‘Impossible’ to create AI tools like ChatGPT without copyrighted material, OpenAI says

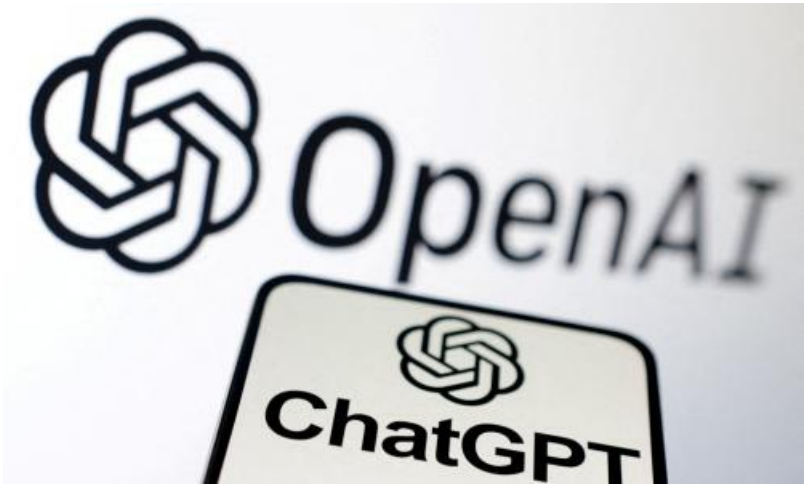
Publication Date: 2024-01-08

Author: Dan Milmo

Section: Technology

Tags: OpenAI, ChatGPT, Artificial intelligence (AI), Computing, news

Article URL: https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-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. Chatbots such as ChatGPT and image generators like Stable Diffusion are “trained” on a vast trove of data taken from the internet, with much of it covered by copyright – a legal protection against someone’s work being used without permission. Last month, the New York Times sued OpenAI and Microsoft, which is a leading investor in OpenAI and uses its tools in its products, accusing them of “unlawful use” of its work to create their products. In a submission to the House of Lords communications and digital select committee, OpenAI said it could not train large language models such as its GPT-4 model – the technology behind ChatGPT – without access to copyrighted work. “Because copyright today covers virtually every sort of human expression – including blogposts, photographs, forum posts, scraps of software code, and government documents – it would be impossible to train today’s leading AI models without using copyrighted materials,” said OpenAI in its submission, first reported by the Telegraph. It added that limiting training materials to out-of-copyright books and drawings would produce inadequate AI systems: “Limiting training data to public domain books and drawings created more than a century ago might yield an interesting experiment, but would not provide AI systems that meet the needs of today’s citizens.” Responding to the NYT lawsuit in a blog post published to its website on Monday, OpenAI said: “We support journalism, partner with news organisations, and believe the New York Times lawsuit is without merit.” Previously, the company said it respected “the rights of content creators and owners”. AI companies’ defence of using copyrighted material tends to lean on the legal doctrine of “fair use”, which allows use of content in certain circumstances without seeking the owner’s permission. In its submission, OpenAI said it believed that “legally, copyright law does not forbid training”. The NYT lawsuit has followed numerous other legal complaints against OpenAI. John Grisham, Jodi Picoult and George RR Martin were among 17 authors who sued OpenAI in September alleging “systematic theft on a mass scale”. Getty Images, which owns one of the largest photo libraries in the world, is suing the creator of Stable Diffusion, Stability AI, in the US and in England and Wales for alleged copyright breaches. In the US, a group of music publishers including Universal Music are suing Anthropic, the Amazon-backed company behind the Claude chatbot, accusing it of misusing “innumerable” copyrighted song lyrics to train its model. Elsewhere in its House of Lords submission, in response to a question about AI safety, OpenAI said it supported independent analysis of its security measures. The submission said it backed “red-teaming” of AI systems, where third-party researchers test the safety of a product by emulating the behaviour of rogue actors. OpenAI is among the companies that have agreed to work with governments on safety testing their most powerful models before and after their

deployment, after an agreement struck at a global safety summit in the UK last year.