Some suggestions to improve robots.txt

Chris Needham chris.needham@bbc.co.uk

Piers O'Hanlon piers.ohanlon@bbc.co.uk

August 2, 2024

The BBC set out its approach to generative AI in [1]: "The emergence of generative AI is expected to herald a new wave of technology innovation that could impact almost every field of human activity. The new tools can generate text, images, speech, music and video in response to prompts from a user, producing new creative possibilities, and potential efficiency gains. Alongside these opportunities, it is clear that generative AI introduces new and significant risks if not harnessed properly. These include ethical issues, legal and copyright challenges, and significant risks around misinformation and bias. The BBC does not believe the current scraping of its content and data without permission in order to train generative AI models is in the public interest, and wants to agree a more structured and sustainable approach with technology companies."

While we recognise that generative AI may bring numerous benefits to the industry, we also recognise that the output of generative AI systems can produce so-called "hallucinations", which may include incorrect or misleading claims, misattribution of sources, and biased points of view is of particular concern to publishers of news and factual content. The problems are not limited to text outputs from large language models. AI-generated images, audio, and video, including deep fake content can mislead while playing on the existing trust relationships people have with publishers [2]. The end result could lead to a further undermining of trust in factual reporting and journalistic output.

In addition, chatbot interfaces to generative AI systems may lead to reduced traffic to publishers, impacting revenues and readership, while at the same time increasing internet centralisation and reducing the incentive to publish on the open web, which could result in an increase in walled gardens.

Many generative AI systems have been developed through large scale scraping of content from the web. In the UK, copyright law allows for text and data mining for non-commercial research [3], and this is currently being reviewed by regulators [4]. In the US, the legal basis for scraping of content is being challenged

in the courts [5]. While the legality of such practices is being debated in some jurisdictions, there are clearer rules in place in the European Union. Article 4 of the Directive on Copyright in the Digital Single Market (EU Directive 2019/790) provides for "Exception or limitation for text and data mining" (TDM), which may be "reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online".

The BBC creates and publishes news, entertainment, and educational content and its terms of service explicitly state that permission is required to use content, media, and metadata from our websites [6].

Robots.txt is the long established mechanism that websites can use to control crawler behaviour, now on the Standards Track at IETF as RFC 9303. This approach has a number of issues which limit its effectiveness in the context of generative AI crawlers:

1. The growth in the number of AI crawlers means that web publishers must actively monitor and block each new one that emerges. Many AI companies publish details on how to block their crawlers using robots.txt, for example, OpenAI [7]. Indeed, the BBC has taken steps to prevent web crawlers such as those from OpenAI and Common Crawl from accessing BBC websites [1].

2. The user-agent string is self-declared and easily spoofed, which means that websites do not have a reliable mechanism for identifying crawlers. Also, the details for how to block each crawler may also change over time, which emphasises the need for active monitoring [8].

3. Robots.txt does not adequately provide a means for signalling of allowed and disallowed usage in a way that is compatible with regulations, such as those in the EU.

4. It has been been observed [9] that not all crawlers respect robots.txt directives, and yet still appear to be accessing content despite having published details on how to block them using robots.txt.

The level of concern regarding the use of published web content for generative AI has led to many different solutions being proposed and developed. This includes ai.txt [10], TDM Reservation Protocol [11], and trust.txt [12], as well as those that embed usage policy metadata in media assets, such as IPTC [13], C2PA [14], or "noai" and "noimageai" HTML meta tags [15]. This points to the need for standardisation, to reduce fragmentation and increase adoption.

In our view, a sustainable solution would allow publishers to indicate how content may be used by purpose, for example, to allow indexing by search engines,

but not for training AI models. This would then apply to all crawlers, and websites could have the option to selectively apply different policies on a per-crawler basis. AI technology is developing and changing rapidly, so any classification of purpose needs to be precise enough to allow or prevent specific uses, but also broad enough to be future proof. Example uses include AI model training, evaluation, verification and validation, AI inference, or retrieval-augmented generation (RAG).

We also need to ensure that such systems can provide for a range of content types including audio, video, text, and data. Streaming media may need special consideration, as these typically use higher level protocols such as DASH or HLS and distribution at scale often uses a multi-CDN approach. Solutions should also not limit themselves to content accessed via HTTP. Other IP protocols such as WebSockets, WebTransport, WebRTC, MoQ, RTP, etc., should be considered.

Finally, to be effective, technical solutions need to be supported by existing and emerging legal frameworks, which also points to the need for alignment at a regulatory level. A more fine-grained, nuanced solution, e.g., through classification of purpose, could help regulatory policy development, to promote innovation while balancing the needs of publishers.

We welcome this IAB initiative to bring together the internet community and look forward to continued discussion to develop potential solutions.

References

[1] https://www.bbc.co.uk/mediacentre/articles/2023/generative-ai-at-the-bbc

[2] https://www.bbc.co.uk/news/technology-66993651

[3] https://www.gov.uk/guidance/exceptions-to-copyright

[4] https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-series-on-generative-ai-and-data-protection/

[5] https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

[6] https://www.bbc.co.uk/usingthebbc/terms-of-use/

[7] https://platform.openai.com/docs/bots

[8] https://www.404media.co/websites-are-blocking-the-wrong-ai-scrapers-because-ai-companies-keep-making-new-ones/

[9] https://www.reuters.com/technology/artificial-intelligence/multiple-ai-companies-bypassing-web-standard-scrape-publisher-sites-licensing-2024-06-21/

[10] https://spawning.substack.com/p/aitxt-a-new-way-for-websites-to-set

[11] https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

[12] https://journallist.net/new-tool-to-deal-with-ai

[13] https://iptc.org/news/meta-announces-support-for-iptc-metadata-in-generative-ai-images/

[14] https://c2pa.org/specifications/specifications/1.4/ai-ml/ai_ml.html

[15] https://www.deviantart.com/team/journal/UPDATE-All-Deviations-Are-Opted-Out-of-AI-Datasets-934500371