WRITTEN EVIDENCE
ICO GENERATIVE AI CONSULTATION

THE LAWFUL BASIS FOR WEB SRACPING TO TRAIN GENERATIVE AI MODELS

Dr Ann Kristin Glenster
February 2024

MINDEROO
CENTRE FOR
TECHNOLOGY
& DEMOCRACY

Generative AI consultation: the lawful basis for web scraping to train Gen AI models

Summary of Written Evidence Submission

1. This is a written submission of evidence to the Information Commissioner's Office ("ICO") consultation on the lawful basis for web scraping to train generative artificial intelligence ("Gen AI") models, closing 1 March 2024.[1]
2. This submission is made by the Minderoo Centre for Technology and Democracy, an independent team of academic researchers at the University of Cambridge, who are radically rethinking the power relationships between digital technologies, society, and our planet.
3. While we do not disagree with the ICO's regulatory approach, in this submission we ask for clarification regarding the use of legitimate interest for web scraping in relation to special category data (Article 9 UK GDPR), data provenance as personal data is scraped by a second data controller, the extent to which the information requirements (Articles 12-14 UK GDPR) are met, and reasonable expectations. We also query how data protection by design and default (Article 25 GDPR) will be implemented regarding the use of personal data in the development of Gen AI models.
4. While the ICO is narrowly seeking evidence on the lawful bases for processing, we seek clarification on the ICO's approach to web scraping and the data processing principles (Article 5 UK GDPR), and further guidance on the safeguarding of researcher access (Article 89 UK GDPR) to data and systems used to develop Gen AI models.

Introduction

5. This is a submission for the ICO's consultation for evidence regarding the lawful basis for processing web scraped data to train Gen AI models. In soliciting views on its regulatory approach, the ICO is particularly interested in the use of legitimate interest (Article 6(1)(f) UK GDPR) as the legal basis for the processing of training data.

6. The training data is most likely harvested from accessible sources, from either direct 'web scraping' or from the use of data that has been 'web scrapped' by another data controller. The ICO defines web scraping as involving:

> "the use of automated software to 'crawl' web pages, gather, copy and/or extract information from those pages, and store that information (e.g., in

[1] https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-first-call-for-evidence/.

a database) for further use. The information can be anything on a website – images, videos, text, contact details, etc."[2]

7. According to the approach taken by the ICO, when relying on legitimate interest to use web scraped data to train Gen AI models, developers must be able to:

    - "Evidence and identify a valid and clear interest;
    - Consider the balancing test particularly carefully when they do not or cannot exercise meaningful control over the use of the model; and
    - Demonstrate how the interest they have identified will be realised, and how the risks to individuals will be meaningfully mitigated, including their access to their information rights."[3]

Legal Basis for web scraping for AI Models

8. The ICO states that developers of Gen AI models must ensure that the collection of personal data to be used as training data complies with data protection. The ICO suggests that legitimate interest (Article 6(1)(f) UK GDPR) can be used as the legal basis in "some circumstances" if the data controller demonstrates: (1) that the purpose of the processing is legitimate; (2) that the processing is necessary for that purpose; and (3) that the individual's interests do not override the legitimate interest of the data controller.

9. It would be helpful to know which other circumstances the ICO believes would make a different legal basis appropriate for the use of personal data in training Gen AI models. For example, it is difficult to see how legitimate interest could be used in regard to special category data (Article 9 UK GDPR),[4] given the fact that special category data can only be processed with explicit consent of the data subject (Article 9(2)(a) UK GDPR).[5] This is relevant to the consultation in that indiscriminate web scraping (for example using a web crawler) is unlikely to differentiate between ordinary personal data and special category data. We therefore argue that legitimate interest cannot be used for special category data and thus, unless there is a robust mechanism to distinguish these two forms of data at the point of collection, it is doubtful that web scraping could be based on Article 6(1)(f) UK GDPR.

10. It is a further concern that while the ICO's regulatory approach sets out three steps which must be satisfied by the data controller, it does not concern itself with the provenance of the data. The ICO's regulatory approach could be interpreted to suggest that personal data was made available on the Internet for the purpose of being web scraped by a third party at an unknown point in the future. This is unlikely to be the case. Instead, the original data controller will have relied on one of the six legal bases, none of which would allow for the further processing of web scraping (Article 6(4) and Recital 50 UK GDPR). However, we accept that there is a legal lacuna in that the *original* data controller does not carry out the

---

[2] ibid.

[3] ibid.

[4] Special category data is "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."

[5] There are exceptions to the consent requirement in Article 9(2)(a) UK GPDR, but none apply to the use of personal data to train Gen AI models.

the web scraping and thus cannot be legally reliable for its further processing by a different data controller; the second data controller is collecting personal data that is in the public domain, and therefore readily available regardless of the purpose for which it was posted on the Internet. Clarification of the ICO's approach to the legality of the provenance of the personal data that is being web scraped for the purposes of training Gen AI models would therefore be welcomed.

11. In terms of the balancing act of weighing the interests of the individual against those of the data controller, the ICO suggests that both upstream and downstream risks and harms should be considered. Upstream risks include a loss of control over personal data or and a negative impact on fairness. Downstream risks and harms are potential for dissemination of inaccurate information which may cause distress or reputational damage. There is also a risk of exposure to malignant actors, such as scammers and hackers.

12. The ICO consultation highlights that the individual's interests would include being protected from downstream uses will "respect data protection and people's rights and freedoms."[6] This would necessitate ensuring that individuals received information about the processing of their personal data (Articles 12-14 and Recital 61 UK GDPR) throughout the lifecycle of the use of the data. In this context, it is notable that the Polish Supervisory Authority ("SA") in March of 2019 fined a commercial company €220,000 for failing to inform the public of how it web scraped 7.6 million public records. The SA found that placing an information notice on the company website did not meet the legal threshold for 'disproportionate effort' under Article 14(5)(b) GDPR.[7] It is therefore difficult to see how individuals can feasibly be informed of the collection and use of their personal data as training data in a manner that will satisfy the requirements in Articles 12-14 UK GDPR. Further elucidation of the ICO's approach to determining what would constitute a disproportionate effort under Article 14(5)(b) UK GDPR, and especially Articles 13(3) and 14(3) UK GDPR concerning further processing, in this regard would be helpful.

13. This is particularly relevant in regard to the second step of the balancing test set out by the ICO of identifying and weighing the interest of the individuals against the legitimate interest of the data controller. According to Recital 47 UK GDPR: "…the existence of a legitimate interest would need careful assessment including whether a data subject can reasonably expect at the time and in the context of the collection of the personal data that processing for that purpose may take place. The interests and fundamental rights of the data subject could in particular override the interest of the data controller where personal data are processed in circumstances where data subjects do not reasonably expect further processing."

---

[6] Supra note 1.

[7] https://www.insideprivacy.com/data-privacy/polish-supervisory-authority-issues-gdpr-fine-for-data-scraping-without-informing-individuals/.

14. It would be helpful if the ICO could clarify what would constitute a reasonable expectation in this regard as it is unlikely that individuals would expect their personal data to be subjected to large-scale web scraping for the training of Gen AI models. This is a particularly thorny issue in cases where the original processing which led to the personal data being made available on the Internet was obtained using consent (Article 6(1)(a) UK GDPR) as the legal basis for that processing. For example, it is difficult to see how the original data controller would be able to fulfil the obligation to provide a mechanism for the withdrawal of consent (Article 7 UK GDPR) for personal data that had subsequently been web scraped by a second data controller.

15. It may be argued that training data is different from input data and as the data is likely to only be 'machine read', it would qualify as pseudonymised data (Article 4(5) UK GDPR[8]). However, pseudonymisation is not used in the GDPR as a way by which to disapply the data protection requirements, but rather as a measure to mitigate individuals' exposure to risk (Article 25 and Recitals 28 and 78 UK GDPR).

16. If the technical system has been designed in accordance with the requirements of data protection by design and default (Article 25 GDPR) it should be impossible for a data controller to extract identifiable personal data from the training data if the processing of that data is wholly automated. Questions remain outstanding, however, in relation to the classification and labelling of that data, particularly in relation to is granularity, and the capacity of the algorithm (as it 'trains' itself) to link or combine that data in ways which may relate or link to an identifiable natural person.

Further Consultations

17. The use of web scrapped personal data raises significant questions for the feasible enforcement and application of data protection. Further clarification is needed regarding the obligations data controllers will have in relation to special category data, further processing, pseudonymisation, information requirements, user rights, data protection by design and default, and the data processing principles, especially 'purpose limitation' (Article 5(1)(b) UK GDPR), 'data minimisation' (Article 5(1)(c) GDPR), and 'storage limitation' (Article 5(1)(e) UK GDPR). We would recommend further ICO consultations into these aspects of data protection as they relate to the development of Gen AI models in the UK.

18. We are concerned that personal data is being scraped and used to build large-scale Gen AI models without robust research into the potential effects this may have on individuals' fundamental interests and societal trust in these technologies. Thus, to ensure that Gen AI models developed in the UK are responsible, ethical, legal, safe, fair, and trustworthy, we urge the ICO to issue guidance to support researcher access to data (Article 89 UK GDPR) to ensure that independent researchers are able to study these models before they are put into the market and throughout their lifecycle.

---

[8] 'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

About the Minderoo Centre for Technology and Demoracy

The Minderoo Centre for Technology and Democracy is an independent team of academic researchers at the University of Cambridge, who are radically rethinking the power relationships between digital technologies, society and our planet.

For more information visit www.mctd.ac.uk

Get in touch: email us at minderoo@crassh.cam.ac.uk