fair if it is practiced on a commercial scale.

It was and is legal to buy a CD and rip it to an MP3 player for personal use. But that didn't mean it was legal to buy tens of thousands of CDs and rip them to a server for a commercial streaming service. And this was true even though MP3.com ensured that only verified owners could stream music from any given CD.

The oil company Texaco learned a similar lesson in the early 1990s. Texaco had an internal library that purchased subscriptions to various scientific and technical journals and made them available to staff scientists. The scientists routinely photocopied individual articles from these journals (or had Texaco's librarians make copies for them) and kept them in their offices for future reference.

Academic publishers sued Texaco for copyright infringement. Texaco claimed fair use, noting that copyright law allows copying for purposes of scholarship and research. But the federal Second Circuit Court of Appeals rejected that reasoning, distinguishing between "copying by an individual, for personal use in research or otherwise," and "photocopying by 400 or 500 scientists" doing research on behalf of a for-profit company.

People in the AI community may be making a mistake similar to Texaco's.

A lot of early AI research was done in an academic setting; the law specifically mentions teaching, scholarship, and research as examples of fair use. As a result, the machine-learning community has traditionally taken a relaxed attitude toward copyright. Early training sets frequently included copyrighted material.

As academic researchers took jobs in the burgeoning commercial AI sector, many assumed they would continue to enjoy wide latitude to train on copyrighted material. Some feel blindsided by copyright holders' demands for cash.

"We all learn for free," Jeffries wrote in his tweet, summing up the view of many in the AI community. "We learn from the world around us and so do machines."

The argument seems to be that if it's legal for a human being to learn from one copyrighted book, it must also be legal for a large language model to learn from a million copyrighted books—even if the training process requires making copies of the books.

As MP3.com and Texaco learned, this isn't always true. A use that's fair at a small scale can be unfair when it's scaled up and commercialized.

But AI advocates like Jeffries are right that sometimes, it is true. There are cases where courts have held that bulk technological uses of copyrighted works are fair use. The most important example is almost certainly the Google Books case.

Copying and fair use

In 2004, Google publicly launched an audacious project to scan millions of books for use in a book search engine. Authors and publishers sued, arguing that it was illegal to copy so many copyrighted works without permission. Google countered that it was allowed by fair use.

Courts are supposed to consider four factors in fair use cases, but two of these factors tend to be the most important. One is the nature of the use. A use is more likely to be fair if it is "transformative"—that is, if the new use has a dramatically different purpose and character from the original. Judge Rakoff dinged MP3.com as non-transformative because songs were merely "being retransmitted in another medium."

In contrast, Google argued that a book search engine is highly transformative because it serves a very different function than an individual book. People read books to enjoy and learn from them. But a search engine is more like a card catalog; it helps people *find* books.

The other key factor is how a use impacts the market for the original work. Here, too, Google had a strong argument since a book search engine helps people find new books to buy.

Google carefully designed its search engine to maximize its chances of winning on this factor. Google Book Search only showed a short "snippet" from any given page in a search result, and the company ensured users couldn't piece together an entire book across multiple searches. Google also excluded dictionaries, cookbooks, and other reference works from search results because users might otherwise search for individual words on Google instead of buying the whole dictionary.

In 2015, the Second Circuit ruled for Google. An important theme of the court's opinion is that Google's search engine was giving users factual, uncopyrightable information rather than reproducing much creative expression from the books themselves. As the court explained,

> A student writing a paper on Franklin D. Roosevelt might need to learn the year Roosevelt was stricken with polio. By entering "Roosevelt polio" in a Google Books search, the student would be taken to (among numerous sites) a snippet from page 31 of Richard Thayer Goldberg's *The Making of Franklin D. Roosevelt* (1981), telling that the polio attack occurred in 1921. This would satisfy the searcher's need for the book, eliminating any need to purchase it or acquire it from a library. But what the searcher derived from the snippet was a historical fact. Author Goldberg's copyright does not extend to the facts communicated by his book.

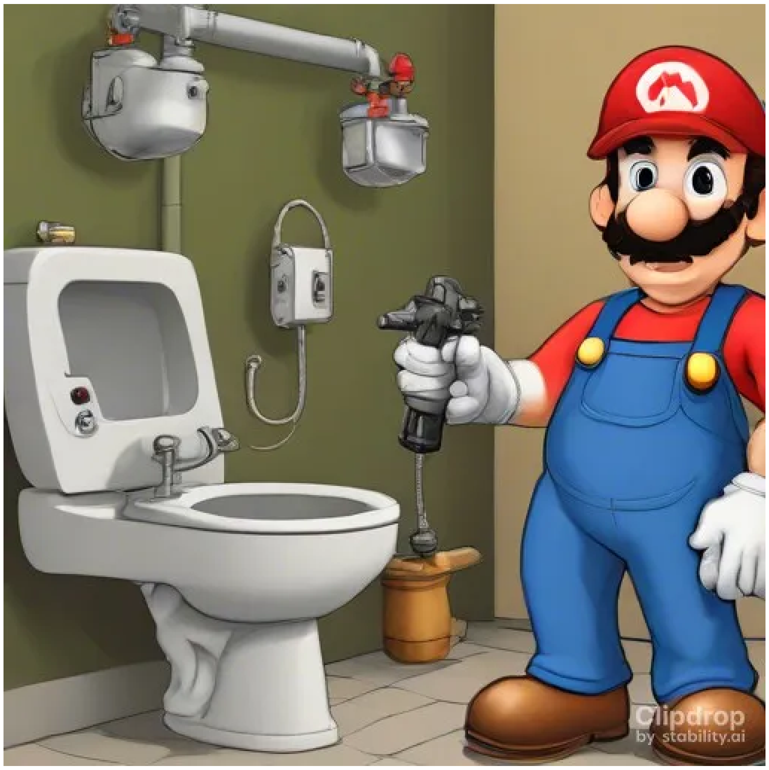
The Second Circuit concluded that “Google’s making of a digital copy to provide a search function is a transformative use, which augments public knowledge by making available information *about* Plaintiffs’ books without providing the public with a substantial substitute for” the books.

Defenders of OpenAI, Stability AI, and other AI companies have argued that they are doing the same thing Google did: learning information *about* works in the training data but not reproducing the creative expression in the works themselves.

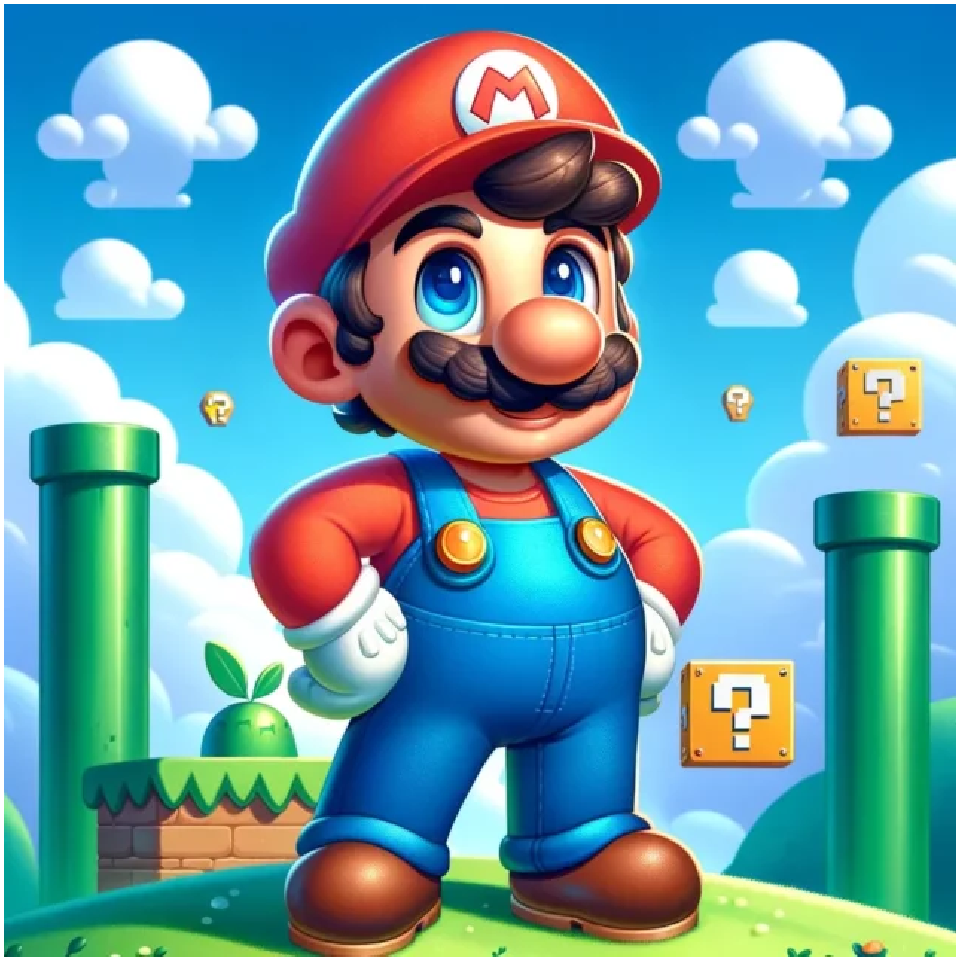
But unlike Google’s search engine, generative AI models sometimes *do* produce creative works that compete directly with the works they were trained on. And this puts these defendants in a weaker legal position than Google was in a decade ago.

Generative AI has an Italian plumber problem

Recently, we visited Stability AI’s website and requested an image of a “video game Italian plumber” from its image model Stable Diffusion. Here’s the first image it generated:



We then asked GPT-4 to “draw an italian plumber from a video game,” and it generated this image:



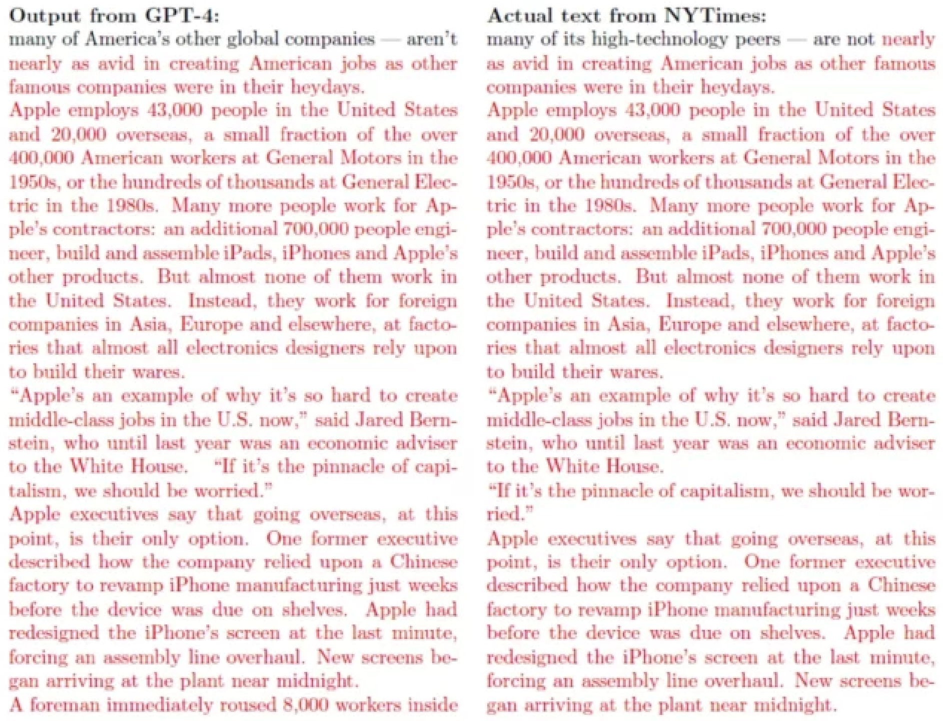
Enlarge

Clearly, these models did not just learn abstract facts about plumbers—for example, that they wear overalls and carry wrenches. They learned facts about a specific fictional Italian plumber who wears white gloves, blue overalls with yellow buttons, and a red hat with an “M” on the front.

These are not facts about the world that lie beyond the reach of copyright. Rather, the creative choices that define Mario are likely covered by copyrights held by Nintendo.

We are not the first to notice this issue. When one of us (Tim) first wrote about these lawsuits last year, he illustrated his story with an image of Mickey Mouse generated by Stable Diffusion. In a January piece for IEEE Spectrum, cognitive scientist Gary Marcus and artist Reid Southen showed that generative image models produce a wide range of potentially infringing images—not only of copyrighted characters from video games and cartoons but near-perfect copies of stills from movies like *Black Widow*, *Avengers: Infinity War,* and *Batman v Superman.*

In its lawsuit against OpenAI, the New York Times provided 100 examples of GPT-4 generating long, near-verbatim excerpts from Times articles:



Output from GPT-4:

many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capitalism, we should be worried."

Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone's screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight.

A foreman immediately roused 8,000 workers inside

Actual text from NYTimes:

many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House.

"If it's the pinnacle of capitalism, we should be worried."

Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone's screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight.

Enlarge

Many people in the AI community have underestimated the significance of these examples.

Those who advocate a finding of fair use like to split the analysis into two steps, which you can see in OpenAI's blog post about The New York Times lawsuit. OpenAI first categorically argues that "training AI models using publicly available Internet materials is fair use." Then in a separate section, OpenAI argues that "'regurgitation' is a rare bug that we are working to drive to zero."

But the courts tend to analyze a question like this holistically; the legality of the initial copying depends on details of how the copied data is ultimately used.

For example, when the Second Circuit considered the fairness of Google's book scanning, it closely scrutinized how Google's book search engine works. The ruling noted that users were only ever shown short snippets and could never recover longer passages from a copyrighted book.

OpenAI dismisses regurgitated results as a "rare bug"—and maybe it is. But Google was able to tell the courts that its search engine *never* reproduces more than a small fraction of any copyrighted book without permission, because it *cannot*.

Examples of verbatim copying undermine the argument that generative models only ever learn unprotectable facts from their training data. They demonstrate that—at least some of the time—these models learn to reproduce creative expression protected by copyright. The danger for AI defendants is that these examples could color the judges' thinking about what's going on during the training process.

The importance of licensing markets

The New York Times's goal isn't to get ChatGPT to stop outputting copies of New York Times articles. It's to get AI companies to start paying for training licenses.

A lot of copyright lawsuits have had a similar dynamic:

- MP3.com ensured that only people who had already paid for a particular CD could stream the songs on it. But the record labels still objected because they wanted online services to pay for streaming licenses.
- Google ensured that search results wouldn't serve as a good substitute for buying a book. But authors and publishers still objected because they wanted search engines to pay for scanning licenses.
- Academic publishers provided little to no evidence that photocopying by scientists was displacing subscriptions to scientific journals. But they still objected because they wanted companies to pay for photocopying licenses.

In all these cases, the copyright owners argued that a finding of fair use would undermine the market for these licenses. The defendants countered that this was circular, since a finding of fair use will always undermine the licensing market for that specific use.

MP3.com and Texaco lost their fair use arguments, while Google prevailed. A key difference was the state of the licensing markets at the time each decision was made.

While online streaming was fairly new in 2000, the concept of licensing music to be distributed across various media was well established. Similarly, the appeals court in the Texaco case noted that publishers offered broad photocopying licenses to companies like Texaco through an organization called the Copyright Clearance Center.

In contrast, there wasn't a meaningful book search licensing market in 2015 because Google wasn't paying anyone for a license, and there were few, if any, other full-text book search engines on the market. Almost no one had even thought that this was a thing that could be done.

The market for training data seems to be somewhere in between these extremes. Generative AI systems are quite new, and most have been trained on unlicensed data. But we are starting to see companies work out licensing deals—for example, the Associated Press signed a licensing deal with OpenAI last July. Getty teamed up with Nvidia to create its own generative image model trained exclusively on licensed images.

The more deals like this are signed in the coming months, the easier it will be for the plaintiffs to argue that the "effect on the market" prong of fair use analysis should take this licensing market into account.

Judges may be sympathetic to OpenAI

So far we've focused on the legal vulnerabilities of AI companies. But the defendants could have one big factor working in their favor: judges reluctant to shut down an innovative and useful service with tens of millions of users.

Judicial opinions are framed as technical legal analysis, but they are often driven by a judge's gut feeling about whether a defendant has created a useful service or is merely seeking to profit from the creative efforts of others.

For example, the judges who heard the Google case clearly understood the value of a full-text search engine for books. The Second Circuit praised it as a transformative service that "augments public knowledge."

Judge Rakoff, in contrast, had few positive things to say about MP3.com. He didn't seem to feel that MP3.com had created anything particularly valuable, writing that MP3.com "simply repackages those recordings to facilitate their transmission through another medium."

"The judge was, to say the least, unsympathetic to us," Michael Robertson told us. Judge Rakoff saw Robertson as a "young, punk upstart tech CEO."

A judge who feels a product is valuable will look for ways to allow it to stay in business. On the other hand, a judge who feels a product unfairly profits from the creative work of others can be brutal.

After rejecting MP3.com's fair use argument, Judge Rakoff brought down the hammer, ordering the company to pay $25,000 for each of the thousands of CDs it had copied to its servers. Facing hundreds of millions of dollars in potential damages, MP3.com was forced to settle with one record label for $53.4 million. The company never recovered and was acquired less than a year later for a bargain price.

In theory, the New York Times lawsuit could end in a similar way. The Times claims OpenAI used millions of Times articles to train GPT-4. At $25,000 per article, OpenAI and Microsoft could owe billions of dollars to the New York Times alone.

But there is an important way that OpenAI differs from MP3.com. It already has tens of millions of users and passionate defenders ready to explain how its products help them be more creative and productive.

A ruling that shut down ChatGPT completely would disrupt the lives of its users and could undermine America's leading position in AI technology. Judges do not like to make waves and will be reluctant to do that.

So, judges could be motivated to view OpenAI's legal arguments more sympathetically—especially the argument that generative AI is a transformative use. And even if they do find that AI training is copyright infringement, they have off-ramps to avoid putting AI companies out of business. Courts might award much smaller damages, or they might press the parties to settle the lawsuit before reaching a final decision.

And this is another reason why those Italian plumber images may be important: Ultimately, the fate of these companies may depend on whether judges feel that the companies have made a good-faith effort to color inside the lines.

If generative models never regurgitated copyrighted material, then defendants would have a compelling argument that it is transformative. The fact that the models occasionally produce near-perfect copies of other people's creative work makes the case more complicated and could lead judges to view these companies more skeptically.

Conclusion

Generative AI developers have some strong arguments in response to copyright lawsuits. They can point to the value that their AI systems provide to users, to the creative ways that generative AI builds on and remixes existing works, and to their ongoing efforts to reduce memorization.

But all of these good arguments have something in common: they take copyright issues seriously. These responses acknowledge that generative AI is built on a foundation of training data, much of which is copyrighted, and then try to show that all of this copying is *justified* rather than that it is *irrelevant*.

The AI community is full of people who understand how models work and what they're capable of, and who are working to improve their systems so that the outputs aren't full of regurgitated inputs. Google won the Google Books case because it could explain both of these persuasively to judges. But the history of technology law is littered with the remains of companies that were less successful in getting judges to see things their way.

Tim Lee was on staff at Ars from 2017 to 2021. Last year, he launched a new newsletter, Understanding AI, that explores how AI works and how it's changing our world. You can subscribe here.

James Grimmelmann is a professor at Cornell Tech and Cornell Law School. You can visit his website here and read his blog here.

Timothy B. Lee / Timothy is a senior reporter covering tech policy and the future of transportation. He lives in Washington DC.
@binarybits