



May 1<sup>st</sup>, 2024

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

## Written Evidence Submission on the Accuracy of Training Data and Model Outputs

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 with respect to the third chapter of its generative AI and data protection consultation series, focussing on how the accuracy principle applies to the outputs of generative AI models and the impact that the accuracy of training data has on the output.<sup>1</sup>

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

The UK GDPR provides individuals the right to have data held about them by organisations kept up to date and as accurate as possible. We argue that whilst the level of statistical accuracy required for use cases relying on AI will vary, caveats exist which may limit the utility of accuracy as a performance metric, and further measures should be taken to support the integration of sub-perfect AI. We also argue that the ICO should pay due regard to the technical feasibility and, therefore, limitations of fulfilling data policy. Specifically, we make the following points:

1. The ICO is correct that the level of statistical accuracy required of model output depends on the use case of the model;
2. The ICO should recognise alternative metrics to statistical accuracy;
3. The ICO should refine what information it expects AI developers to disclose about model accuracy and reliability;
4. The ICO should adjust requirements for developers, deployers, and end-users that would not be technically feasible; and
5. The ICO should consider the impact of removing personally identifiable information on the accuracy principle.

---

<sup>1</sup> "Generative AI third call for evidence: accuracy of training data and model outputs" Information Commissioner's Office

### **THE ICO IS CORRECT THAT THE LEVEL OF STATISTICAL ACCURACY REQUIRED OF MODEL OUTPUT DEPENDS ON THE USE CASE OF THE MODEL**

We agree with the ICO's assessment that AI systems may not always produce output that is statistically accurate, and the acceptable level of statistical accuracy will depend on the specific intended use case of the system. For example, AI systems involving high-stakes decision-making, such as producing a performance review of an employee based on productivity metrics, should require a high degree of statistical accuracy (i.e., correctly reflect their performance) and data protection accuracy (i.e., contain correct information about the employee) because of the potential negative impact caused by an inaccurate output.

### **THE ICO SHOULD RECOGNISE ALTERNATIVE METRICS TO STATISTICAL ACCURACY**

There are, however, two caveats to this assessment. First, statistical accuracy is not always the best metric to apply in a given circumstance. Accuracy forms part of wider classification metrics that also include precision, recall, F1 scores, and Area Under ROC (Receiver Operating Characteristic) Curve. All of these can play a role in deciding whether an AI system is suitable for decision-making. Accuracy by itself is insufficient to determine how performant a model is, and indeed, a model's accuracy may not actually reflect its true performance. For example, Google runs a foundational course in machine learning that explains the nuances between classification metrics, as well as the pitfalls of considering accuracy alone. In its example, a model classifying tumours on its face is 91 percent accurate.<sup>2</sup> However, a deeper dive into this value shows that whilst it may correctly identify 90 tumours as benign, out of the 9 actual malignant tumours, only 1 was correctly diagnosed. In this case, the imbalanced nature of the dataset caused the stark number of false negatives and highlights that precision and recall would offer more suitable metrics for measuring performance on this basis.

Whereas statistical accuracy measures both true positives and true negatives amongst all outcomes, precision measures only true positives amongst both true positives and true negatives, focussing on the accuracy of those positive predictions. Recall by contrast looks at the impact of false negatives, which in the above example would flag the fact that only 1 out of 9 actual malignant tumours are correctly classified. An F1 score is the combination of precision and recall of a classification model, which, where there is an imbalanced dataset, offers greater insight into model reliability, rather than simply relying on the statistical accuracy metric, as the F1 score considers errors like false negatives and false positives.

What this indicates is that the utility of performance metrics will vary depending on the balance and makeup of input data, and the type of model used. Generative AI can perform classification tasks, but it is not limited in this regard. Therefore, the ICO should allow and encourage developers of generative AI systems and deployers of generative AI systems to use the most

---

<sup>2</sup> "Classification: Accuracy" Google, Machine Learning Foundational Course [Accessed April 29 2024]



appropriate metrics based on input data and intended model use rather than only relying on accuracy.

When considering performance metrics, the ICO should be careful not to require organisations to hold AI systems to a higher level of performance than humans. Human error can also lead to costly, and sometimes grave, mistakes.<sup>3</sup> To address potential mistakes, the ICO should promote safeguards around the integration of sub-perfect AI in a wider system, including human-in-the-loop oversight, clarity around the attribution of liability, and openness around the use of AI in decision-making.

### **THE ICO SHOULD REFINE WHAT INFORMATION IT EXPECTS AI DEVELOPERS TO DISCLOSE ABOUT MODEL ACCURACY AND RELIABILITY**

We agree with the ICO's suggestions that organisations with consumer-facing applications using generative AI clearly communicate about the general accuracy and reliability of their application's output. We also agree that developers of AI models should make this information available to its users. Providing this information will help consumers and other users make more informed decisions.

However, we disagree with the ICO's recommendation that deployers should provide clear information about the statistical accuracy of their application and its intended use. Given the above, making end users aware of statistical accuracy would do little to inform them of the limitations of model outputs and, thus, the reduction of AI harm. Instead, deployers should communicate to end users the safeguards they are putting in place and the impact of these safeguards on creating more reliable AI systems and AI-driven decision-making.

We also disagree with the ICO's recommendation that developers should "ensure the model is not used by people in a way which is inappropriate for the level of accuracy." Many developers, including those of open-source AI models, do not have this level of oversight or control over deployers, including when those deployers run the AI model on their own proprietary data. Those deployers may face their own requirements under the accuracy principle, depending on the use case, but they should be separate from those imposed on AI model developers.

### **THE ICO SHOULD ADJUST REQUIREMENTS FOR DEVELOPERS, DEPLOYERS, AND END-USERS THAT WOULD NOT BE TECHNICALLY FEASIBLE**

Research is underway to uncover how exactly a deep learning model relies on data inputs at each layer to make a decision. This research is termed "mechanistic interpretability" and seeks to go beyond the mathematics currently understood to reverse engineer an output so that people can achieve a deeper understanding of precisely how a model works.

---

<sup>3</sup> "The third-leading cause of death in US most doctors don't want you to know about" Ray Sipherd, CNBC.com, Feb 22 2018

Understanding the relationship between input and output data would go far towards compliance with existing regulatory regimes. The process to achieve mechanistic interpretability, however, is ongoing, and whilst a technically feasible, widespread methodology is being developed, efforts are better focused on mitigating risk associated with model outputs rather than how a model uses inputs to generate the output. The ICO should approach its consultation with this in mind, placing greater emphasis on actions that are presently technically feasible and typically within the control of the deployers of AI systems, rather than the developers themselves.

For example, the ICO expects developers to understand and document the impact that the accuracy of training data has on generative AI model outputs. This process would be significantly limited without first deploying the model to a use case, of which the deployers would have most oversight as to its impact. Similarly, there are limitations, as mentioned above, that would prevent developers from fully understanding the relationship between input data and model outputs, restricting their value in this process of understanding and documenting the impact of data on specific outputs in a use case.

The ICO also expects clear communication of certain steps to deployers and end users to ensure the lack of accuracy at the training stage does not result in negative impacts on individuals at deployment phase. This requirement blurs the lines between the role of developer, and the role of deployer, and whilst we agree that communication is necessary to improve the reliability of AI models, it is incorrect to place burdens on developers to communicate with end-users where there is no direct relationship. This would dramatically blur the lines in terms of both accountability and liability, ultimately making it more difficult for end users to seek redress where they suffer a negative consequence of generative AI.

Instead, the ICO should provide clear boundaries about how the roles of developer, deployer, and user interact, paying due regard to the relationships between each, and the technical capabilities available to them.

### **THE ICO SHOULD CONSIDER THE IMPACT OF REMOVING PERSONALLY IDENTIFIABLE INFORMATION ON THE ACCURACY PRINCIPLE**

The ICO should consider how the practice of removing personally identifiable information (PII) will impact the accuracy principle. Given that little is known about how precisely a model relies on data inputs to generate model outputs, the removal of PII should also diminish the application of the accuracy principle and in turn the right to rectification, which gives individuals the right to rectify any incorrect information an organisation may hold about them. This is because if PII is removed, an individual should have no concerns about inaccurate information being used as there would be no way of exposing the relationship between that information and the individual. This in turn makes redundant the right to rectification. We therefore encourage the ICO to clarify its understanding of the application of the accuracy principle where developers of generative AI models remove PII.