March 1, 2024

Information Commissioner's Office
generative.ai@ico.org.uk

Written Evidence Submission on the Lawful Basis for Web Scraping to Train Generative AI Models

On behalf of the Center for Data Innovation, we are pleased to submit this response to the Information Commissioner's Office (ICO) call for evidence in respect to the first chapter of its generative AI and data protection consultation series, focussing on the lawful basis for web scraping to train generative AI models.

The Center for Data Innovation studies the intersection of data, technology, and public policy. With staff in Washington, London, Ottawa, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximise the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

EXECUTIVE SUMMARY

In this submission, we make the following points:

1. The ICO should re-evaluate the article 6(1) legal bases for lawful processing under the UK General Data Protection Regulation to include article 6(1)(e) as a lawful basis for web scraping data in public sector AI;
2. The ICO should consider the broader purpose of building an AI model as opposed to the purpose the model is used for in the application of legitimate interests;
3. We support the ICO's necessity assessment that web scraping is the least intrusive way to train generative AI models;
4. The ICO should assume a reasonable expectation of processing by data subjects for data wilfully made available to the public via the Internet; and
5. Downstream uses of a trained model do not engage in the same level of processing of personal data and should not be subject to the same controls as developers training the original model.

1. THE ICO SHOULD RE-EVALUATE THE ARTICLE 6(1) LEGAL BASES FOR LAWFUL PROCESSING UNDER THE UK GENERAL DATA PROTECTION REGULATION TO INCLUDE ARTICLE 6(1)(E) AS A LAWFUL BASIS FOR WEB SCRAPING DATA IN PUBLIC SECTOR AI

1.1. We agree with the Information Commissioner's Office (ICO) assessment that the legal basis for training generative artificial intelligence (AI) models with web-scraped data falls under article 6(1) of the UK General Data Protection Regulation (GDPR). Of the six legal bases however, we believe that in addition to the application of article 6(1)(f), article 6(1)(e) also applies.

1.2. Article 6(1)(e) of UK GDPR allows for the processing of personal data where it is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller. We believe there is scope for the applicability of this article to several different use cases in the public sector, which would rely on web scraping to train generative AI models.

1.3. Given the benefit that generative AI poses to the public sector, clearly establishing a legal basis under this article would afford better clarity to both public bodies and private sector companies providing products for the public sector on their ability to use web-scraped data to train models for public sector use.

1.4. For example, a wealth of case law exists on the Internet, freely available through sources such as the British and Irish Legal Information Institute (Bailii). Databases of case law exist covering not only England and Wales but jurisdictions such as the United Arab Emirates, Ireland, and the Cayman Islands. This type of data would be crucial to a model trained for the purpose of justice administration, given that this particular sector is notorious for its lack of digitalisation. Web scraping offers an out-of-the-box alternative that would encourage AI uptake in the justice sector.

1.5. Similarly, models based on web scraping can be used to provide insight at the governmental level, such as the TrueInflation project, which tracked Argentine inflation when official statistics were not credible. This work is invaluable for research and accountability in an era of widespread misinformation and disinformation, and it is made possible through web scraping.

1.6. Therefore, we recommend re-evaluating the article 6(1) legal bases to expand the consideration of public sector AI and training used for models servicing the public.

2. THE ICO SHOULD CONSIDER THE BROADER PURPOSE OF BUILDING AN AI MODEL AS OPPOSED TO THE PURPOSE THE MODEL IS USED FOR IN THE APPLICATION OF LEGITIMATE INTERESTS

2.1. Data controllers must pass a three-part test to determine if they meet the legitimate interest basis to process data, the first of which is the purpose test (i.e., is there a valid interest?).

2.2. We disagree with the ICO's assessment that developers training AI models must "frame the interest in a specific, rather than open-ended way." The ICO argues that developers are unable to ensure downstream compliance with data protection if they do not know at the outset the purpose of the model.

2.3. We think this framing is wrong and disregards as a possibility the broader purpose of simply training a model. This situation is akin to the web scraping involved with search engines.

2.4. Web scraping for search engine functionality is necessary to provide an index of possible search results. However, the search engine provider cannot anticipate the downstream uses of its product (i.e., what users search). Despite the potential for misuse, the ICO does not prevent search engines from scraping to provide this service.

2.5. The ICO should take the same approach towards developers training foundation models. The purpose of training a model to provide some functionality, even if the developer does not know its specific uses, should be sufficient to create a valid interest. Any downstream violations of data protection law should be treated similarly to the results shown with a search engine.

3. WE SUPPORT THE ICO'S NECESSITY ASSESSMENT THAT WEB SCRAPING IS THE LEAST INTRUSIVE WAY TO TRAIN GENERATIVE AI MODELS

3.1. The second step in the ICO's consideration of legitimate interest is the necessity test (i.e., is web scraping necessary given the purpose?).

3.2. We agree with the ICO's assessment that most generative AI training is only possible using the volume of data obtained through large-scale scraping.

3.3. We would also like to highlight that there appears to be no practical alternative to scraping given the volume of data needed. Arguments have been made about the use of synthetic data to fine-tune models, which can only extract so much from the Internet;

however, synthetically generated data is done from other types of AI models, which first must be trained on real data.[1]

3.4. Moreover, scraping data publicly accessible on the Internet has little marginal impact on individual rights.

3.5. Voluntary, non-regulatory standards are respected by AI development companies such as OpenAI, including the Robots Exclusion Protocol, which restricts web crawlers by disallowing access to certain sites. In fact, nearly 20 percent of the top 1,000 websites in the world are blocking crawler bots that gather data for AI services.[2] Similarly, Adobe has proposed a "Do Not Train" metadata label for creators to inform companies of their position.[3]

3.6. These voluntary efforts are a big indicator that rights can still be maintained whilst enabling innovation through web-scraping, and that the private sector often takes it upon itself to strengthen rights protection.

4. THE ICO SHOULD ASSUME A REASONABLE EXPECTATION OF PROCESSING BY DATA SUBJECTS FOR DATA WILFULLY MADE AVAILABLE TO THE PUBLIC VIA THE INTERNET

4.1 The final part of the legitimate interest test is to evaluate the balance of competing interests between individual rights and the interest of the generative AI developer.

4.2 The ICO holds the view that if individuals would not reasonably expect processing of their personal data, or doing so would cause unjustified harm, their interests are likely to override the legitimate interests of the developer.

4.3 Applying this logic, personal data made purposefully accessible on the Internet would reasonably expect some processing activity. Search engines also engage in the "invisible processing" of web-scraping, for which people do have a reasonable expectation if they are able to access their own data via the search engine. Therefore, the existence of data on the Internet should itself be an indication of a minimum level expectation of processing.

4.4 In addition, the fact that it has been made public would drastically limit the likelihood of unjustified harm since it was purposefully made publicly available and, therefore, potentially widely accessible.

---

[1] "Why computer-made data is being used to train AI models," Madhumita Murgia, *Financial Times*, July 19, 2023.
[2] "Major websites are blocking AI crawlers from accessing their content," Sara Fischer, *Axios*, August 31, 2023.
[3] "Responsible innovation in the age of generative AI," Dana Rao, Adobe Blog, March 21, 2023.

4.5 Data not made publicly available by the data subject yet accessible on the Internet is difficult to navigate, particularly without effective monitoring and enforcement of takedowns. Whilst the risk of unjustified harm is greater, any downstream infringements resulting from this initial action should be attributed to the original infringer (the individual or company wrongly making the data accessible on the Internet), rather than subsequent users of that data who would be unaware of its origins and unable to ascertain these origins using reasonable means.

4.6 Similarly, the unjustified harm highlighted in the downstream risks portion of the consultation covers broad risks that are not generative AI-specific. False information can be generated just as easily by humans as it can by generative AI, and the simple fact the information is available on the Internet in the first instance is the main cause.

4.7 The very nature of the data being available to scrape on the Internet should automatically satisfy the balancing test in this context. The ICO should, therefore, assume a reasonable expectation of processing unless specifically stated otherwise, through such mechanisms mentioned above in point three.

5. DOWNSTREAM USES OF A TRAINED MODEL DO NOT ENGAGE IN THE SAME LEVEL OF PROCESSING OF PERSONAL DATA AND SHOULD NOT BE SUBJECT TO THE SAME CONTROLS AS DEVELOPERS TRAINING THE ORIGINAL MODEL

5.1 The consultation segregates risk mitigation of the balancing test based on how a model is made available on the market. Whilst different distribution models can employ different potential controls, the ICO should be careful not to penalise distributions that do not subscribe to the same level of control.

5.2 We believe the ICO's role does not extend beyond the data pre-processing stage, as highlighted in Figure 1 of the consultation, because once a model is trained and ready to be distributed, there is no longer a need to use the original personal data used for training.

5.3 Once a model is trained, it loses the level of granularity needed at the data pre-processing stage. Further infringements on individual rights are impossible because any downstream users are unable to access the data trained on it through the model.

5.4 Therefore, at the distribution stage, the concerns about open-source availability are not as severe as one might think. Developers releasing open-source models do not expose personal data because this is not useful for further development. Instead, model weights are exposed, which do not touch on personal data processing.

5.5 Similarly, further fine-tuning would simply break the chain of causation. If new data is injected into a distributed model, the third-party developers would be responsible for any GDPR compliance because of their own processing activity.

5.6 The more likely case for fine-tuning, however, is through synthetic data, which typically achieves better results or voluntarily with reinforcement learning by human feedback (RLHF).[4] Neither of these solutions involve the processing of personal data whatsoever.

5.7 The same analysis would also be applied to distributions made available through an application programming interface (API). Whilst developers maintain a level of control over the original model and thus can introduce additional privacy walls or safeguards, the availability of a trained model via an API makes no difference on personal data implications than an open-source distribution. Thus, the fact an API distribution does not impose any further privacy safeguards should not be an issue.

5.8 In evaluating the risk to individual rights based on different distributions, the ICO should carefully consider exactly what processing is done with each distribution. We argue the most significant, potentially harmful point of processing data is done within the control of the original developers at the data pre-processing stage, to which the remit of the ICO would apply. Subsequent distributions are unlikely to engage in the same level of processing, and any new processing activity would shift obligations away from the original distributors.

[4] See footnote 1.