Response by the Centre for Information Policy Leadership to the Information Commissioner's Office's Consultation on the Lawful Basis for Web Scraping to Train Generative AI Models

Submitted March 1, 2024

The Centre for Information Policy Leadership (CIPL) welcomes the opportunity to respond to the Information Commissioner's Office's (ICO) Consultation on the lawful basis for web scraping to train Generative AI (Gen AI) models.

For more than 20 years, CIPL has been on the forefront of promoting organisational accountability and a risk-based approach as cornerstones of effective data protection law, policy, and oversight. As noted in our white paper *Ten Recommendations for Global AI Regulation*[1], CIPL advocates that any regulatory approach to AI should seek to protect fundamental human rights and minimise risks of harm to individuals and society, while enabling the responsible development and deployment of AI. CIPL has recently published a report entitled *Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework*[2], which showcases how 20 leading organisations are developing accountable AI programs and best practices for developing and deploying AI on the ground through the lens of CIPL's Accountability Framework.

CIPL's feedback on the Consultation will focus on 1) the ICO's analysis 2) the technical and organisational measures developers should implement to limit the ways customers can use Gen AI models, and 3) the aspects of this topic CIPL would like the ICO to consider in future publications.

1) **Do you agree with the analysis presented in this document?** *Yes.*

CIPL is still shaping our thinking on public policy and governance of Gen AI given the fast evolution of the technology. At the same, we can draw on durable principles from our work to date in order to share perspectives for this consultation. CIPL previously assessed the application of GDPR to AI[3] and addressed key tensions in applying data protection principles to AI.[4] In the latter paper, we also emphasised the need for regulators to evolve the interpretation of GDPR principles in light of technological developments to ensure they remain valid and fit for purpose for Gen AI technologies. The ICO consultation seeks to do just that, and we welcome the approach.

CIPL agrees that legitimate interest can be a valid, and appropriate, legal basis for scraping publicly available personal data and processing this data for the purpose of training Gen AI models. Use of

---

[1] CIPL, "Ten Recommendations for Global AI Regulation", October 2023, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ten_recommendations_global_ai_regulation_oct2023.pdf.

[2] CIPL, "Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework", February 2024, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_building_accountable_ai_programs_23_feb_2024.pdf.

[3] CIPL, "Artificial Intelligence and Data Protection: How the GDPR Regulates AI," March 2020, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf.

[4] CIPL, "Second Report: Hard Issues and Practical Solutions," February 2020, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_hard_issues_practical_solutions_english_feb20.pdf.

legitimate interest as a legal basis for training Gen AI models requires putting in place appropriate guardrails to keep data safe, controls and accountability measures, as well as robust oversight and enforcement.

In order to rely on legitimate interest, a controller must be able to demonstrate that there is a specific valid interest to process personal data in this way. Similar to the 2013 opinion of Advocate General Jääskinen in the context of search engines, which informed the landmark CJEU *Costeja* decision,[5] it would similarly seem appropriate for controllers developing Gen AI models to rely on legitimate interest as a ground for processing personal data sourced from the internet for the purpose of training the models, given the broad societal benefits the technology can bring . The ICO has long supported a broad interpretation of what can constitute a legitimate interest, including societal benefits, and the potential benefits of Gen AI appear to be large across the economy and society in fields as diverse as medicine,[6] science, agriculture and business.[7] More specifically, in this case there is a legitimate interest to ensure robust, accurate, safe and non-biased functioning of algorithms by training models using large, diverse and high quality sets of data.

CIPL acknowledges that at present many developers rely on web scraping of publicly available data for training Gen AI models. Publicly available data is at the core of how many Gen AI models are trained; it is foundational to model quality and functionality. Gen AI requires data to learn how language incorporates concepts about relationships between people and the world. A Gen AI model is not so much a "database" holding data, but rather, an algorithm that has learned patterns and relationships in data and uses them to predict the next probable words or images in a sequence.

However, controllers must put in place demonstrable policies and procedures to ensure that personal data are processed responsibly. This includes instituting safeguards against models being used to harm the fundamental rights of individuals whose personal data may be processed within the models, such as by enabling users to derive inferences on specific people's sensitive characteristics.[8] Removing personal data from the data collection and training stage of Gen AI is an

---

[5] Opinion of Advocate General Jääskinen, https://curia.europa.eu/juris/document/document.jsf?docid=138782&doclang=EN, para 95. To quote from this paragraph: "As to the criteria relating making data processing legitimate in the absence of a data subject's consent (Article 7(a) of the Directive), it seems obvious that provision of internet search engine services pursues as such legitimate interests (Article 7(f) of the Directive), namely (i) making information more easily accessible for internet users; (ii) rendering dissemination of the information uploaded on the internet more effective; and (iii) enabling various information society services supplied by the internet search engine service provider that are ancillary to the search engine, such as the provision of keyword advertising. These three purposes relate respectively to three fundamentals rights protected by the Charter, namely freedom of information and freedom of expression (both in Article 11) and freedom to conduct a business (Article 16). Hence, an internet search engine service provider pursues legitimate interests, within the meaning of Article 7(f) of the Directive, when he processes data made available on the internet, including personal data."

[6] Andrew Myers, "Doctors Receptive to AI Collaboration in Simulated Clinical Case without Introducing Bias," https://hai.stanford.edu/news/doctors-receptive-ai-collaboration-simulated-clinical-case-without-introducing-bias.

[7] Adam Zewe, "Explained: Generative AI," *MIT News,* https://news.mit.edu/2023/explained-generative-ai-1109.

[8] *Google Spain SL and Google Inc v Agencia Espanola de Proteccion de Datos (AEPD) and Mario Costeja Gonzalez*, Case C-131/12, https://curia.europa.eu/juris/document/document_print.jsf?doclang=EN&docid=152065, at para 80. To quote from this paragraph: "It must be pointed out at the outset that, as has been found in paragraphs 36 to 38 of the present judgment, processing of personal data, such as that at issue in the main proceedings, carried out by the operator of a search engine is liable to affect significantly the fundamental rights to privacy and to the

industry-wide effort, which currently still requires robust research and innovation efforts for its development. [9] While significant volumes of data are often required for effectively training Gen AI models, large-scale web scraping of personal data should only be considered if there are no reasonable alternatives.

Furthermore, as the ICO notes, legitimate interest is likely to be the most practicable legal basis for training Gen AI models on personal data scraped from publicly accessible sources. Other legal bases such as consent appear impracticable (e.g. for seeking and withdrawing consent) and may result in incomplete or non-representative training data.

The risk-assessment requirement (or "balancing test", see next question) inherent in the legitimate interest basis also provides for more accountability and controls than any other basis.

2) **As we explain in the consultation, the legitimate interests test could be met if technical and organisational measures to limit the use of the Gen AI model are in place. Do you agree with the analysis we have presented? If yes, what measures should a developer implement to limit the ways in which its customers can use the Gen AI model?** *Yes.*

The final part of the 'three-part' test to meet the legitimate interest basis requires a balancing test as to whether individuals' interests override the interest being pursued, which effectively entails performing a risk assessment on the proposed processing activity. CIPL notes that in addition to the right to privacy, individuals have other fundamental rights that could be put at risk from inaccurate or biased outputs of AI models. It is critical that controllers have in place demonstrable safeguards to protect all these fundamental rights and mitigate risks when training Gen AI models using web-scraped personal data (e.g., performing risk assessments and DPIAs, ensuring data integrity, and providing appropriate transparency).

Regardless, controllers must ensure not just the appropriate legal basis, but compliance with all the other provisions of the GDPR, such as data security, transparency, and rights of individuals, as well as non-infringement of intellectual property, contract and other laws.

In many contexts, it may be advisable that developers of Gen AI models generally build in controls to prevent users from building detailed profiles of individuals, retrieving sensitive information about them, or generating their likenesses, as is already the case for a number of Gen AI tools in broad public use.

Model developers should also be incentivised to use privacy-enhancing technologies where feasible and appropriate, including, but not limited to, synthetic data and federated learning.

---

protection of personal data when the search by means of that engine is carried out on the basis of an individual's name, since that processing enables any internet user to obtain through the list of results a structured overview of the information relating to that individual that can be found on the internet — information which potentially concerns a vast number of aspects of his private life and which, without the search engine, could not have been interconnected or could have been only with great difficulty — and thereby to establish a more or less detailed profile of him. Furthermore, the effect of the interference with those rights of the data subject is heightened on account of the important role played by the internet and search engines in modern society, which render the information contained in such a list of results ubiquitous."

[9] For example, Google announced the first Machine Unlearning Challenge with a broad group of academic and industrial researchers aiming to advance the state of the art in machine unlearning and encourage the development of efficient, effective and ethical unlearning algorithms (https://blog.research.google/2023/06/announcing-first-machine-unlearning.html).

CIPL also asks the ICO to appropriately define key terms such as 'developer' and 'deployer'. The consultation refers to 'customers' and 'third parties', which we assume intend to refer to deployers, but may lead to confusion.

3) This is the first in a series of publications on the ICO's analysis of personal data processing involved in Gen AI. What aspects of this topic would you like us to consider in future publications?

- Address the broad range of tensions between AI technologies and data protection principles, such as the data minimisation principle, legal basis for sensitive data processing, transparency and explainability, rights of individuals (access, objection, deletion), rules on automated decision taking, and international data transfers.[10] In particular, analysis of how to address special category data when training Gen AI models, including whether there is a valid legal basis organisations can rely upon, given that article 9 of the UK GDPR prevents web scraping of special category data. Including guidance on how to address countries where images are considered sensitive data.
- Practical guidance on data ethics and the risk assessment/balancing test for Gen AI, that in addition to the risks to individuals takes also into account the risk of not deploying AI (loss of opportunity) and benefits from deploying AI.
- The role of certifications, codes of conduct, and other accountability tools for responsible Gen AI development and deployment.
- Best practices in red-teaming and other adversarial testing approaches for Gen AI, as well as other approaches to mitigate potential risks associated with potential use of sensitive training data.
- Possible unintended consequences arising from training Gen AI on web-scraped data, such as any limitations associated with this approach related to accuracy and bias.
- Specific considerations in the context of open source versus closed source Gen AI models.
- We recommend that the ICO focus on applications built upon Gen AI models and their uses, as well as the underlying model.
- The ICO could leverage and inform the work of the UK AI Safety Institute, and continue to cooperate with other regulators in this space through the Digital Regulation Cooperation Forum.

[10] See discussion of these issues in CIPL's *Artificial Intelligence and Data Protection: Delivering Sustainable AI Accountability in Practice, First Report* (https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_first_ai_report_-_ai_and_data_protection_in_tension__2_.pdf) and *Second Report* (https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_hard_issues_practical_solutions_english_feb20.pdf) .