Google

Chi Onwurah MP for Newcastle upon Tyne Central and West
Chair, Science, Innovation and Technology Committee
House of Commons,
London, SW1A 0AA

2 April 2025

Dear Chair,

I would like to thank you for your letter of Wednesday 12 March 2025, requesting additional detail on Google's response to the UK copyright consultation. We welcome this opportunity to share more with the Committee now that the consultation deadline has passed.

Our response outlines how a reformed and balanced copyright framework will help facilitate the UK's ambitions to become a leader for AI investment and innovation, while also delivering for the UK creative sectors.

The UK is rightly positioning itself to be at the forefront of AI technologies to reap the societal and economic benefits of these developments, estimated at up to £400bn by 2030[1]. We welcome the UK's ambition to seize this opportunity and become a world leader on AI, and the recent AI Opportunities Action Plan is a positive statement of intent. But for AI companies of all sizes to flourish, an enabling copyright framework will be crucial.

We are clear in our view that the training of AI models is a non-expressive use of open web content, which boosts economic growth, fosters scientific advancement, and enables the creation of valuable new content. AI platforms and services will supplement and support creativity, not replace it.

Foundational language models learn by training on diverse information and data from the open web - learning in much the same way as people do.

Training a model requires billions of data points to provide information, culturally-relevant content, varied perspectives and different forms, and uses of language to help guard against bias and take into account societal norms. However, generative AI models are not databases or information retrieval systems - there are no copies of the training content in the model itself. The model is simply generating a statistical estimation of a satisfactory response.

As such, there are no specific content pathways in the model and the use of content is non-expressive. The problem is that the legal status of non-expressive uses is not articulated clearly in current UK copyright law. Such legal certainty is important given the scale of investment required for AI development.

To ensure the UK delivers on its ambition to be an 'AI maker', we believe the Government should deliver an internationally competitive copyright regime that:

- Maintains the current non-commercial research exception
- Enables TDM for any purpose with rights reservation for rightsholders
- Provides the UK with a competitive advantage by introducing a commercial research exception

While Option 2 in the Government's consultation is the most competitive option and would support the Government's ambitions to be an AI leader, we believe rightsholders should have choice and control, and

[1] Google UK Economic Impact report

acknowledge that Option 3 would deliver on this aim. As we lay out in detail in our response, appropriate technical solutions exist (e.g. in the form of an existing and effective opt out mechanism) and are working to facilitate the detail of that proposal. However, with respect to transparency requirements referenced in Option 3, getting the detail right will be critical to ensure that any such provisions do not hinder AI development and impact the UK's competitiveness in this space.

Therefore, our preferred approach is an amended Option 3, which allows TDM for any purpose with rights reservation for rightsholders, *while* introducing a commercial research exception.

There is a clear need for a balanced approach that encourages AI innovation while respecting the rights of content creators. We hope for reforms that provide legal clarity, support technological advancement, and foster collaboration between AI developers and rightsholders, to the benefit of the UK.

You asked whether our response contained suggestions for "simple technical means for creators to exercise their rights, either individually or collectively".

The consultation itself stated that over half of news publishers block the main generative AI web-crawlers using the robots.txt standard.[2] In our response, we outlined that these mechanisms are well-understood and used on websites of all kinds and sizes. A study from Reuters Institute at Oxford University supports that nearly a quarter of news publishers are already blocking Google's AI crawler, and half of sites are already blocking OpenAI's crawler[3]. Nor is the use of robots.txt limited to use by news publishers. Companies like YouTube use robots.txt to control how hosted content is accessed by web crawlers.

We support web publishers in exercising their rights under UK law to prohibit or authorise use of their content for training AI models through the tools available. In September 2023, Google launched Google-Extended, a new control that web publishers can use to manage whether their sites help improve Gemini and Vertex AI generative APIs, including future generations of models that power those products. Making simple, scalable and well-recognised controls, like Google-Extended, available through robots.txt is an important step in providing transparency and control.

Any web publisher can implement Google-Extended to opt out of model training. It uses the well-established robots.txt technology with which publishers are already familiar, and it does not impact a site's inclusion or ranking in Google Search. Publishers such as The Guardian have implemented Google-Extended and are still clearly available on Search. It is easy for anyone to check that Google Extended has been applied to a web publisher site by adding /robots.txt to the web address.

Overwhelmingly, the feedback we have received from site owners is that they understand robots.txt and its mechanisms, and are able to use it appropriately to achieve their goals and exercise their rights. Independent studies such as the open-source Web Almanac for 2024 have shown that the vast majority of the web uses these mechanisms, and they also use them specifically for the newer AI crawlers. These mechanisms are well-understood, and used on websites of all kinds and sizes.

Your letter also asked whether our response contained "proposals to utilise technology to help deliver an entirely transparent rights environment".

As outlined above, we believe an opt-out is an effective and workable mechanism to deliver choice and transparency for rightsholders. We would also note that certain types of transparency measures can be unworkable for AI firms of all sizes, including start-ups.

---

[2] *"82: There is a growing use of rights reservation protocols and standards by both AI developers and rightsholders. Over half of news publishers block the main generative AI web-crawlers using the robots.txt standard. AI developers generally respect this standard and offer various implementations of it. For example, Google, Microsoft Bing and OpenAI each adopt their own approach."* Copyright and AI: Consultation

[3] University of Oxford Reuters Institute, 'How many news websites block AI crawlers?', February 2024.

The absence of a requirement under international and UK copyright law to register works or transfers of rights, the inconsistent quality of rights ownership metadata, and a generally low threshold for copyright protection, make it impossible for AI developers to identify and document the copyright status of third party works used for training and make such information available to rightsholders.

Moreover, there are major risks associated with the UK requiring the disclosure of training data:

- It could equip bad actors with the knowledge to attack the model, causing security concerns.
- The selection and preparation of training data includes valuable information protected by trade secrets. Certain transparency requirements raise the risk of the disclosure of sensitive information and thus jeopardises investment, including in and by start-ups, in the development and deployment of AI models here in the UK.

Under current legal frameworks, disclosure is already ensured in specific enforcement proceedings. This disclosure is purposeful, meaning it concerns a specific subject matter and a specific rights holder, and is subject to the normal checks and balances of the judicial system, including proportionality, protection of trade secrets and confidentiality of proceedings. This approach is welcome as it provides safeguards for all parties involved, helping to build trust. We would urge caution against unnecessary divergence from current legal frameworks due to the risks and practical hurdles outlined.

Finally, there are practical hurdles associated with any requirement to publish detailed transparency reports. This is because this reporting is effectively asking developers to disclose and summarise all the content on a web that is constantly changing and dynamic. This makes it impossible to produce and maintain an up-to-date summary.

Next steps

We have attached our full consultation response to this letter for the Committee's reference. We will be making this available to the public in due course. We are also available to join future evidence sessions the Committee may hold on this topic.

We look forward to continuing this conversation with you in the coming months.

Best wishes,
Didi Denham

Government Affairs and Public Policy, Google UK