NOTEWORTHY

AI IS USING COPYRIGHTED MATERIAL

5.02.2025

We know it for years: LLMs are trained by copyrighted material. But we should never forget: Aaron Swartz, a copyright activist lost his life. And so did Suchir Balaji (his parents do not believe in a suicide). And another activist Alexandra Elbakayan is being prosecuted for years.

So how can LLMs of all kind now make money of copyrighted text and images bypassing all rules? The Guardian about OpenAI

> The developer OpenAI has said it would be impossible to create tools like its groundbreaking chatbot ChatGPT without access to copyrighted material, as pressure grows on artificial intelligence firms over the content used to train their products.

The New York Times about Suchir Balaji

> But after the release of ChatGPT in late 2022, he thought harder about what the company was doing. He came to the conclusion that OpenAI's use of copyrighted data violated the law and that technologies like ChatGPT were damaging the internet. In August, he left OpenAI because he no longer wanted to contribute to technologies that he believed would bring society more harm than benefit.

Are there still copyright rules in place?

Probably. Getty Images is now suing Stable Diffusion, Facebook is using LibGen although they had to pay recently 30m penalties. Universal Music filed a lawsuit against Anthropic and NYT against OpenAI. At least a dozen of court cases are ongoing.

But I haven't heard so far of any action of a major medical publishers against any AI company (including the company who sued Elbakayan). They must have a different strategy –

instead of suing they just sell their content even behind the back of the authors. This is what Christa Dutton found out.

> One of those tech companies, Microsoft, paid Informa, the parent company of Taylor & Francis, an initial fee of $10 million to make use of its content “to help improve relevance and performance of AI systems,” according to a report released in May… Another publisher, Wiley, also recently agreed to sell academic content to a tech company for training AI models. The publisher completed a “GenAI content rights project” with an undisclosed “large tech company,” according to a quarterly earnings report released at the end of June

But can publishers just do this without asking authors? authorsalliance.org has an answer.

> In a lot of cases, yes, publishers can license AI training rights without asking authors first. Many publishing contracts include a full and broad grant of rights–sometimes even a full transfer of copyright to the publisher for them to exploit those rights and to license the rights to third parties.

We had been too naive.

Or we have been blackmailed.

14/23/25

There was never fair use … https://arstechnica.com/tech-policy/2025/03/openai-urges-trump-either-settle-ai-copyright-debate-or-lose-ai-race-to-china/ … while I now fear that this will be decided by politics not by courts.

20/3/2025

https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/ writes

Meta employees acknowledged in their internal communications that training Llama on LibGen presented a “medium-high legal risk,” and discussed a variety of “mitigations” to mask their activity.

leading to the paradoxical situation

LibGen and other such pirated libraries make information more accessible, allowing people to read original work without paying for it. Yet generative-AI companies such as Meta have gone a step further: Their goal is to absorb the work into profitable technology products that compete with the originals.