Nearly One-Third of the World's Biggest News Sites Have Blocked AI Crawlers from Accessing their Content

Posted on

September 18, 2023. Jastra Kranjec.

After the ChatGPT launch in November last year, companies and consumers worldwide started using generative artificial intelligence (AI) to automate tasks, write documents, do market research, or even basic coding.

However, the rise of large language models and generative AI has also pushed into the spotlight the problem of news sites, publishers, and intellectual property holders who see their data being collected by AI crawlers. And while there are still no clear regulatory rules controlling AI's use of copyrighted material, some of the world's largest news websites have taken matters into their own hands.

According to data presented by AltIndex.com, nearly one-third of the world's top 50 news sites have blocked AI crawlers from accessing their content, and their number continues rising.

CNN, New York Times, Daily Mail, Reuters, and Bloomberg Have All Blocked At least One AI Crawler

AI companies send crawlers to collect data to train their models and provide information for chatbots. However, as data is one of their core advantages, many of the world's largest news websites have become extremely cautious, especially since there is generally no upside to handing over their data to AI crawlers.

The entire situation escalated last month after OpenAI had launched its GPTBot crawler to collect data to enhance its language models. Although the AI company promised that paywalled content would be excluded from websites, several high-profile news sites, including CNN, Reuters, and the New York Times, blocked GPTBot. Their number continued growing in the following weeks.

According to a Kirwan Digital Marketing Agency survey, 28% of the top 50 news sites worldwide have blocked at least one AI crawler by the end of last month. In regional comparison, the picture is a bit different. For example, 24%, or twelve out of fifty largest news sites in the United States, have blocked at least one AI crawler, far more than in the United Kingdom, where only three of 21 leading sites did the same. In India, the percentage of top new sites unwilling to hand over their data to AI companies is much higher, with one-third blocking at least one AI crawler.

One in Five Top News Sites has Blocked GPTBot

Although most of the world's 50 leading news sites still haven't taken action on blocking, the study showed GPTBot is the number one choice among those who have. Statistics show the brainchild of OpenAI has been blocked 22% of the time across the top 50 news sites, with Bloomberg, Reuters, Business Insider, Washington Post, the New York Times, and CNN as the top names on this list.

CCBot has been blocked about half as often as the GPTBot, with a 10% share across the top 50 news sites. The survey also showed ChatGPT had been blocked by only one website, that of the Washington Post, the same as AnthropicAI, which has been blocked by only the UK's NewsNow.

Overall, the New York Times, Washington Post, Reuters, and NewsNow lead in blocking AI crawlers from accessing their content, with each news site blocking two AI bots.

Jastra is an editor, writer, and PR specialist with years of experience in news, research, and report writing. Over the years, she has covered different topics and markets, including social media, digital content, the creator economy and the entertainment industry.