Consultation Response

I broadly agree with the CR as outlined in this consultation, but I have two concerns that I do not feel the CR and the Interpretative Notes address.

Opt-Out

On page 17, in paragraph 3.3, Effective controls 1(a)(i) you write:

> make any changes to necessary to the existing Google-Extended control in order for it to enable publishers to opt their Search Content out of the training of generative AI models

This is one of the first times were "opt…out" is mentioned.

What it does not address is what is to happen to content that has been crawled before the new controls are in place. Should an opt-out of published content also mean that if it was previously crawled, be removed from the existing data sets?

I think it ought to be, otherwise the proposed controls can be easily pre-empted by gathering the data set now, while these proposed controls are not in place yet.

In addition, I would expect that publishers, when they notice that their published content is used inaccurately, or against their wishes, they still would be able to make sure their content is no longer used for "the training of generative AI models and grounding of broader generative AI services".

Third Party Datasets

On page 18, in Effective controls 5, you write:

> The CMA expects Google not to, for example, pay a third party to crawl the website of a published that has opted out of its Search Content being used by Google through these controls. However, the CMA considers that it would be reasonable for Google to acquire such content through **open-source datasets**, where these datasets have obtained content legally, given the nature of such sources.

And on page 29, paragraphs 4.26(b)(ii) and 4.27, you write:

> 4.26(b)(ii)
>
> circumventing publisher choices by acquiring opted-out content through other means. For example, in principle Google could pay a third party to crawl an opted-out website.
>
> 4.27
>
> We have therefore also included a high-level requirement addressing these risks at paragraph 3 in the Publisher CR.

The *open-source* datasets that you could be referring too, even though they have been obtained legally, often do lack the attribution and source information that is required to be transparent to users of general search, in such way that attribution can be provided. The original source would **not** be the open-source dataset itself, but where they obtained *their* data.

The Publisher CR and Interpretative Notes make no such reference to this. And therefore this still seems as a quite possible circumvention of at least some of the provisions as outlined in the Publisher CR.