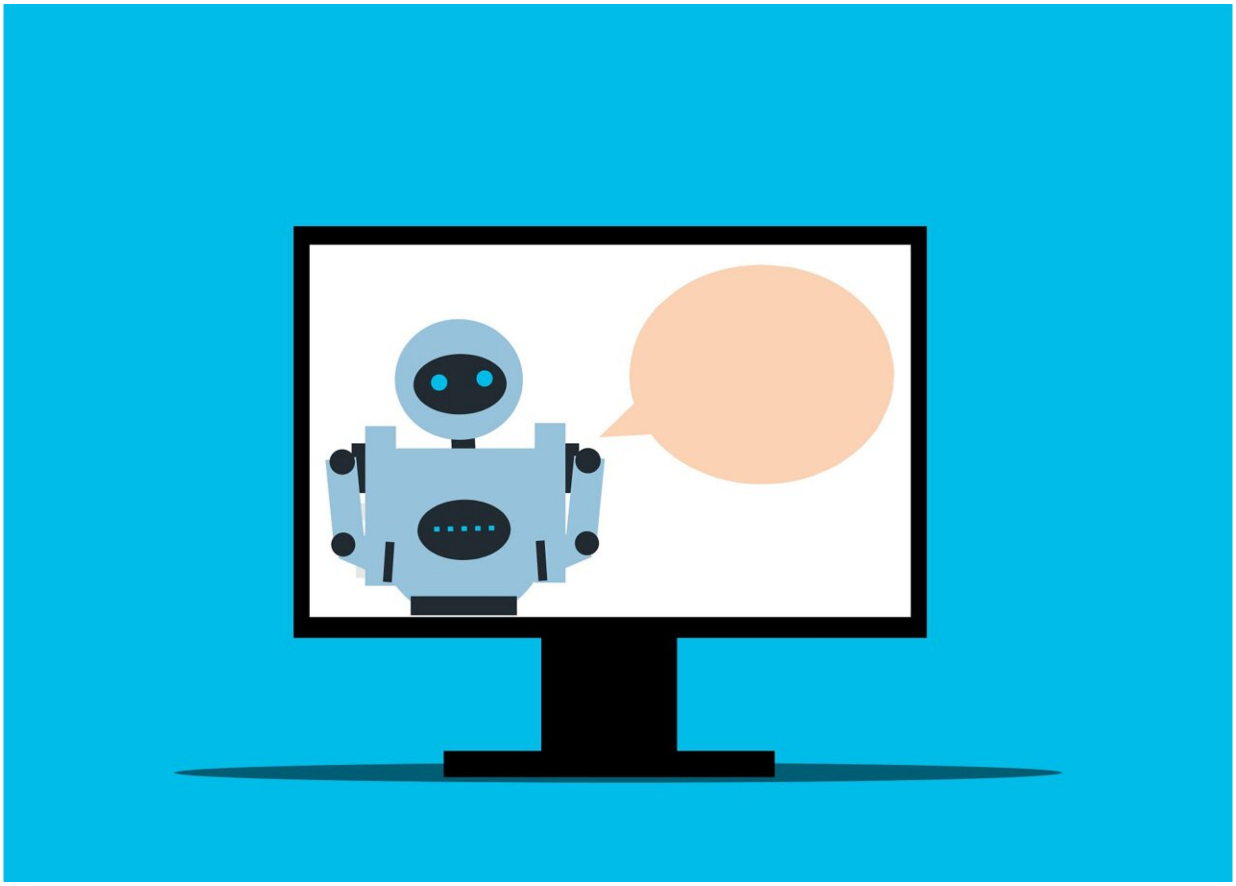
TechXplore

The New York Times' lawsuit against OpenAI could have major implications for the development of machine intelligence

January 11 2024, by Mike Cook



Credit: Pixabay/CC0 Public Domain

In 1954, the Guardian's science correspondent reported on "electronic brains", which had a form of memory that could let them retrieve information, like airline seat allocations, in a matter of seconds.

Nowadays, the idea of computers storing information is so commonplace that we don't even think about what words like "memory" really mean. Back in the 1950s, however, this language was new to most people, and the idea of an "electronic brain" was heavy with possibility.

In 2024, your microwave has more computing power than anything that was called a brain in the 1950s, but the world of artificial intelligence is posing fresh challenges for language—and lawyers. Last month, the New York Times newspaper filed a lawsuit against OpenAI and Microsoft, the owners of popular AI-based text-generation tool ChatGPT, over their alleged use of the Times' articles in the data they use to train (improve) and test their systems.

They claim that OpenAI has infringed copyright by using their journalism as part of the process of creating ChatGPT. In doing so, the lawsuit claims, they have created a competing product that threatens their business. OpenAI's response so far has been very cautious, but a key tenet outlined in a statement released by the company is that their use of online data falls under the principle known as "fair use". This is because, OpenAI argues, they transform the work into something new in the process—the text generated by ChatGPT.

At the crux of this issue is the question of data use. What data do companies like OpenAI have a right to use, and what do concepts like "transform" really mean in these contexts? Questions like this, surrounding the data we train AI systems or models like ChatGPT on, remain a fierce academic battleground. The law often lags behind the behavior of the industry.

If you've used AI to answer emails or summarize work for you, you might see ChatGPT as an end justifying the means. However, it perhaps should worry us if the only way to achieve that is by exempting specific corporate entities from laws that apply to everyone else.

Not only could that change the nature of the debate around copyright lawsuits like this one, but it has the potential to change the way societies structure their legal system.

Fundamental questions

Cases like this can throw up thorny questions about the future of legal systems, but they can also question the future of AI models themselves. The New York Times believes that ChatGPT threatens the long-term existence of the newspaper. On this point, OpenAI says in its statement that it is collaborating with news organizations to provide novel opportunities in journalism. It says the company's goals are to "support a healthy news ecosystem" and to "be a good partner".

Even if we believe that AI systems are a necessary part of the future for our society, it seems like a bad idea to destroy the sources of data that they were originally trained on. This is a concern shared by creative endeavors like the New York Times, authors like George R.R. Martin, and also the online encyclopedia Wikipedia.

Advocates of large-scale data collection—like that used to power Large Language Models (LLMs), the technology underlying AI chatbots such as ChatGPT—argue that AI systems "transform" the data they train on by "learning" from their datasets and then creating something new.

Effectively, what they mean is that researchers provide data written by people and ask these systems to guess the next words in the sentence, as they would when dealing with a real question from a user. By hiding and then revealing these answers, researchers can provide a binary "yes" or "no" answer that helps push AI systems towards accurate predictions. It's for this reason that LLMs need vast reams of written texts.

If we were to copy the articles from the New York Times' website and charge people for access, most people would agree this would be "systematic theft on a mass scale" (as the newspaper's lawsuit puts it). But improving the accuracy of an AI by using data to guide it, as shown above, is more complicated than this.

Firms like OpenAI do not store their training data and so argue that the articles from the New York Times fed into the dataset are not actually being reused. A counter-argument to this defense of AI, though, is that there is evidence that systems such as ChatGPT can "leak" verbatim excerpts from their training data. OpenAI says this is a "rare bug".

However, it suggests that these systems do store and memorize some of the data they are trained on—unintentionally—and can regurgitate it verbatim when prompted in specific ways. This would bypass any paywalls a for-profit publication may put in place to protect its intellectual property.

Language use

But what is likely to have a longer-term impact on the way we approach legislation in cases such as these is our use of language. Most AI researchers will tell you that the word "learning" is a very weighty and inaccurate word to use to describe what AI is actually doing.

The question must be asked whether the law in its current form is sufficient to protect and support people as society experiences a massive shift into the AI age. Whether something builds on an existing copyrighted piece of work in a manner different from the original is referred to as "transformative use" and is a defense used by OpenAI.

However, these laws were designed to encourage people to remix, recombine, and experiment with work already released into the outside world. The same laws were not really designed to protect multi-billion-dollar technology products that work at a speed and scale many orders of magnitude greater than any human writer could aspire to.

The problem with many of the defenses of large-scale data collection and usage is that they rely on strange uses of the English language. We say that AI "learns", that it "understands", that it can "think". However, these are analogies, not precise technical language.

Just like in 1954, when people looked at the modern equivalent of a broken calculator and called it a "brain", we're using old language to grapple with completely new concepts. No matter what we call it, systems like ChatGPT do not work like our brains, and AI systems don't play the same role in society that people play.

Just as we had to develop new words and a new common understanding of technology to make sense of computers in the 1950s, we may need to develop new language and new laws to help protect our society in the 2020s.

Provided by The Conversation

Citation: The New York Times' lawsuit against OpenAI could have major implications for the development of machine intelligence (2024, January 11) retrieved 4 May 2024 from https://techxplore.com/news/2024-01-york-lawsuit-openai-major-implications.html