28 November 2024

Dear EDPB Chair, EDPB Members, and the EDPB Secretariat,

We are writing to provide the views of Privacy International (PI) in relation to the EDPB's forthcoming opinion on artificial intelligence (AI) models, following a request submitted to the EDPB by the Irish Data Protection Authority (DPA) under Article 64(2) GDPR.

## Introduction

Privacy International (PI) is a global, not-for-profit organization that campaigns against companies and governments who exploit our data and technologies. We have long been involved in the development and enforcement of data protection law in Europe and other countries and advocate for its strict enforcement to limit harms arising from the exploitation of personal data and improve access to privacy and security.

PI's assessment of AI models is that they have been trained based on the processing of personal data without an adequate legal basis and that they are unable to uphold individuals' rights under the GDPR. Their development and operation have therefore been in breach of these rights. New technologies designed in a way that cannot uphold people's rights must not be permitted just for the sake of innovation.

The approach taken by the EDPB towards generative AI models may have important downstream repercussions for the future of people's information rights online. If the balance it strikes is wrong with respect to emerging practices, then people stand to have their rights under the GDPR further violated by other new and emerging technologies. The growth of generative AI and LLMs is likely to further drive business models that depend on large scale scraping, processing, and potentially exploitation, of personal data with limited regard for people's rights and interests. If the interest in LLMs and products trained on vast datasets continues growing, it is likely new incumbents would be incentivised to invest in mass scraping capabilities – increasing their prevalence and potentially their harm, unless steps are taken to require less intrusive practice.

As such, we submit that a strong position should be taken with respect to generative AI models. It is unacceptable to rely on untested, unproven and uncertain additional technology (such as 'machine unlearning') to try to fulfil people's rights.

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

1. The fundamentally general nature of AI models creates problems for the legitimate interest test;
2. The risks of an overly permissive approach to the legitimate interests test;
3. Web scraping as 'invisible processing' and the consequent need for transparency;
4. Innovative technology and people's fundamental rights;
5. The (in)adequacy of filters and other similar safeguards; and
6. Opting out of opt outs.

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. The fundamentally general nature of AI models and the legitimate interests test

There may be both business interests and societal interests that could, in theory, qualify as a legitimate interest for scraping data from the web to train a generative AI model. Importantly however, assessment of whether these interests are being met in practice is challenging, perhaps impossible, because AI models are often developed without a particular end use in mind. It is also unclear how upstream development can ensure that its (unknown) downstream use(s) will in practice respect data protection principles and people's rights.

These challenges are fundamentally inherent to the very design of most generative AI models: which is intentionally to be of a general and indiscriminate nature rather than only for a specific purpose. They can be used to draft legal submissions to courts,[1] to generate harmful pornographic content,[2] to provide instructions on building bombs,[3] to produce misleading content about high-profile and/or elected personalities,[4] or for military purposes.[5]

The specific purpose and use of a generative AI model may therefore be impossible to determine at the point that data is collected by scraping and then processed to train the model. Recently developed services already illustrate this aspect of their model, OpenAI for example offers a "GPT store"[6] that provides access to a variety of GPT-based chatbots with widely different purposes, from academic research assistants to text-to-speech tools for maths tutors.

This inherent and inbuilt uncertainty means that relying on legitimate interests to scrape data from the web to build generative AI models is rife with problems. We are not convinced that existing practice (where data collection is indiscriminate and outputs are unpredictable) stands up to the scrutiny and standards established by the GDPR and Data Protection Authorities for the protection of people's rights and the rigour of legitimate interest assessments.

## 2. Risks of an overly permissive approach to the legitimate interest test

The unavoidably generic nature of web scraping for generative AI creates wide-ranging and far-reaching implications of any assessment under the legitimate interest test. If the legitimate interest can be a lawful basis for training generative AI models on web-scraped data, then what other forms of large-scale web scraping of personal data may be allowed under the legitimate interest test? A lack of precision here may leave the door wide open for personal data to be misused or abused in the future in wider contexts as technology develops.

### *Collaterally justifying other forms of large-scale data scraping*

Permitting developers to scrape large amounts of personal data from the web to train AI models could risk collaterally opening the doors for other entities to justify large scale collecting of personal data under the same pretence of the "legitimate interests" of the business. It may even further incentivise and/or legitimise the development of new business models that depend on web

[1] https://www.bbc.co.uk/news/world-us-canada-65735769
[2] https://theconversation.com/ai-generated-pornography-will-disrupt-the-adult-content-industry-and-raise-new-ethical-concerns-226683
[3] https://www.newscientist.com/article/2450838-writing-backwards-can-trick-an-ai-into-providing-a-bomb-recipe/
[4] https://aiforensics.org/work/bing-chat-elections
[5] https://theintercept.com/2024/10/25/africom-microsoft-openai-military/
[6] https://chat.openai.com/gpts

scraping and other large scale and indiscriminate means of data collection, such as that of Clearview AI.[7]

*Different types of data collected*

The legitimate interest test is also difficult, if not impossible, to properly apply to large-scale web scraping in part because blanket scraping cannot easily discern between the types of data it collects. It therefore cannot properly assess the consequences of processing for relevant data subjects. A well-designed crawler and scraper might access content that was inadvertently public, including databases with inadequate protections such as the ones regularly discovered by security researchers and malicious actors. Consequently, we believe there must be limits to permitting the legitimate interest legal basis to apply to large-scale data scraping.

*Third-party deployment of models*

The legitimate interest assessment is further complicated by developers making models available to third parties (whether on an open-source or closed-source basis). The purpose limitation principle requires that developers define why they are processing personal data and only process data for that purpose. However, how can developers demonstrate that their models are meeting their identified purpose when they release their models to third parties who might tailor them to their unique needs beyond the developers' intended purpose? This is particularly problematic for open-weights and open-source models which are freely accessible, fine-tuneable and can be used a wide variety of scenarios.

## 3. Invisible processing and the need for transparency

Web scraping is a form of 'invisible processing': individuals may not be aware that their personal data is being scraped to train a generative AI model as that has taken place almost entirely in secret. Invisible processing can restrict people's knowledge about, and frustrate their ability to exercise, their rights. There is an inherent tension between invisible processing and the exercise of information rights that arises in the context of web scraping and generative AI.

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 at the time they created the data. It may even be beyond people's reasonable expectations that data they provide to a website *today* will be used to train AI models, in light of how poor AI companies have been at explaining the nature of their activities and the sources of their data.[8]

Web scraping by AI developers and use of data scraped by others fundamentally goes against the principles of foreseeability and reasonable expectations.[9] 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, extra effort must be taken to be transparent so that people are able to exercise their rights under the GDPR. Serious questions must be asked as to whether such care and effort can ever reach the transparency standard required by the GDPR. As noted by the Dutch DPA, indiscriminate web

[7] https://privacyinternational.org/legal-action/challenge-against-clearview-ai-europe
[8] https://www.theverge.com/2024/4/6/24122915/openai-youtube-transcripts-gpt-4-training-data-google and https://hdsr.mitpress.mit.edu/pub/xau9dza3/
[9] 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."

scraping almost always violates the GDPR because (in part) of the lack of notification to data subjects that their data is being processed.[10]

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. These cannot be merely vague and general statements as these are unlikely to reach relevant data subjects and do not set out with any granularity or accuracy the categories of personal data concerned or recipients of personal data (required by Arts 14(1)(d) and (e) to be provided), which are potentially extremely large. If web scraping for the purposes of generative AI development can be shown to have a lawful basis, then its extremely high risks mean that strict monitoring and abundant transparency would be proportionate.

As recognised by the exemption in Article 14(5)(b) of the 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 data subjects 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 GDPR.

## 4. Innovative technology and people's fundamental rights

Innovation in the advancement of AI models may have societal value, 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, unproven and uncertain ways of doing so.

The application of legal standards in the early days of a technology can have a critical influence on the development and propagation of practices, with the potential to encourage similarly abusive behaviour in future innovations.

The dependence of AI on web scraped data and the lack of real-time human oversight over its outputs creates unavoidable risks of output harms, including from a data protection perspective. While no technology is entirely secure from hacks or breaches, the key difference in the case of AI models is that the black-box nature of the algorithm means that there are an infinite number of potential vulnerabilities as opposed to "hard-coded" algorithmic logic which can be manually fixed and secured after, for example, a security audit.

We are sceptical about arguments that controllers of AI models cannot or need not comply with rights to access or erasure because they cannot identify individuals in their training data. This does not align with the reality of the product the developers have designed: one that is able to generate, re-produce or hallucinate personal information based on the data they have processed. Nor does it align with the information security researchers have managed to extract from those models using different jailbreaking techniques.

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

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

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 subject to a regulatory complaint in Austria.[11]

Rather than debating evasive questions over whether or not AI developers and/or models *store* any data, the EDPB 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, innovative, 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.

Likewise, all other established rights under the GDPR including to erasure and rectification must be complied with by AI models, even if this presents technical challenges due to their innovative structure. Controllers are not permitted to select which rights to comply with or to limit the extent with which they do so.

## 5. The (in)adequacy of filters and other similar safeguards

Developers may seek to implement technical measures to guard against the known and emerging risks of AI models. However, evidence abounds that technical guardrails are not robust enough to protect against inevitable misuse. Techniques such as 'machine unlearning', pseudonymisation, and 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).

Input and output 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 approached in "traditional" systems. An input offered by a user 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[12] where an input can be abused to gain access to a database. Such 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 same one that also generates the output. But 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

[11] https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it.
[12] https://en.wikipedia.org/wiki/SQL_injection

frequency).[13] The constant discovery of jailbreaks,[14] enabling use of models in unintended or non-authorised (and potentially harmful) ways, illustrate that manipulation.

Researchers from Carnegie Mellon University (CMU) developed 'adversarial attack' methods[15] and concluded that jailbreaking can be automated[16] in such a way that there is an unknown and unlimited number of ways to break in. The CMU research concluded: "it is unclear whether such behaviour can ever be fully patched by LLM providers … It is possible that the very nature of deep learning models makes such threats inevitable. Thus, we believe that these considerations should be taken into account as we increase usage and reliance on such AI models."

Similarly, the UK AI Safety Institute found that LLM safeguards can easily be bypassed[17] where "users were able to successfully break the LLM's safeguards immediately" using basic jailbreaking techniques, and "more sophisticated jailbreaking techniques took just a couple of hours and would be accessible to relatively low skilled actors".

Encouraging developers to implement safeguards in their AI models is therefore not robust enough a mitigation solution based on the inherent fallibility of such "after-the-event" patches to protect against the harms to individuals - they are closing the stable door after the horse has bolted. In any event, it is necessary 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.

## 6. Opting out of opt outs

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"[18] and Spawning have built a Do Not Train Registry which "consolidates machine-readable opt-out methods".[19]

We draw three matters to the EDPB'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

[13] https://cacm.acm.org/opinion/llms-data-control-path-insecurity/
[14] Matt Burgess, "The Hacking of ChatGPT Is Just Getting Started" (13 April 2023), https://www.wired.co.uk/article/chatgpt-jailbreak-generative-ai-hacking; Yuchen Yang, et al., "SneakyPrompt: Jailbreaking Text-to-image Generative Models" (10 November 2023), https://arxiv.org/abs/2305.12082; Rhiannon Williams, "Text-to-image AI models" (17 November 2023), https://www.technologyreview.com/2023/11/17/1083593/text-to-image-ai-models-can-be-tricked-into-generating-disturbing-images/; Jason Koebler, "Google researchers' attack prompts" (29 November 2023), https://www.404media.co/google-researchers-attack-convinces-chatgpt-to-reveal-its-training-data/
[15] Andy Zou, et al., "Universal and Transferable Adversarial Attacks" (20 December 2023), https://llm-attacks.org/
[16] Clint Rainey, "Computer scientists claim" (2 August 2023), https://www.fastcompany.com/90932325/chatgpt-jailbreak-prompt-research-cmu-llms
[17] AI Safety Institute, "AI Safety Institute approach to evaluations" (9 February 2024), https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations
[18] https://openai.com/index/approach-to-data-and-ai/
[19] https://spawning.ai/

therefore be more appropriate and in line with the data protection principles enshrined in the GDPR.

- The EDPB should be wary of the development of a confusing proliferation of approaches to 'opt-outs' and consequent ease with which they could be evaded.
- 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 protected by human rights law and deserve as much, if not more protection than intellectual property rights which are largely commercial assets.

## Final thoughts – who is responsible?

Governance of AI models is likely to be a key battleground for people's data and privacy rights. As the incentives to amass and process ever more data intensify, so do the risks for people's rights as enshrined in national, regional and international laws. The GDPR is technologically neutral, but its interpretation and guidance must keep pace with new and emerging developments.

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 reckless attitude to tech development that disrespects people's rights and believes that it is easier to ask for forgiveness than permission is unacceptable. 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.[20]

Finally, we urge the EDPB 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.

We remain at your disposal should you wish to seek clarification or further detail of any of the issues we raise in this letter.

Yours faithfully,

Privacy International

[20] 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/