OpenAI searches for an answer to its copyright problems

The huge leaps in OpenAI's GPT model probably came from sucking down the entire written web. That includes entire archives of major publishers such as Axel Springer, Condé Nast, and *The Associated Press* — without their permission. But for some reason, OpenAI has announced deals with many of these conglomerates *anyway.*

At first glance, this doesn't entirely make sense. Why would OpenAI pay for something it already had? And why would publishers, some of whom are lawsuit-style angry about their work being stolen, agree?

I suspect if we squint at these deals long enough, we can see one possible shape of the future of the web forming. Google has been referring less and less traffic outside itself — which threatens the existence of the entire rest of the web. That's a power vacuum in search that OpenAI may be trying to fill.

a:hover]:shadow-highlight-franklin dark:[&>a:hover]:shadow-highlight-franklin [&>a]:shadow-underline-black dark:[&>a]:shadow-underline-white">*The New York Times* is prepared to litigate

If OpenAI ingested the entirety of the text-based internet, that means a couple things. First, that there's no way to generate that volume of data again anytime soon, so that may limit any further leaps in usefulness from ChatGPT. (OpenAI notably has not yet released GPT-5.) Second, that a lot of

people are pissed.

Many of those people have filed lawsuits, and the most important was filed by *The New York Times.* The *Times'* lawsuit alleges that when OpenAI ingested its work to train its LLMs, it engaged in copyright infringement. Moreover, the product OpenAI created by doing this now competes with the *Times* and is meant to "steal audiences away from it."

The *Times'* lawsuit says that it tried to negotiate with OpenAI to permit the use of its work, but those negotiations failed. I'm going to take a wild guess based on the math I did above and say it's because OpenAI offered insultingly low sums of money to the *Times.* Its excuse? Fair use – a provision that allows the unlicensed use of copyrighted material under certain circumstances.

Should the newspaper win its case, OpenAI is going to have to pay an absolute minimum of $7.5 billion *in statutory damages alone*

If the *Times* wins its lawsuit, it may be entitled to statutory damages, which start at $750 per work. (I know those figures because – as you may have guessed from my use of "statutory" – they are dictated by law. The paper is also asking for compensatory damages, restitution, and attorneys' fees.) The *Times* says that OpenAI ingested 10 million total works – so that's an absolute minimum of $7.5 billion *in statutory damages alone*. No wonder the *Times* wasn't going to cut a deal in the single-digit millions.

So when OpenAI makes its deals with publishers, they are, functionally, settlements that guarantee the publishers won't sue OpenAI as the *Times* is doing. They are also structured so that OpenAI can maintain its previous use of the publishers' work is fair use – because OpenAI is going to have to argue that in multiple court cases, most notably the one with the *Times*.

“I do have every reason to believe that they would like to preserve their rights to use this under fair use,” says Danielle Coffey, the CEO of the News Media Alliance. “They wouldn’t be arguing that in a court if they didn’t.”

It seems like OpenAI is hoping to clean up its reputation a little. If you’re introducing a new product you want people to pay for, it simply can’t come with a ton of baggage and uncertainty. And OpenAI does have baggage: to make its fair use defense, it must admit to taking *The New York Times’* copyrighted material without permission — which implicitly suggests it’s taken a lot of other copyrighted material without permission, too. Its argument is just that it is legally entitled to do that.

There’s also a question of accuracy. At this point, we all know generative AI makes stuff up. The publisher deals don’t just provide legitimacy — they may also help feed generative AI information that is less likely to result in embarrassing errors.

a:hover]:shadow-highlight-franklin dark:[&>a:hover]:shadow-highlight-franklin [&>a]:shadow-underline-black dark:[&>a]:shadow-underline-white”>Copywrong

A ruling in favor of *The New York Times* can potentially help both Google and OpenAI, as well as Microsoft, which is backing OpenAI. Maybe this was what Eric Schmidt, former Google CEO, meant when he said entrepreneurs should do whatever they want with copyrighted work and “hire a whole bunch of lawyers to go clean the mess up.”

Courts are unpredictable when it comes to copyright law because it kind of works like porn — judges know a violation when they see it. Plus, if there is indeed a trial between *The New York Times* and OpenAI, there will almost certainly be an

appeal on the verdict, no matter who wins.

Court cases take time, and appeals take more time. It will be years before the courts sort all this out. And that's plenty of time for a player like OpenAI to develop a dominant business.

She specifically cites Google as being so big that it can force publishers into its terms

Let's say OpenAI eventually loses. That means all creators of large language models have to pay out. That can get very expensive, very fast — meaning that only the biggest players will be able to compete. It ensconces every established player and potentially destroys a number of open-source LLMs. That makes Google, Microsoft, Amazon, and Meta even more important in the ecosystem than they already dominate — as well as OpenAI and Anthropic, both of which have deals with some of the major players.

There's also some precedent in how big tech companies navigate the rulings against them, says the News Media Alliance's Coffey. She specifically cites Google as being so big that it can force publishers into its terms; as if to underscore her point, a few weeks after our interview, Google was legally declared a monopoly in an antitrust case.

Here's an example of Google's outsize power: In 2019, the EU gave digital publishers the right to demand payment when Google used snippets of their work. This law, first implemented in France, resulted in Google telling publishers it would use only headlines from their work rather than pay. "And so they sent a bunch of letters to French publications, saying waive your copyright protection if you want to be found," Coffey said. "They're almost above the law in that sense" because Google Search is so dominant.

Google is currently using its search dominance to squeeze publishers in a similar way. Blocking its AI from summarizing

people's work means that Google simply won't list them at all, because it uses the same tool to scrape for web search and AI training.

"That would be a real anticompetitive tragedy at the beginning of the ecosystem."

So if the *Times* wins, it seems possible that Google and other major AI players could still demand deals that don't benefit publishers much — while also destroying competing LLMs. "I'm incredibly worried about the possibility that we are setting up an ecosystem where the only people who are going to be able to afford training data are the biggest companies," says Nicholas Garcia, policy counsel at Public Knowledge.

In fact, the existence of the suit may be enough to discourage some players from using publicly accessible data to train their models. People might perceive that they can't train on publicly available data — narrowing competitive dynamics even farther than the bottlenecks that already exist with the supply of compute and experts. "That would be a real anticompetitive tragedy at the beginning of the ecosystem," Garcia says.

OpenAI isn't the only defendant in the *Times* case; the other one is its partner, Microsoft. And if OpenAI does have to pay out a settlement that is, at *minimum*, hundreds of millions of dollars, that might open it up to an acquisition from Microsoft — which then has all the licensing deals that OpenAI already negotiated, in a world where the licensing deals are required by copyright law. Pretty big competitive advantage. Granted, right now, Microsoft is pretending it doesn't really know OpenAI because of the government's newfound interest in antitrust, but that could change by the time the copyright cases have rolled through the system.

And OpenAI may lose because of the licensing deals it negotiated. Those deals created a market for the publishers'

data, and under copyright law, if you're disrupting such a market, well, that's *not* fair use. This particular line of argument most recently came up in a Supreme Court case about an Andy Warhol painting that was found to unfairly compete with the original photograph used to create the painting.

The legal questions aren't the only ones, of course. There's something even more basic I've been wondering about: do people want answer engines, and if so, are they financially sustainable? Search isn't just about finding answers — Google is a way of finding a specific website without having to memorize or bookmark the URL. Plus, AI is expensive. OpenAI might fail because it simply can't turn a profit. As for Google, it could be broken up by regulators because of that monopoly finding.

In that case, maybe the publishers are the smart ones after all: getting the money while the money's still good.

https://www.theverge.com/2024/8/30/24230975/openai-publisher-deals-web-search