April 12, 2024

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

Written Evidence Submission on the Purpose Limitation in the Generative AI Lifecycle

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 second chapter of its generative AI and data protection consultation series, focussing on the purpose limitation in the generative AI lifecycle.[1]

The Center for Data Innovation studies the intersection of data, technology, and public policy. Its mission is to formulate and promote pragmatic public policies designed to maximize 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 argue that the ICO should refine how it applies the purpose limitation principle to the development and deployment of generative AI models. The purpose limitation principle states that organizations should only collect and process data for specified, explicit, and legitimate purposes. However, the ICO should not attempt to apply this principle to general purpose generative AI models where, by definition, the various potential uses of these models are unknown. Instead, it should apply this principle to organisations that fine-tune models for deployment. Specifically, we make the following points:

1. The ICO should clarify its generative AI model lifecycle;
2. The ICO should clarify what is required of organisations in order to perform the compatibility assessment;
3. The ICO should reframe the purpose limitation principle to reflect the nature of generative AI model development; and
4. The ICO should only apply the purpose limitation principle at the deployment stage in the generative AI model lifecycle.

[1] "Generative AI second call for evidence: Purpose limitation in the generative AI lifecycle" Information Commissioner's Office

1. THE ICO SHOULD CLARIFY ITS GENERATIVE AI MODEL LIFECYCLE

1.1. The ICO puts forward a summary of the generative AI model lifecycle ("the lifecycle") in a simplified diagram, for which it bases its analysis of the purpose limitation principle. Unfortunately, the lifecycle is too simplistic to capture the extensive process of developing a generative AI model, in particular, how precisely each stage interacts with data protection concerns.

1.2. The ICO identifies four key stages in the lifecycle where data undergoes processing: data collection and curation for overall model training, data benchmarking for model pre-training, fine-tuning data for model adaptation, and user interaction data as a result of model deployment that feeds back into data collection. These stages fail to capture the other high-level processes that take place, and how developers might manipulate data.

1.3. For example, the ICO's four key stages do not account for data anonymisation to remove personally identifiable information (PII), data cleansing to handle null values, standardisation, manipulation to impute data where it is missing, and handling categorical variables. All of these processes occur prior to using data to train a model, at data collection and curation under the wider umbrella term of "data preprocessing".

1.4. The steps taken prior to training a model are therefore key to data protection analysis because they dictate the extent and type of data used for model training. Data anonymisation is one such technique that can have a significant impact on data protection. For example, Baffle is a company specialising in data compliance, offering techniques to prevent any PII leakage by removing it before model ingestion.[2] This drastically cuts down risks of data privacy exposure, limiting it to the input data fed post-deployment. The ICO should incorporate such techniques into its understanding of the lifecycle, and what state-of-the-art solutions current developers are using to mitigate risk at developmental level.

1.5. Similarly, the ICO conflates fine-tuning with model deployment, which is not reflected in the diagram or analysis. For example, the ICO includes both model pre-training and model adaptation within the model training phase of the lifecycle. It, however, goes on to state that "[A]fter the initial training of the generative AI model, an application is built based on it or a fine-tuned version of it, enabling its deployment in the real world. This means that one core model can give rise to many different applications." This analysis suggests there exists a stage where only model pre-training takes place to achieve a "core model." Applying this analysis, model adaptation, which involves training the model on fine-tuning data, would then only occur once an organisation using the model specifies the use case. This two-step process of model training indicates the possibility of two separate training

[2] "Preventing PII Leakage through Text Generation AI Systems" Min-Hank Ho, Baffle, December 7 2023

stages. The ICO should accurately articulate the lifecycle upon which it applies the purpose limitation principle.

1.6. There also appear to be discrepancies in diagrams between the first call for evidence covering the lawful basis for web scraping and the second call.[3] The ICO should clearly define what it understands to be the generative AI model lifecycle, including what it means by data curation, benchmarking, and fine-tuning, the explicit stages that take place in the lifecycle, and what data processing actually takes place at each stage.

2. THE ICO SHOULD CLARIFY WHAT IS REQUIRED OF ORGANISATIONS IN ORDER TO PERFORM THE COMPATIBILITY ASSESSMENT

2.1. The compatibility assessment is the process of determining whether data used for the original purpose of data processing can be reused without defining a new purpose for further processing.

2.2. The ICO recommends that where a developer has no direct relationship with a data subject, they should use public messaging campaigns and prominent privacy information in the compatibility assessment to increase awareness. The suggestion that generative AI developers with no direct relationship to data subjects, of which the majority, if not all, will fall into, use these methods is both unclear and unrealistic, and the ICO should specify what it expects of developers.

2.3. For example, it is unclear what public messaging campaigns or prominent privacy information entails, and to what extent is acceptable. OpenAI maintain a general FAQ on its privacy policies, such as one covering how ChatGPT and its other language models are developed, going into detail on where training data is sourced and how it is processed.[4] A similar statement is provided for the use of data post deployment.[5] It would be helpful for other generative AI developers if the ICO took a position on this as to whether that is sufficient for satisfying the compatibility assessment. Furthermore, the ICO should be careful not to restrict developers in such a fashion that they dedicate more time to public awareness than the development of safe and responsible generative AI. Doing so would set back the UK on the international stage and lead to a less competitive tech ecosystem.

2.4. Similarly, a different approach should be taken for the open-source community, who may lack the resources to engage in public messaging campaigns or devise extensive privacy

[3] "Generative AI first call for evidence: The lawful basis for web scraping to train generative AI models" Information Commissioner's Office

[4] "How ChatGPT and our language models are developed" OpenAI, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

[5] "How your data is used to improve model performance" OpenAI, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

information in using widely available pre-trained models. The ICO should, however, take note that a pre-trained generative AI model neither stores nor creates new datasets during pre-training or model adaptation.[6] Therefore, data processing activities in the open-source community are typically limited to fine-tuning data, and cases where human input post-deployment is used to continue training the model, actions of which would likely be captured by existing approaches to data protection compliance through privacy notices.

3. THE ICO SHOULD REFRAME THE PURPOSE LIMITATION PRINCIPLE TO REFLECT THE NATURE OF GENERATIVE AI MODEL DEVELOPMENT

3.1. The purpose limitation principle outlines that at the outset, the purposes of processing must be clear, said purposes must be recorded as part of the documentation obligations associated with data protection as well as specified in an organisation's privacy information, and that an organisation can only use personal data for a new purpose where it is either compatible with the original purpose, it obtains consent, or the new purpose complies with a clear obligation or function set out in law.

3.2. The principle applies easily when using private datasets with explicit consent, or public datasets where it has been made clear that they can be used for various stated purposes. This is because the source of the data can be identified, so compliance with the principle is straightforward.

3.3. The same cannot be said for web-scraped data, and the ICO should consider adapting the principle for the specific needs of generative AI model development. According to the ICO definition, web scraping involves the use of automated software to crawl web pages to gather, copy and/or extract information and store it for further use. Training a generative AI model requires vast amounts of data to offer a baseline from which to build upon, which is why web-scraping offers an attractive solution. For example, OpenAI's GPT-3 is trained on 300 billion tokens collected from a number of datasets, of which the largest portion is from Common Crawl, an open repository of web crawl data.[7] The ICO itself acknowledged that the least intrusive way to train a generative AI model is through web-scraping, offering the greatest access to the quantity of data needed for a model to function in a useful way.

3.4. Based on the quantity of data needed for training, the principle is ill-fit to apply to model development. As web-scraping is a largely automated process, there is no direct relationship between data subjects and the developers. It is, therefore, difficult to comply with the purpose limitation principle if a new purpose, such as training a model with a

[6] "Large language models (LLM)" Xabier Lareo, European Data Protection Supervisor
[7] "Open AI's GPT-3 Language Model: A Technical Overview" Chaun Li, Lamba, June 3 2020

different purpose on the same data, is incompatible with the original purpose for which the data was scraped and where consent is impossible to obtain.

3.5. In such a situation, at the first stage of the development process prior to deployment, the principle should assume consent from data subjects where said data was web-scraped in a lawful way (i.e., paying attention to any opt-outs signalled according to the Robots Exclusion Protocol, and only scraping from sources where public access is granted).[8] Given the rise in popularity of using generative AI models by the general public and the understanding that such models require extensive amounts of data, it is reasonable to assume that any data made publicly available on the Internet maintains a reasonable expectation of processing for this broad purpose of model training.

3.6. Taking this approach would not remove the requirements of the original principle at later stages in the development lifecycle, such as at deployment where a more specific purpose is known. However, it would alleviate burdens from organisations in trying to comply with requirements not fit for purpose and drafted without generative AI in mind. Additionally, this approach recognises there are privacy concerns associated with web scraping, in agreement with the ICO's previous joint statement on web scraping, but limits its application to where it will have greatest impact.[9]

3.7. In its joint statement, the ICO outlines several privacy concerns from using scraped data, including targeted cyberattacks, identity fraud, monitoring, profiling and surveilling individuals, unauthorised political or intelligence gathering purposes, and unwanted direct marketing or spam. The existence of a large language model of itself is unlikely to realise these privacy threats, instead behaving in a similar fashion to search engines that make use of web-scraped data to display search results. As discussed above, large language models do not store or create datasets from its training data, and developers employ techniques at the data pre-processing stage to remove PII, meaning it is impossible to have data leaks of PII. Therefore, the greatest risks to privacy arise not when the core model is trained, but when the model is fine-tuned and deployed for a specific use case. By applying the purpose limitation principle to deployment, this approach pinpoints privacy concerns to the use of data for the known use case, which is likely where privacy risks are most potent.

3.8. Therefore, the ICO should expand or reframe the principle to better fit this new form of data processing where in the majority of cases, data subjects are not known or are impossible to contact. This would likely mean that in cases where a large language model

[8] "In the Wake of Generative AI, Industry-Led Standards for Data Scraping are a Must" Morgan Stevens and Daniel Castro, Center for Data Innovation, September 1 2023
[9] "Join statement on data scraping and the protection of privacy", Information Commissioner Office, August 24 2023

is being trained, the ICO accepts the purpose of “training a large language model” as sufficient for establishing a core model, given the amount of data required to train it. Following this training period, the purpose limitation principle is engaged for use cases built on the core model that requires fine-tuning.

4. THE ICO SHOULD ONLY APPLY THE PURPOSE LIMITATION PRINCIPLE AT THE DEPLOYMENT STAGE IN THE GENERATIVE AI MODEL LIFECYCLE

4.1. The ICO highlights three scenarios for processing activities: a single organisation develops both a generative AI model and the application built on top of it, an organisation builds the model then provides a fine-tuned version to another organisation that builds an application on top of it and, an organisation develops the model then builds an application on it based on the specifications of another organisation.

4.2. Firstly, we believe the ICO’s assessment should include a fourth scenario that is also common practice, namely that an organisation develops a generative AI model, and another organisation or individual performs its own fine-tuning on the model before deploying it in an application. This type of practice is common within the open-source community where large language models are freely available to build on. Hugging Face is one such organisation, providing the infrastructure to build and deploy AI applications. GitHub also hosts a number of repositories exploring fine-tuning using openly available models.[10] The ICO should therefore expand its analysis to cover this practice of development.

4.3. Secondly, the relatively higher costs required to pre-train a generative AI model compared to fine-tuning an AI model means that significantly more developers will fine-tune AI models than pre-train AI models..[11] Because of this, most data protection concerns are likely to be directed after pre-training, when fine-tuning data is used to adapt a model. In its analysis, the ICO separates model training with model deployment. Given the demarcation made between initial model training or “pre-training”, and subsequent fine-tuning for deployment to a use case, we believe model adaptation falls within the model deployment phase, and that ICO should only apply the principle at this stage.

4.4. The purpose of fine-tuning is to take a broadly trained model and provide it with the subject matter expertise required for the specific use case. The fine-tuning data is typically a smaller subset of data with very specific data types relevant to that use case. It is most likely at this stage that a purpose is known. For example, a large language model trained on web-scraped data may be fine-tuned by an online retailer on a proprietary

[10] “Fine-tuning Mistral 7B using QLoRA” BrevDev, GitHub, https://github.com/brevdev/notebooks/blob/main/mistral-finetune.ipynb

[11] “What Large models Cost You – There Is No Free AI Lunch” Carig S. Smith, Forbes, Sep 8 2023

dataset, such as the online retailer's own customer chat logs. This practice can be seen in IBM's Watson Assistant, a chatbot for enterprises that builds on company content.[12] The application of the principle here aligns nicely with the goals of the principle, namely to give data subjects an understanding of how exactly their data is being processed. There is also a direct relationship between the retailer and its customers, which makes informing data subjects of their data processing much easier.

4.5. This approach would also complement the instance of one model fulfilling multiple purposes through different deployments. Rather than requiring a known purpose at the very outset of model development, it is only required at deployment, meaning a single model can satisfy the principle through different use cases.

4.6. The ICO should consider limiting application of the principle to this fine-tuning, deployment stage in the lifecycle, that would better reflect the practical data processing activities taking place in model development.

[12] "Build conversational AI chatbots infused with new generative AI capabilities" IBM