ICO consultation on Accuracy of training data and model outputs

The British Copyright Council (BCC) represents those who create, hold interests, or manage rights in literary, dramatic, musical, and artistic works. The following response has been developed with our membership which include professional associations, industry bodies and trade unions which collectively represent the voices of over 500,000 creators and performers, spanning the creative industries.

These rights holders include many individual freelancers, sole traders, and SMEs, as well as larger corporations within the creative and cultural industries. Our members also include collecting societies which represent rights holders, and which provide licensed access to works of creativity. Our members, individual creators or corporations have been using AI for over a decade as a tool assisting their creative activities or business administration.

We have commented in detail to the first and the second chapter of the ICO series of consultations on generative AI and data protection. We reiterate the main point asking for mutual respect from AI developers when processing personal data. From a policy perspective, only a level playing field for all parties of the value chain will enable the establishment of a successful market in which AI developers innovate and prosper in tandem with the creative sector and society overall.

The centrepiece of a fair market is compliance with the legal framework. For our members this means mainly copyright and related rights (including anti circumvention of technological protection measures), trademark, data protection, privacy, non-discrimination and contractual obligations.

In the same way as in the first two chapters of this series of consultations on generative AI, the legality of the acquisition of data and subsequent processing and the ability of individuals to exercise their data protection and privacy rights is a fundamental consideration. For the purposes of this response, we refer to "AI Developers" to include AI developers of training models, adapting models, deploying models and each separate organisation involved in the AI lifecycle. Personal data of BCC creators includes data such as names, likeness, voice, as well as potentially sensitive category data.

For more details please refer to the BCC submissions. In summary without express permission relating to every individual purpose of data processing, there is no lawful basis for AI developers to process personal data.

Input.

Accuracy initially relates to the input data. AI developers as data controllers and/ or data processors are responsible and accountable for maintaining the accuracy of the data ingested. This means that they are required to maintain accurate records of such data. Individual data subjects need to be in a position to assess whether the data which AI developers process in the machine learning stage continue to be accurate. The records of the data need to be auditable for the individual creator to check whether they have been processed on a lawful basis and that they continue to be accurate.

Therefore, we reiterate our concerns that transparency is required for the processes by which data is mined to allow for the source (and therefore the accuracy of the source) to be identified.

We are concerned that text and data mining in the ingestion process may lead users wrongly to rely on AI applications to provide factually accurate information – when in fact the application has allowed for the manipulation of the voice or image or other personal data about a creator/ artist to misappropriate, create unauthorised look and soundalike performances or to imply false endorsement and otherwise be entirely misleading.

The focus of advice to date issued by the ICO concerns protections needed for processing after access. We submit that the manipulation enabled by AI applications increases the need for greater transparency over the source of data accessed (particularly when the data links to copyright works).

This is support for AI developers when they are an "authorised" user and are required to provide information about the reliability of their output.

We expect data protection problems related to the correlation of factually accurate information and the origin of a human created work which, after processing during the deep learning through an AI application might relate inaccurate information about its original creator.

Output.

Accuracy is importantly recognised as relating to output as well. We generally agree with the ICO analysis on accuracy. However, we question how new transparency lines should be drawn when assessing the test of "whether the data needs to be accurate depends on the purpose of the processing".

In terms of personal data enshrined in data which are copyright protected works, legal processing must ensure that the data is not manipulated or changed in ways which lead to misinformation without sources being able to be checked and verified and relevant access permissions enabling validation alongside the other recommendations set out in the consultation.

Of these recommendations, our members particularly welcome labelling the outputs as generated by AI when issues of accuracy are relevant. We note that it is suggested this could be done for example by ”embedding metadata in the output or making imperceptible alterations to it to record its origin (sometimes referred to as “watermarking” and “data provenance”)”. Further discussion of practicalities, again linked to transparency is needed.

Overall appropriate and manageable systems for the labelling of content generated by AI as such is a key ask for the creative sector; a fair and competitive market requires a level playing field between AI generated and human created works and performances.

The protection of consumers as data subjects should generally be central in any consideration of data protection in the context of generative AI. Furthermore, consumers need to be able to base their decision for a product or service on an informed choice whether they acquire an AI generated or a human created work. Evidently, this is in addition to the need to protect the personal data of the creators whose works and performances are being copied to train the machine.

We agree also with the concluding advice by the ICO in particular that AI developers must, “where appropriate:

- seek and document the views of individuals whose data you will be processing during the AI lifecycle, or their representatives, unless there is a good reason not to;
- consult all relevant internal stakeholders;
- consult with your processor, if you use one; and consider seeking legal advice or other expertise.”

We stress once again that the AI lifecycle must include documentation of the access to data to be used to train the AI application and respect for the data owners of the data which is to be accessed.

We are looking forward to further engagement with the ICO.