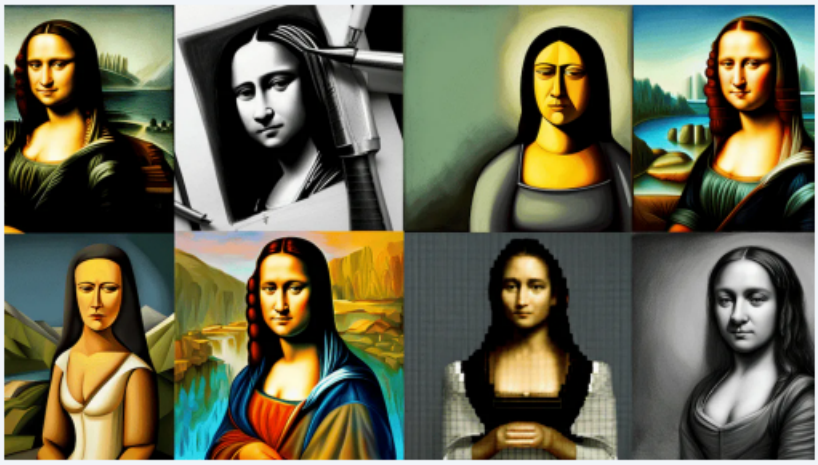
The Batch > AI & Society > Article

Copyright Owners Take AI to Court

Artists and writers sue big tech companies over copyright infringement.

AI & Society | Generative Media | Generative Modeling | Regulations | Google | OpenAI | ChatGPT | Stability AI | Midjourney | Github | Microsoft



AI models that generate text, images, and other types of media are increasingly under attack by owners of copyrights to material included in their training data.

What's happening: Writers and artists filed a new spate of lawsuits alleging that AI companies including Alphabet, Meta, and OpenAI violated their copyrights by training generative models on their works without permission. Companies took steps to protect their interests and legislators considered the implications for intellectual property laws.

Lawsuits and reactions: The lawsuits, which are ongoing, challenge a longstanding assumption within the AI community that training machine learning models is allowed under existing copyright laws. Nonetheless, OpenAI responded by cutting deals for permission to use high-quality training data. Meanwhile, the United States Senate is examining the implications for creative people, tech companies, and legislation.

- Unnamed plaintiffs sued Alphabet claiming that Google misused photos, videos, playlists, and the like posted to social media and information shared on Google platforms to train Bard and other systems. One alleged that Google misused a book she wrote. The plaintiffs filed a motion for class-action status. This action echoes an earlier lawsuit against OpenAI filed in June.
- Comedian Sarah Silverman joined authors Christopher Golden and Richard Kadrey in separate lawsuits against Meta and OpenAI in a United States federal court. The plaintiffs, who are seeking class-action status, claim that the companies violated their copyrights by training LLaMA and ChatGPT, respectively, on books they wrote.
- In a similar lawsuit authors Paul Tremblay and Mona Awad allege that OpenAI violated their copyrights.
- OpenAI agreed to pay Associated Press for news articles to train its algorithms — an arrangement heralded as the first of its kind. OpenAI will have access to articles produced since 1985, and

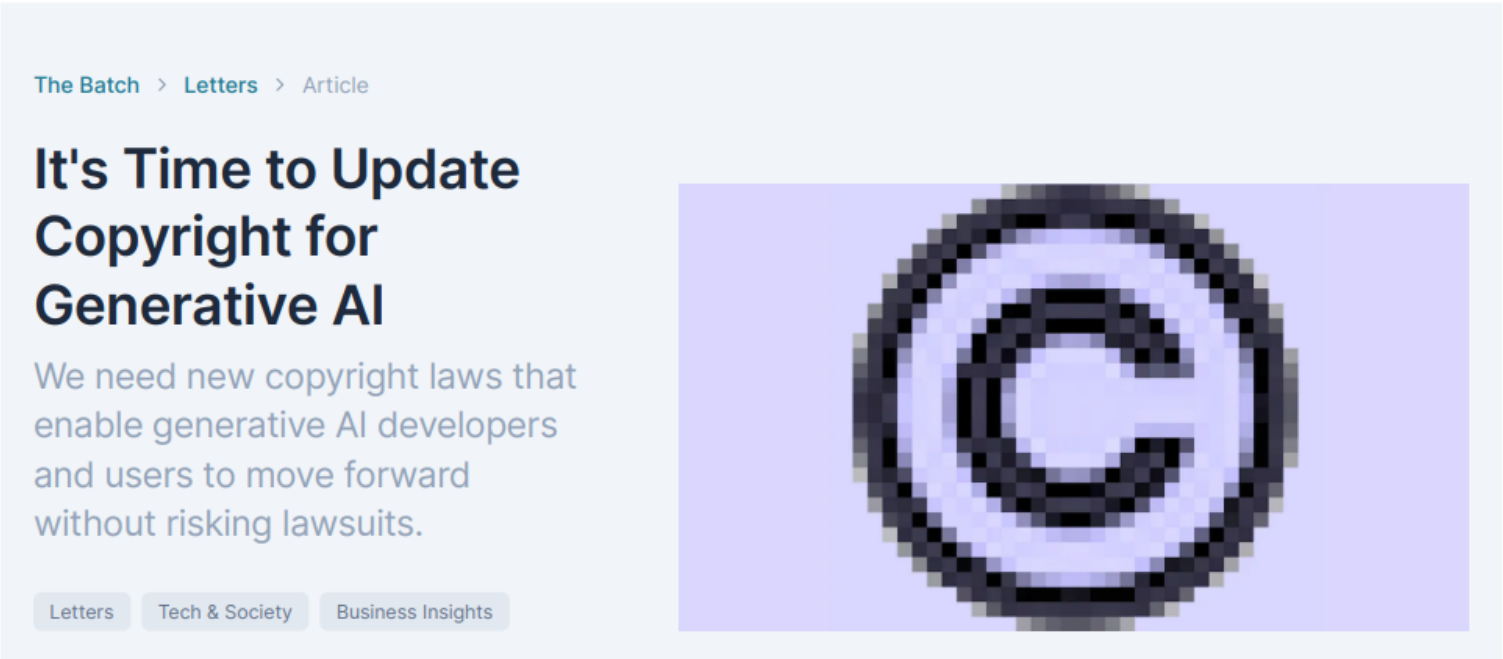
Associated Press will receive licensing fees and access to OpenAI technology. In a separate deal, OpenAI extended an earlier agreement with Shutterstock that allows it to train on the stock media licensor's images, videos, and music for six years. In return, Shutterstock will continue to offer OpenAI's text-to-image generation/editing models to its customers.

- A U.S. Senate subcommittee on intellectual property held its second hearing on AI's implications for copyright. The senators met with representatives of Adobe and Stability AI as well as an artist, a law professor, and a lawyer for Universal Music Group, which takes in roughly one-third of the global revenue for recorded music.

Behind the news: The latest court actions, which focus on generated text, follow two earlier lawsuits arising from different types of output. In January, artists Sarah Anderson, Kelly McKernan, and Karla Ortiz (who spoke in the Senate hearing) sued Stability AI, Midjourney, and the online art community DeviantArt. In November, two anonymous plaintiffs sued GitHub, Microsoft, and OpenAI saying the companies trained the Copilot code generator using routines from GitHub repositories in violation with open source licenses.

Why it matters: Copyright laws in the United States and elsewhere don't explicitly forbid use of copyrighted works to train machine learning systems. However, the technology's growing ability to produce creative works, and do so in the styles of specific artists and writers, has focused attention on such use and raised legitimate questions about whether it's fair. This much is clear: The latest advances in machine learning have depended on free access to large quantities of data, much of it scraped from the open internet. Lack of access to corpora such as Common Crawl, The Pile, and LAION-5B would put the brakes on progress or at least radically alter the economics of current research This would degrade AI's current and future benefits in areas such as art, education, drug development, and manufacturing to name a few.

We're thinking: Copyright laws are clearly out of date. We applaud legislators who are confronting this problem head-on. We hope they will craft laws that, while respecting the rights of creative people, preserve the spirit of sharing information that has enabled human intelligence and, now, digital intelligence to learn from that information for the benefit of all.



Dear friends,

Many laws will need to be updated to encourage beneficial AI innovations while mitigating potential harms. One example: Copyright law as it relates to generative AI is a mess! That many businesses are operating without a clear understanding of what is and isn't legal slows down innovation. The world needs updated laws that enable AI users and developers to move forward without risking lawsuits.

Legal challenges to generative AI are on the rise, as you can read here, and the outcomes are by no means clear. I'm seeing this uncertainty slow down the adoption of generative AI in big companies, which are more sensitive to the risk of lawsuits (as opposed to startups, whose survival is often uncertain enough that they may have much higher tolerance for the risk of a lawsuit a few years hence).

Meanwhile, regulators worldwide are focusing on how to mitigate AI harm. This is an important topic, but I hope they will put equal effort into crafting copyright rules that would enable AI to benefit more people more quickly.

Here are some questions that remain unresolved in most countries:

- Is it okay for a generative AI company to train its models on data scraped from the open internet? Access to most proprietary data online is governed by terms of service, but what rules should apply when a developer accesses data from the open internet and has not entered into an explicit agreement with the website operator?
- Having trained on freely available data, is it okay for a generative AI company to stop others from training on its system's output?
- If a generative AI company's system generates material that is similar to existing material, is it liable for copyright infringement? How can we evaluate the allowable degree of similarity?
- Research has shown that image generators sometimes copy their training data. While the vast majority of generated content appears to be novel, if a customer (say, a media company) uses a third-party generative AI service (such as a cloud provider's API) to create content, reproduces it, and the content subsequently turns out to infringe a copyright, who is responsible: the customer or the cloud provider?
- Is automatically generated material protected by copyright, and if so, who owns it? What if two users use the same generative AI model and end up creating similar content — will the one who went first own the copyright?

Here's my view:

- I believe humanity is better off with permissive sharing of information. If a person can freely access and learn from information on the internet, I'd like to see AI systems allowed to do the same, and I believe this will benefit society. (Japan permits this explicitly. Interestingly, it even permits use of information that is not available on the open internet.)
- Many generative AI companies have terms of service that prevent users from using output from their models to train other models. It seems unfair and anti-competitive to train your system on others' data and then stop others from training their models on your system's output.

- In the U.S., "fair use" is poorly defined. As a teacher who has had to figure out what I am and am not allowed to use in a class, I've long disliked the ambiguity of fair use, but generative AI makes this problem even more acute. Until now, our primary source of content has been humans, who generate content slowly, so we've tolerated laws that are so ambiguous that they often require a case-by-case analysis to determine if a use is fair. Now that we can automatically generate huge amounts of content, it's time to come up with clearer criteria for what is fair. For example, if we can algorithmically determine whether generated content overlaps by a certain threshold with content in the training data, and if this is the standard for fair use, then it would unleash companies to innovate while still meeting a societally accepted standard of fairness.

If it proves too difficult to come up with an unambiguous definition of fair use, it would be useful to have "safe harbor" laws: As long as you followed certain practices in generating media, what you did would be considered non-infringing. This would be another way to clarify things for users and generative AI companies.

The tone among regulators in many countries is to seek to slow down AI's harms. While that is important, I hope we see an equal amount of effort put into accelerating AI's benefits. Sorting out how we should change copyright law would be a good step. Beyond that, we need to craft regulations that clarify not just what's not okay to do — but also what is explicitly okay to do.

Keep learning!

Andrew Ng