Privacy International's response to ICO's fourth call for evidence on Generative AI: engineering individual rights into generative AI models

June 2024

Privacy International (PI)[1] is providing input to the Information Commissioner's Office (ICO)'s fourth call for evidence on generative AI models. This submission builds on our response to the ICO's first call for evidence in this series.[2]

Summary

PI's assessment of the major generative AI models is that they cannot uphold individuals' rights under the UK GDPR. Their development and operation have therefore been in breach of these rights and the only way to fully apply the law is to require deletion of infringing data and re-training of models on compliant data. New technologies designed in a way that cannot uphold people's rights cannot be permitted just for the sake of innovation.

As such, while PI broadly agrees with the ICO's analysis, we submit that a stronger position should be taken with respect to generative AI models. It is unacceptable to rely on untested, unproved and uncertain additional technology (such as 'machine unlearning') to try to fulfil people's rights. Until they offer adequate means for people to exercise their rights, companies developing generative AI should not be allowed to distribute their models or offer services relying on those models.

In the remainder of this submission, we bring the following matters to the ICO's attention. Together they demonstrate how peoples' rights are undermined by generative AI and indicate why the ICO must take a firm approach:

- A high bar must be set for transparency, with significant improvement on past and current practice necessary;
- Input and output filters are inherently unreliable because they cannot exhaustively cover every use case. An illustrative example is the constant discovery of ways to "jailbreak" LLMs;
- Measures to protect personal data must be at least as strong as any measure devised to protect copyrighted materials (such as opt-outs and filters);
- Privacy by default and design should be implemented so as to not place the onus on individuals to take action to prevent invasive practices.

[1] Privacy International (PI) is a London-based non-profit, non-governmental organization (Charity Number: 1147471) that researches and advocates globally against government and corporate abuses of data and technology..

[2] https://privacyinternational.org/advocacy/5263/pi-response-ico-consultation-web-scraping-generative-ai

A high bar for transparency and the right to be informed

The development of generative AI has been dependent on the scraping and processing of publicly available data in ways that could not have been reasonably predicted by the owners and producers of this data. Scraping and processing for generative AI have taken place almost entirely in secret, with leading AI companies showing reluctance to be transparent about their activities. Given the role that generative AI is shaping up to take in our society and economy, it is unacceptable for such a data intensive (and privacy invasive) technology to have such a poor approach towards transparency.

PI firmly agrees with the ICO's position that "the processing of personal data to develop generative AI models is likely to be beyond people's reasonable expectations at the time they provided data to a website." It may even be beyond people's reasonable expectations that data they provide to a website *today* will be used in that way, in light of how poor AI companies have been at explaining the nature of their activities and the sources of their data. Web scraping by AI developers and use of data scraped by others fundamentally goes against the principles of foreseeability and reasonable expectations.[3] It can be readily distinguished from crawling by search engines, which have been around since the early days of web 2.0.

Given that people have no way of knowing that their data has been processed in the first place (and so no reason to be looking for information about it), extra care and effort must be taken to be transparent about it as people cannot exercise any of their data rights if they are unaware that data about them is held in the first place. Serious questions must be asked both by the ICO and by AI developers as to whether any such care and effort can ever reach the transparency standard required by the UK GDPR. In fact, the Dutch DPA has recently published guidelines that state that web scraping almost always violates the GDPR because (in part) of the lack of notification to data subjects that their data is being processed.[4]

In addition to the generally unexpected and hidden nature of web-scraping, the scale and potential societal impact of generative AI data processing means that developers have weighty obligations to make information about their scraping and processing publicly available and understandable. We agree with the ICO that "vague statements about data sources (eg just 'publicly accessible information) are unlikely to help individuals understand whether their personal data may be part of the training dataset or who the initial controller is likely to be." We add that they are unlikely to (1) reach relevant data subjects nor (2) set out with any granularity or accuracy the categories of personal data concerned nor recipients of personal data (required by Arts 14(1)(d) and (e) to be provided), which are potentially extremely large.

[3] According to the Art 29 WP Guidelines on transparency, "a central consideration of the principle of transparency [...] is that the data subject should be able to determine in advance what the scope and consequences of the processing entails and that they should not be taken by surprise at a later point about the ways in which their personal data has been used."

[4] https://autoriteitpersoonsgegevens.nl/actueel/ap-scraping-bijna-altijd-illegaal

Poor practice to date by AI developers on this front further justifies the ICO taking a strong stance. OpenAI's privacy policy, for example, provides extremely limited information on the core of its data collection and processing, i.e. mass scraping from the public internet – none of the information provided is helpful to any data subject to understand the processing of their own data[5] – and hypes the benefits of LLMs while glossing over the risks of mass data scraping.[6] It fails to clarify exactly what data the company holds – for example, sentences such as "ChatGPT does not copy or store training information in a database"[7] are potentially misleading if OpenAI does have such data stored.

As recognised by the exemption in Article 14(5)(b) of the UK GDPR to the requirement to provide information to data subjects when data was not collected directly from them, in some circumstances it may be disproportionate to provide billions of people with individual information about how their data has been processed. However, this exemption applies "in particular" to "processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes", subject to a number of safeguards in Article 89. Commercial AI developers' activities do not fall within these, and hence cannot avail themselves of the disproportionality exemption. Any argument about the proportionality or feasibility of providing individuals with the requisite level of transparency therefore cannot be entertained in a way that complies with the UK GDPR.

Accessing data

The ICO analysis identifies that developers may "argue they are not able to respond to requests [to access a copy of the personal data held about someone] because they cannot identify individuals (in the training data or anywhere else)". We are sceptical about such arguments and suggest that the ICO consider them within the context of the

[5] OpenAI, Europe privacy policy (15 December 2023), https://openai.com/policies/eu-privacy-policy/: "Personal Data We Receive From Other Sources: We collect information from other sources, like information that is publicly available on the internet, in particular to develop the models that power our Services. We also receive information from our trusted partners, such as security partners to protect against fraud, abuse, and other security threats to our Services or marketing vendors who provide us with information about potential customers of our business services." OpenAI, How ChatGPT and our language models are developed (29 May 2024), https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed: "A large amount of data on the internet relates to people, so our training information does incidentally include personal information. We don't actively seek out personal information to train our models."

[6] OpenAI, How ChatGPT and our language models are developed (29 May 2024), https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed: "We use training information lawfully. Large language models have many applications that provide significant benefits and are already helping people create content, improve customer service, develop software, customize education, support scientific research, and much more. These benefits cannot be realized without a large amount of information to teach the models. In addition, our use of training information is not meant to negatively impact individuals, and the primary sources of this training information are already publicly available. For these reasons, we base our collection and use of personal information that is included in training information on legitimate interests under privacy laws like the GDPR, as explained in more detail in our Privacy Policy. We have also completed a data protection impact assessment to help ensure we are collecting and using this information legally and responsibly."

[7] ibid.

product the developers have designed: one that is able to generate, re-produce or hallucinate personal information based on the data they have processed.

A simple engagement with a generative AI chatbot (eg asking 'who is X?') demonstrates that they can provide information about individual people (in particular where those people have an unusual name and/or large online profile, even if not a public figure). This carries risks of harm for that individual whether that information is true (breach of data rights) or false/hallucinated (defamation, misinformation, discrimination). The latter is already subject to a regulatory complaint in Austria.[8]

Rather than debating technical questions over whether or not AI developers and/or models store any data, the ICO should focus on the fact that personal data is processed whenever it is being collected, provided by users or generated by the model. From the perspective of the data subject, it makes no difference whether (mis)information about them has been regurgitated, hallucinated, inferred or looked-up in a database: what matters is the harm suffered.

It is problematic – and potentially obtuse and disingenuous – for AI developers to argue that they do not have to provide people with access to personal data about them just because it is being stored and processed in a new way. Their products are able to identify individuals, which means they are processing personal data relating to those individuals. Data that can be inferred from collected data qualifies as personal data, whether the inference was made by a human, a simple algorithm or a neural network.

Innovative technology and people's rights

The ICO analysis recognises that the new technological context of generative AI may require the use of new tools to protect people's rights. Innovation here may be valuable, but it must not come at the cost of an erosion in legal standards such as transparency and the effectiveness of peoples' rights. While new technologies may entail or require new ways of ensuring rights are being met, it is unacceptable to rely on untested, unproved and uncertain ways of doing.

Filters for LLMs will always be breakable

The ICO specifically asks for "views on whether input and output filters are a sufficient mechanism for enacting people's rights". PI's view is that filters are inherently limited and cannot be relied on as a way of protecting people's rights. This is not a comment on the current quality of filters, but rather of the very design of LLMs.

Filters by design rely on strictly defined parameters which by nature cannot cover the infinite ways that LLMs process input and generate output. One can make a comparison with how security is approach in "traditional" systems. If an input if offered to a user, it will be sanitised, inspected and submitted to filters to ensure that nothing provided by the user in this input leads to a disfunction or vulnerability in the system. A classic example is the SQL injection[9] where an input can be abused to gain access to a database. Such

[8] Noyb, ChatGPT provides false information about people, and OpenAI can't correct it (29 April 2024), https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it.
[9] https://en.wikipedia.org/wiki/SQL_injection

situations can be avoided and risks minimised because developers know exactly how and where the input will be processed by the system. Nonetheless, and despite these risks being known for decades, vulnerabilities are still regularly discovered.

In the case of LLMs, the system with which the user interacts directly is the one that also generates the output and developers do not fully understand how it will be used to generate the output, meaning there are potentially infinite ways to make LLMs behave differently than the developers intended. As explained by Bruce Schneier, this means that commands can always be manipulated (like payphones that could be tricked into giving free calls by whistling at a certain frequency).[10] The constant discovery of jailbreaks, enabling use of models in unintended or non-authorised ways, illustrate that manipulation.

Because LLMs do not separate their commands (prompts) and the data itself, or because of the dependence on input and the inscrutable way data is being generated by models, filters will always be breakable. If the model has been trained on certain data, there are or will certainly be a way to access it in the future.

Opting out

A number of mechanisms seek to facilitate people opting out of information being used to train generative AI. In addition to users of generative AI being able to opt out of their inputs being used to train, there are also wider approaches that are conceptually similar to robots.txt files or the Do Not Track (DNT) / Global Privacy Control (GPC) header fields. The idea is to signal to AI developers that the relevant material should not be used for training generative AI models.

For example, OpenAI's Media Manager is "a tool that will enable creators and content owners to tell us what they own and specify how they want their works to be included or excluded from machine learning research and training"[11] and Spawning have built a Do Not Train Registry which "consolidates machine-readable opt-out methods".[12]

We draw three matters to the ICO's attention in relation to these:

- An opt-out model may improperly reflect the surprising, intrusive and far-reaching nature of the processing, in particular for data produced before 2022. An opt-in model may therefore be more appropriate and in line with the data protection principles enshrined in the UK GDPR.
- The ICO should assess the case for their being a standard for opt-outs in order to prevent a confusing proliferation of approaches and consequent ease with which they could be evaded. Any standard may need legal underpinning to have full effect.
- At the very least, any standard or approach adopted to protect copyrighted material should also be used to protect personal data. Data protection rights are

[10] https://cacm.acm.org/opinion/llms-data-control-path-insecurity/
[11] https://openai.com/index/approach-to-data-and-ai/
[12] https://spawning.ai/

protected by human rights law and deserve as much, if not more protection than intellectual property rights which are largely commercial assets.

Machine unlearning and other new technologies

Techniques such as 'machine unlearning', pseudonymisation, relying on AI itself to develop safety mechanisms such as filters and other privacy-enhancing technologies should only be relied on where they can be shown to meet current legal standards (ie rather than being better than any other way of doing it).

As with opt-outs, any technological safeguards for copyrighted data should apply at least as strongly for personal data. It will be important for AI developers to be more open about how their technology works (eg through greater access to sandboxes or the source code itself) if they want people to be assured of the effectiveness of their safeguards.

Final thoughts – who is responsible?

The creation of a new technology (in this case, generative AI) does not change the law. Many of the rights people have under the GDPR are not for results of best effort and therefore the argument that something is technically hard or novel provides no defence against non-compliance. This may be especially important where the new technology is widespread and the subject of considerable societal and economic upheaval. A high bar for transparency is needed if generative AI is to be a trusted and valuable contributor to society.

LLMs are not about to disappear, but rights have been already violated in their creation. A gung-ho attitude to tech development that runs roughshod over people's rights and believes that it is easier to ask for forgiveness than permission must not be tolerated. There are lessons to learn from the harms that have arisen from poor regulation of social media companies to refute the idea that AI developers cannot or should not be held responsible for how their products work and the material they generate.[13]

Finally, we urge the ICO to be careful of placing too much onus and responsibility on individual control and action. Especially in an area where people are ill-equipped to access information or understand the technicalities, we must not be reliant on people seeking out information and exercising their rights. Invasive – and potentially illegal – practices should be stopped at the outset, not only once people have objected to them.

Deletion of the datasets collected unlawfully and the models trained on those is the only way to redress the past and future harms while sending a strong message to any future technology.

[13] See https://www.create.ac.uk/blog/2024/05/29/new-working-paper-private-ordering-and-generative-ai-what-can-we-learn-from-model-terms-and-conditions/ and https://www.technologyreview.com/2024/03/13/1089729/lets-not-make-the-same-mistakes-with-ai-that-we-made-with-social-media/