The Guardian blocks ChatGPT owner OpenAI from trawling its content

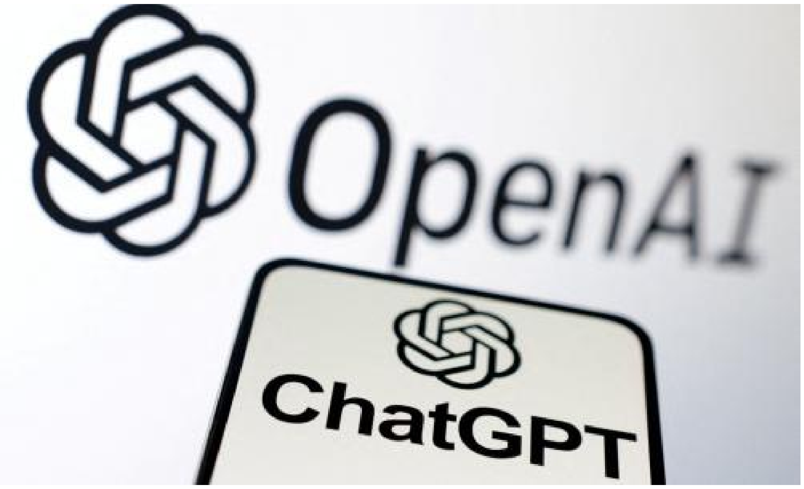
Publication Date: 2023-09-01

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Computing, The Guardian, Newspapers & magazines, ChatGPT, OpenAI, Google, news

Article URL: https://www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content



The Guardian has blocked OpenAI from using its content to power artificial intelligence products such as ChatGPT. Concerns that OpenAI is using unlicensed content to create its AI tools have led to writers bringing lawsuits against the company and creative industries calling for safeguards to protect their intellectual property. The Guardian has confirmed that it has prevented OpenAI from deploying software that harvests its content. Generative AI technology – the term for products that generate convincing text, image and audio from simple human prompts – has dazzled the public since a breakthrough version of its ChatGPT chatbot launched last year. However, fears have arisen about the potential mass-production of disinformation and the way in which such tools are built. The technology behind ChatGPT and similar tools is “trained” by being fed vast amounts of data culled from the open internet, including news articles, which enable the tools to predict the likeliest word or sentence to come after the user’s prompt. OpenAI, which does not disclose the data that helped build the model behind ChatGPT, announced in August that it will enable website operators to block its web crawler from accessing their content, although the move does not allow material to be removed from existing training datasets. A number of publishers and websites are now blocking the GPTBot crawler. A spokesperson for Guardian News & Media, publisher of the Guardian and Observer, said: “The scraping of intellectual property from the Guardian’s website for commercial purposes is, and has always been, contrary to our terms of service. The Guardian’s commercial licensing team has many mutually beneficial commercial relationships with developers around the world, and looks forward to building further such relationships in the future.” According to Originality.ai, which detects AI-generated content, news websites now blocking the GPTBot crawler, which takes data from webpages to feed into its AI models, include CNN, Reuters, the Washington Post, Bloomberg, the New York Times and its sports site the Athletic. Other sites that have blocked GPTBot include Lonely Planet, Amazon, the job listings site Indeed, the question-and-answer site Quora, and dictionary.com. This week British book publishers urged Rishi Sunak to protect the intellectual property rights of creative industries by adding it to the agenda at the November summit on AI safety being hosted in the UK. A letter from the Publishers Association, which represents publishers of digital and print books as well as research journals and educational content, asked the prime minister to make clear that intellectual property law must be respected when AI systems are being built. In July Elon Musk imposed limits on his Twitter platform, now rebranded X, to address what he claimed were “extreme levels of data scraping” by AI firms building their models. He tweeted that “almost every company doing AI” was taking “vast amounts of data” from Twitter, which Musk said was forcing the company to deploy more servers – at a cost – to cope with the demand. However, Musk has also confirmed that he will use public tweets to train models developed by his newly announced AI startup, xAI. Google's privacy policy now states that the company, which uses web crawlers to help find search results for users, may collect publicly available information to train models for Google's AI products, which include the Bard chatbot. This week Meta, the owner of Facebook and Instagram as well as a major AI developer, introduced a new policy that allows users to say they if they do not want their personal information used for training AI models. OpenAI has been contacted for comment.