FACTSHEET
February 2024

DOI: 10.60625/risj-xm9g-ws87

How Many News Websites Block AI Crawlers?

Richard Fletcher

Key findings

In this factsheet we describe the proportion of top news websites in ten countries that block AI (artificial intelligence) crawlers. Companies such as OpenAI (who make ChatGPT) and Google (who make Gemini, formerly called Bard) use crawlers to scrape data from websites to train large language models (LLMs). They can also be used to retrieve information from websites in real time in response to user requests. But if websites choose to block crawlers (and no other arrangement is made), none of this is permitted, with consequences for the quality of the underlying models and their ability to retrieve information from the web.

We find:

- By the end of 2023, 48% of the most widely used news websites across ten countries were blocking OpenAI's crawlers. A smaller number, 24%, were blocking Google's AI crawler.
- Almost every website (97%) that decided to block Google's AI crawler was also blocking OpenAI's crawlers.
- The proportion of news websites that blocked OpenAI varied considerably by country, ranging from 79% in the USA to just 20% in Mexico and Poland. For Google, the figures ranged from 60% in Germany to 7% in Poland and Spain.
- During 2023, none of the websites we examined had reversed their decision after deciding to block.
- News outlets with a relatively large online news reach were slightly more likely to be blocking AI crawlers than those with a relatively small reach.
- All types of news outlet were blocking, but the websites of legacy print publications were more likely to be blocking than those of either broadcasters or digital-born outlets.
- Comparing our findings to other work suggests that news publishers are more likely to block compared to popular websites more generally.

Background

Web crawlers (sometimes referred to as 'spiders' or 'bots') automatically browse the web, systematically collecting data as they go. They can be used for a variety of purposes. Search engines, for example, rely on data collected by their web crawlers to index pages on the web, so they can respond quickly to search queries.

AI companies like OpenAI can also use crawlers to collect data from the web to train their models. LLMs need to be trained on huge volumes of data to work well, and the web is an important source of high-quality text and audio-visual data. For example, when a team of *Washington Post* journalists and researchers at the Allen Institute for AI analysed Google's C4 data set, which has been used to instruct some LLMs (Schaul et al. 2023), the broad category of 'News and Media' – which included news publishers, but also, for example, Wikipedia, Scribd, Goodreads, and more – accounted for 13% of tokens. Once trained, LLMs like GPT can produce outputs and respond to questions from people via interfaces such as ChatGPT. While models do not need to be 'connected' to the internet to do this, once trained they can also be linked up to the web, allowing them to retrieve information from websites in real time that can then be used as prompts as part of the outputs. In this way, LLMs can be used as an alternative to other kinds of search.

However, for a variety of possible reasons, news publishers may not want their content to be used by AI companies. For example, some – like the *New York Times* (Grynbaum and Mac 2023) – think they should be financially compensated for the use of their content to train AI models. Or, if people use AI to get the latest news from the web, brands may worry about incorrect outputs (or 'hallucinations') being attributed to them, or that users won't be linked back to the publishers for them to monetise. Other publishers may not be worried by any of these potential risks, or they may actively *want* to be included because they want their journalism to feature when people use generative AI for news-related purposes (Maher 2024). And a few – like Axel Springer – have already struck deals with companies such as OpenAI, permitting them to respond to user queries with news from their websites (Sisani and Sommerfeld 2023). In parallel, news publishers all over the world are currently experimenting with AI tools to see whether they can create new user experiences, improve efficiency, and cut costs.

If publishers do not want AI companies to be able to access their online content, they have the option of blocking their web crawlers. Publishers can use the robots.txt file on their website to instruct web crawlers to stay away (though compliance with the instruction is voluntary). When OpenAI released its latest web crawlers on 7 August 2023, they also provided instructions on how to block them (OpenAI, n.d.), giving publishers the ability to opt out, as did Google on 28 September 2023 (Romain 2023).

Tracking which publishers are blocking therefore tells us about the relationship between publishers and AI companies at a time when many observers believe that AI will transform the information landscape. It also helps us understand how well future models are likely to perform when it comes to news, and thus how useful AI will be to the public as a way of getting news.

To track how many websites are blocking the most prominent AI crawlers we examined their robots.txt file over time. We did this by automatically examining the archived robots.txt files from the Internet Archive's Wayback Machine for every available day in 2023 for the 15 most widely used online news sources according to the 2023 Reuters Institute *Digital News Report* (Newman et al. 2023) in ten countries: Brazil, Denmark, Germany, India, Mexico, Norway, Poland, Spain, the UK, and the US.[1]

Previous Research

We are not the first to do this. Originality.ai, a company that develops AI and plagiarism checkers, currently tracks the proportion of the world's 1,000 most popular websites that block AI crawlers using the Wayback Machine (Originality.ai, n.d.). At the time of writing, they find that around one-third block OpenAI, around one-fifth block Common Crawl and one in ten block Google. However, the tracker does not focus on news publishers specifically, and includes categories of website (e.g. e-commerce) that have little reason to block. Journalist Ben Welsh

[1] Some of the lists of 15 most widely used online news sources published in the 2023 *Digital News Report* sometimes include generic options such as 'Local or regional newspaper online'. These were excluded and replaced with the online source with the next highest reach. Some lists include international or non-domestic sources (e.g. the BBC in the US) and aggregators (e.g. Yahoo! News and MSN) as these are among the 15 most widely used online news sources in several countries. For India, the list is of the 15 most widely used online news sources among the English-speaking population.

automatically examines the robots.txt files of 1,156 news publishers every day, checking to see whether they are blocking OpenAI, Google AI, and Common Crawl, sharing the results on his website (Welsh, n.d.). At the time of writing, the results show that around 50% of news websites tracked block OpenAI, and around 40% block Google AI and Common Crawl. This suggests that news websites are more likely to block compared to popular websites more generally. However, it is important to note that around 75% of sites checked are from the US, and it is unclear if this skews the results. Furthermore, it is not clear whether there are any systematic differences in blocking between different types of news publisher.

Results

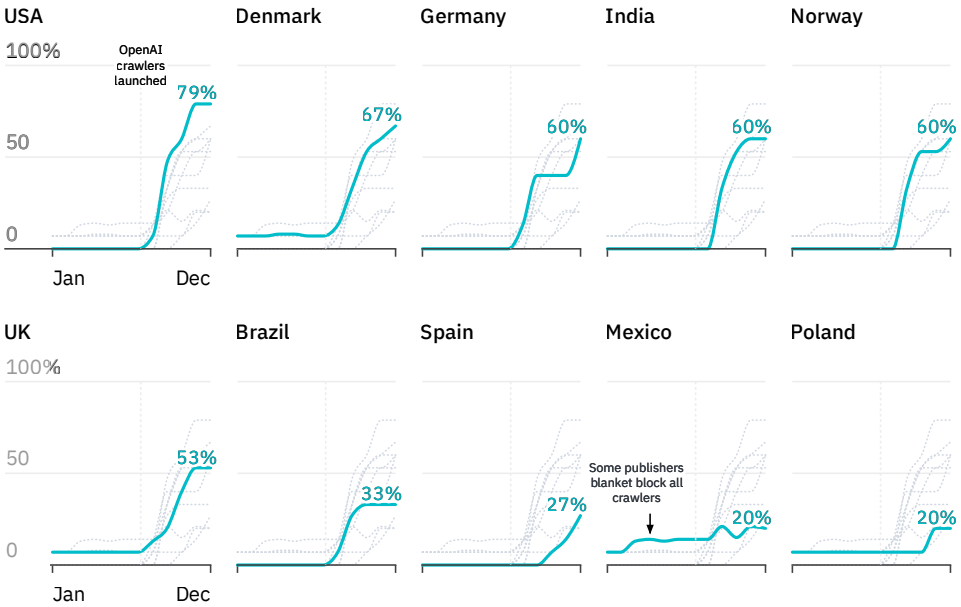
Differences by country

Examining the 15 most widely used online news sources in ten countries, we find that by the end of 2023, 48% of top news websites across ten countries were blocking OpenAI's crawlers. Around half as many (24%) were blocking Google's AI crawler.[2]

The headline figures mask very large differences by country. The proportion of top online news websites blocking OpenAI ranged from 79% in the US, to just 20% in Mexico and Poland (Figure 1). For Google, the proportion blocking their AI crawler ranged from 60% in Germany to 7% in Poland and Spain (Figure 2). In general, outlets in the Global North were more likely to be blocking than those in the Global South. (Interestingly, the figures are aligned with attempts to index countries in terms of AI capabilities and preparedness, such as those published by Tortoise (n.d.) and Oxford Insights (n.d.), both of which rank the US first.)

In every country apart from Germany, where the figure was 60% for both, more top news websites blocked OpenAI's crawlers than Google's. Moreover, almost every website that blocked Google AI also blocked OpenAI (97%). This could be because ChatGPT is more prominent and widely used than Bard/Gemini, or it could be because the OpenAI crawler was released first. But it is also possible that publishers are more cautious about blocking Google in case it affects their prominence in search results – even though there are separate crawlers for search and AI.

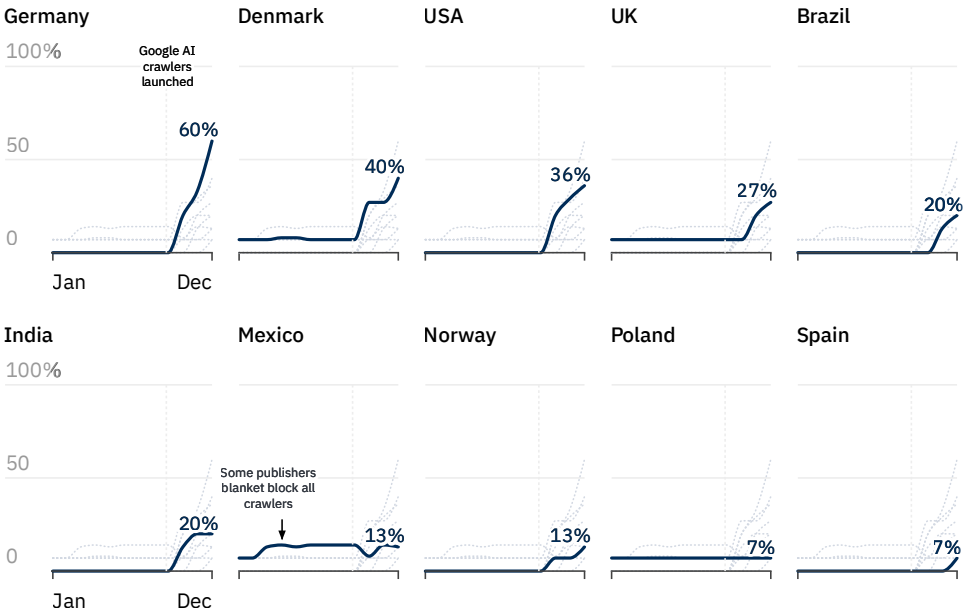
Figure 1. Proportion of top news websites blocking OpenAI's crawlers during 2023



Note: Based on analysis of the robots.txt files from the 15 most widely used online news websites in each country (according to the 2023 Reuters Institute *Digital News Report*) sourced from the Wayback Machine. December 2023 data available for all websites except the *Washington Post* in the USA.

[2] A website is counted as blocking OpenAI's crawlers if it blocks either 'GPTBot', 'ChatGPT-User' or all crawlers using robots.txt. A website is counted as blocking Google AI if it blocks 'Google-Extended' or all crawlers. December 2023 data available for all websites except the *Washington Post* in the USA.

Figure 2. Proportion of top news websites blocking Google's AI crawler during 2023



Note: Based on analysis of the robots.txt files from the 15 most widely used online news websites in each country (according to the 2023 Reuters Institute *Digital News Report*) sourced from the Wayback Machine. December 2023 data available for all websites except the *Washington Post* in the USA.

Differences over time

Looking at the data over time, we can see that in most countries at least some publishers started blocking OpenAI's crawlers as soon as they were released – but in Spain, Mexico, and Poland publishers acted later (Figure 1).[3] We see a similar picture for blocking Google AI, but in Mexico and Poland there's no evidence that any of the top news websites responded to the launch of the Google crawler by blocking it (Figure 2). Some publishers in these countries were de facto blocking the Google AI crawler, but only because they have a longstanding policy of blocking all web crawlers, and hence the proportion blocking being more than 0% before the crawler was even launched.

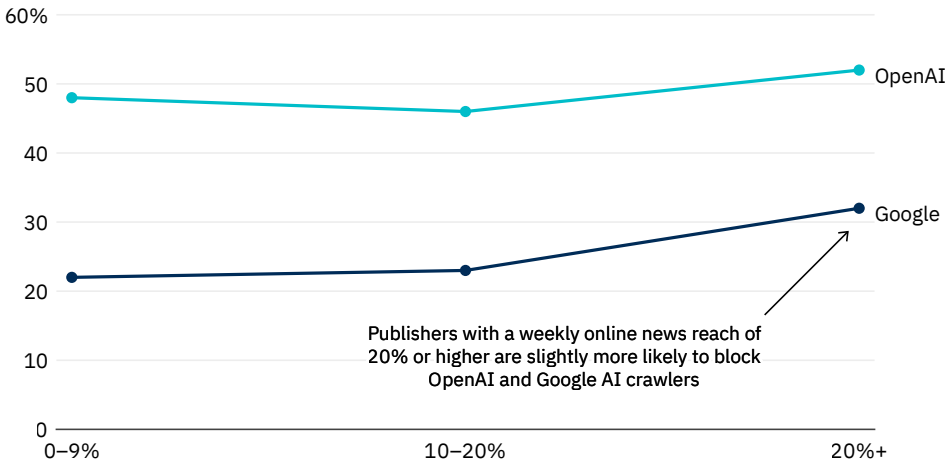
During 2023, no website unblocked either an OpenAI or Google AI crawler once the decision had been made to block. The dips in the trendline in Mexico in Figures 1 and 2 are due to missing data from the Wayback Machine rather than websites unblocking. However, if more publishers strike deals with AI companies, or if the downsides of blocking start to outweigh the upsides, we could see a reversal in this trend in the future.

Differences by publisher

If we focus on differences by publisher, we see that some categories of publisher were more likely to block AI crawlers than others. First, outlets with a relatively large online reach, according to the RISJ's 2023 *Digital News Report* (Newman et al. 2023), were slightly more likely to be blocking than outlets with a relatively small reach (Figure 3). Among outlets with an online news reach of 20% or higher, 32% were blocking the Google AI crawler by the end of 2023. However, the figure was just 22% for those with a reach of less than 10%. The differences for blocking OpenAI are much smaller.

[3] Data over time is analysed and presented on a monthly basis to account for missing or erroneous data from the Wayback Machine. Because some archived robots.txt files are missing or old, we collect all data from each day in a given month and use the most commonly occurring file to represent that month. However, if data is missing for an entire month, then this can lead to dips in trendlines, as in Mexico.

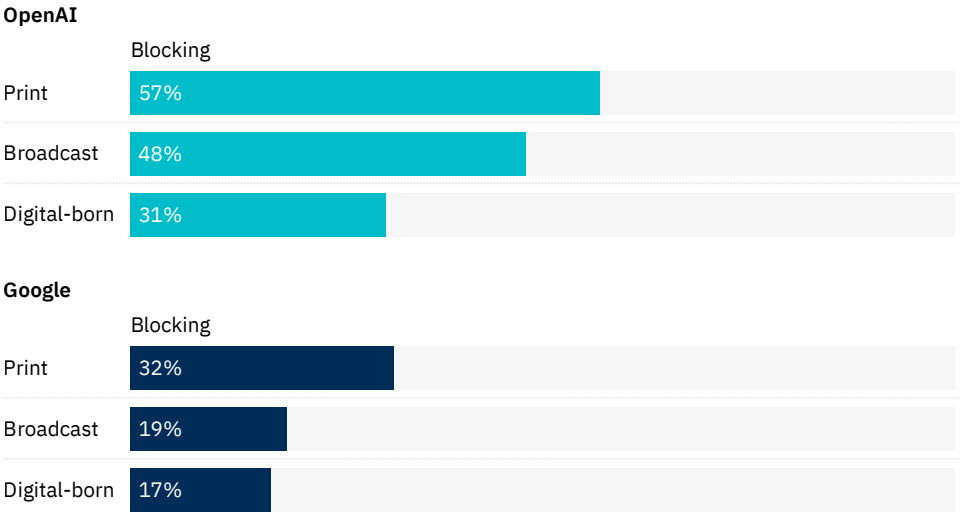
Figure 3. Proportion of top news websites blocking AI crawlers by the end of 2023
By weekly online news reach



Note: Based on analysis of the robots.txt files from the 15 most widely used online news websites in each country (according to the 2023 Reuters Institute *Digital News Report*) sourced from the Wayback Machine. Weekly online news each figures taken from the 2023 Reuters Institute *Digital News Report*. December 2023 data available for all websites except the *Washington Post* in the USA.

We see larger differences by outlet type. We group outlets into three categories: legacy print publications (e.g. newspapers like the *New York Times* and magazines like *Der Spiegel*), television and radio broadcasters (e.g. the BBC and CNN), and digital-born outlets (including HuffPost and Yahoo!). Over half (57%) of the websites of legacy print publications were blocking OpenAI's crawlers by the end of 2023, compared to 48% of television and radio broadcasters, and around one-third (31%) of digital-born outlets (Figure 4). The pattern for Google was similar, with print outlets more likely to be blocking (32%) than broadcasters (19%) and digital-born (17%) outlets.

Figure 4. Proportion of top news websites blocking AI crawlers by the end of 2023
By outlet type



Note: Based on analysis of the robots.txt files from the 15 most widely used online news websites in each country (according to the 2023 Reuters Institute *Digital News Report*) sourced from the Wayback Machine. December 2023 data available for all websites except the *Washington Post* in the USA.

Conclusion

In this factsheet we have shown that by the end of 2023 around half of the most widely used news websites across ten countries were blocking OpenAI and Google's AI crawlers. Furthermore, those blocking were disproportionately legacy print outlets and outlets with a larger reach. This means that newer models are less likely to be trained on news output from newspaper and magazine publishers, and those outlets that are more widely used by the public. This could have consequences for both the quality and relevance of AI outputs when it comes to news, both from the models themselves and in terms of what they are able to retrieve from the web.

However, it is important to keep in mind that this is just a snapshot of the situation at the end of 2023. This is a fast-moving area, and the situation is likely to change, even in the short term, especially as some publishers look to strike deals with AI companies and new products are being developed all the time.

References

Grynbaum, M. M., Mac, R. 2023. 'The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work'. *New York Times*, 27 December. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Maher, B. 2024. 'Politico embraces generative AI web crawlers with website redesigns'. *Press Gazette*, 1 February. https://pressgazette.co.uk/platforms/politico-relaunches-european-sites-first-party-data-ai-crawler-readability/

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., Nielsen, R. K. 2023. *Reuters Institute Digital News Report 2023*. Oxford: Reuters Institute for the Study of Journalism.

OpenAI. (n.d.). *GPTBot*. https://platform.openai.com/docs/gptbot (Accessed 8 February 2024).

Originality.ai. (n.d.). *Websites That Have Blocked OpenAI's GPTBot CCBot Anthropic Google Extended – 1000 Website Study*. https://originality.ai/ai-bot-blocking (Accessed 8 February 2024).

Oxford Insights. (n.d.). *Government AI Readiness Index 2023*. https://oxfordinsights.com/ai-readiness/ai-readiness-index/ (Accessed 8 February 2024).

Romain, D. 2023. 'An update on web publisher controls'. *The Keyword*, 28 September. https://blog.google/technology/ai/an-update-on-web-publisher-controls/

Schaul, K., Chen, S. Y., Tiku, N. 2023. 'Inside the secret list of websites that make AI like ChatGPT sound smart'. *Washington Post*, 19 April. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Sisani, A., Sommerfeld, J. 2023. 'Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism'. *Axel Springer*, 13 December 13. https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism

Tortoise. (n.d.). *The Global AI Index*. https://www.tortoisemedia.com/intelligence/global-ai/ (Accessed 8 February 2024).

Welsh, B. (n.d.). *Who blocks OpenAI, Google AI and Common Crawl?* https://palewi.re/docs/news-homepages/openai-gptbot-robotstxt.html (Accessed 8 February 2024).

Acknowledgements

The author would like to thank Marina Adami and Rasmus Kleis Nielsen for their feedback and input.

About the Author

Richard Fletcher is Director of Research at the Reuters Institute for the Study of Journalism.