Generative AI call for evidence: allocating controllership across the generative AI supply chain

September 2024

Overview

The British Copyright Council (BCC) represents those who create, perform, hold interests, or manage rights in literary, dramatic, musical, and artistic works. The following response has been developed with our membership which includes 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.

Many BCC members are creators who increasingly work with AI technologies as assistive tools linked to the works they create. On the other hand, many creators are extremely concerned with good reason, that AI-outputs are, and will be used without recognition of the personal data copied in data used for training new AI systems or permission from the human authors and the authors who make the arrangements for the creation of the works through the use of AI applications.

As such, transparency over how creative works and performances and the personal data associated with these works can be ingested and adapted throughout this process, particularly for works including personal data which are protected by copyright, will be increasingly important.

AI technologies are developing so rapidly that traditional silos for regulation applied to data protection on the one hand and application of copyright law on the other need to be questioned if uses of data and "content" which inform trusted AI applications are to be trusted and support true economic growth within the UK in the future.

Important "transparency" requirements can be accomplished by respecting and enhancing data protection law frameworks. IP licensing safeguards will remain vital to protect against the unfair use and devaluation of copyright protected work.

Situation Analysis

It is currently the case that creative works and the personal data associated with them (be it the likeness of a person depicted in a photograph, or metadata recording the name(s) or authors or contributors) are frequently being ingested for training generative AI applications without respect for contractual terms and conditions or technical protection measures, such as robots.txt, which should be respected in any crawling processes. Metadata is, in fact, of key importance in data ingestion as it allows the machine to match imagery with textual description, thus enabling it to create artistic works from text prompts.

Content and the underlying metadata are protected against copying and crawling through website terms and conditions as well as technical protection measures. Owners of the content and related databases rely on both mechanisms not only to protect their intellectual property, but also as part of their compliance with their obligations as data controllers. When the mechanisms are circumvented, as is the case in data mining for AI models, it is essential for regulation:

- to impose a robust transparency mechanism by which those responsible for crawling
- and demonstrate how they respect protections put in place by companies to protect against unauthorised processing of data held by them.

The transparency obligation should extend to copyrighted works copied for the purpose of mining personal data as the two – personal data and copyright – are for many of our members – inseparable.

While it is important to understand AI's impact in specific areas, regulating AI technology in silos would not address the all-encompassing nature of the technology which relies on simultaneous input on data protected by overlapping legal regimes (including intellectual property, personal information, commercial "personality" rights and online safety, to name but a few).

The robust protection of personal data would require a close re-examination of the current derogations from GDPR so as not exempt the use of indicia of personality in avatars created by generative AI systems merely because such use may be a result of an artistic, literary or journalistic activity. Given the wider societal risks of deep fakes, all uses of likeness, voice and other physical traits ought to require obtaining a specific and informed consent of the individual in question.

The latest call for views focuses on the allocation of roles and responsibilities in the generative AI supply chain. With advances in technology rapidly increasing the ways in

which data identifiers link to personal and sensitive data about individuals many of the systems applied for this increasingly overlap with the identifiers used across the creative industries to trace data and content which is also protected by copyright.

The consultation states that the roles and responsibilities under data protection law are not influenced by other legal regimes such as intellectual property law or competition law. However, the rapid development of AI models which require training on the use of data, including personal and sensitive data, means that such a siloed approach to regulatory responsibilities is becoming increasingly untenable.

Data Protection law already provides for controllers to explain why and how they process personal and sensitive data. It is generally recognised that the use of proprietary content is key for the training of general-purpose AI models, because the content is the most important source to generate training data in sufficient quantity and with sufficient quality.

This has been expressly recognised under the EU AI Act (Regulation (EU) 2024/1689) with providers of general-purpose AI models "having a particular role and responsibility along the AI value chain".

Within the UK the structures for reporting and governance applied are already relevant to the authorising the control of, and permitted processing of, personal data. This suggests that the ICO is well placed as a Regulator to have oversight of, or at least compliance duties linked to, providers of AI models publishing information in template form (which would explain the purposes of an intention to collect and process data and content to be used for training general-purpose AI models, in advance of being able to collect any such data).

If the ICO doesn't take on such a role, oversight of the important GDPR exceptions and caveats already recognised for permitted processing of data will become increasingly practically challenged.

In addition, the ICO would be well placed to ensure that general-purpose AI developers have to recognise that trade secrets cannot serve as a blanket justification for not disclosing information about data and content which they plan to use to train new general-purpose AI models. With respect to data which falls within the legitimate interests of copyright holders, a right holder's legitimate interest to know if its content has been used and if this use was lawful always prevails.

We therefore hope that the ICO will recognise that there is a vital transparency role to be played in supporting effective future development of general-purpose AI models within the UK. This transparency will help models to be transparently developed to allow for the trusted data and content required to be licensed and used effectively for the benefit of consumers and users generally.

Example of the need for transparency regulation

This has been recently exemplified by the unlicensed use of the “voice” of Scarlet Johansson by ChatGPT, who incidentally rejected a request for the use of her voice by ChatGPT.

This highlights genuine concern that technical means for data mining (particularly linked to AI generative training) mean that a person/business which is a regulated data controller, is increasingly facing a scenario whereby the data for which they are responsible as a “controller” is being copied without consent – and crucially without transparency over the way in which data may then be used by the “new” controller. Even in case of complete transparency of the data management, an individual creator cannot monitor the compliance with all data protection rules. The allocation of responsibility for data protection compliance should be for all data controllers involved.

Moreover, the extraction of creative works such as text, film, image and music from publicly available websites, even in cases where such practices are explicitly prohibited in their terms and conditions, compounds the issue even further. The fact that creative works may be publicly accessible online for a specific authorised use does not mean it is "publicly available" for scraping. The creator remains the original controller until, and unless an agreement is reached mutually to change this.

The solution: A level playing field

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 suggestion that data processing controls of one party can be overridden by another on the basis that data duplicated simply makes them “jointly controlled” does not work in such a scenario.

Transparency to provide the basis for commercial dialogue and permissions to be established for any accepted joint control must be enabled with the support of the Transparency template provisions proposed above.

The centrepiece of a fair market is compliance with the wider legal framework.

For our members this includes the data protection framework. This increasingly integrates with oversight of copyright and related rights (including any circumvention of legitimate technological protection measures), trademark, privacy, non-discrimination and contractual obligations.

Many of our members work to compile and produce new copyright protected works under existing regulatory structures which governs and supports editorially overseen, verified and trusted work which is vital to the economic success of the UK creative

sector. However, without mechanisms to know when general-purpose AI developers are looking to copy and process data linked to their catalogues, commercial structures to support a system of trusted and approved data in and therefore application of a trusted base in terms of outputs, misinformation and mistrust will result.

In the same way as in the first four 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. This remains crucial as we look to define allocation of roles and responsibilities in the generative AI supply chain.

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, directorial style as well as potentially sensitive category data.

We consequently agree with the ICO's restatement in early consultations that "as part of complying with the lawfulness principle of data protection, developers need to ensure their processing": is not in breach of any laws; and has a lawful basis under UK GDPR".

Transparency

In the absence of effective “transparency” requirements covering initial mining by a developer and the subsequent processing thereafter linked to distribution of a generative AI model (whether “open access” or “closed access”), there are real risk to the livelihood, safety and control of the individual whose personal data has been mined.

For example, the original (initially authorised) controller is currently often unable to identify how the information is ingested for use in outputs thereafter. The subsequent use by a processor and/ or AI model may then exacerbate possible harm to the individual by association with information or “services” or over which they have no control and no locus in terms of any direct relationship with the AI developer. This interaction and the roles and responsibilities must be transparent and by agreement.

The role of the controller

Due to the technological implications of AI development and the wholesale and opaque collection of personal data without notice to data subject, the active participants in the AI supply chain (which excludes holders of repositories data which is mined) are unlikely to be meeting their obligations as data controller under GDPR. This reinforces the unaccountability, lack of transparency and fairness as regards data subjects.

We are aware that the ICO has already issued papers summarising the "accountability and governance implications of AI". However, this addresses how companies should assess the risks posed to the processing of personal data by the [AI] processors.

However, addressing risks is clearly difficult if not impossible if the original/ true controller of data is entirely unaware of when/ and where the data for which they are responsible is being "taken"/processed by a third party. This is currently very often the case.

For the original controller to prepare and apply an appropriate risk assessment as envisaged by GDPR rules as per above, transparency and disclosure of DPIA's prepared and linked to AI LLM development must be an easily available as practical option for controllers before their data is mined or reproduced by the AI developer.

Means and purposes of distinct processing activities

Currently, if a controller is not able to know when their data is being mined, there is significant risk to them for "misuse" or processing of the data beyond originally permitted scope.

Despite some revised practices, questions remain over whether third party controllers who own "risk protection" mechanisms for example whether in the form of paywalls or robots.txt or other technical protection measures) are being observed.

This risks an original controller becoming an unwitting joint controller. This presents the further challenge around allocation of responsibility at this point.

It is the case then without greater enforceable transparency requirements linked to proposed use of mined data linked to AI developments, in advance of mining, original controllers will be unable to establish where their own protection and processing responsibilities begin and end.

Governance

We believe the Information Commissioner's Office (ICO) has a duty as a regulator to scrutinise the practices of generative AI application developers and that responsibilities cannot be side stepped by arguing that solutions can be found under different "siloes" of applicable law.

This is because of the almost universal and integral role that AI developments will play for the overall UK economy in the future.

The BCC is of the view that it is vital that the ICO takes on the responsibility of upholding data protection standards for data processing by general purpose AI developers which is extremely important to the creative industry (as well as society at large to protect

individual data subjects) particularly in the absence of a dedicated regulator safeguarding the interests of creators.

We agree that there should be a delineation of practices at the ingestion stage, when data including personal data undergoes processing. Compliance would mandate that developers ensure their processing activities are not only in compliance with pertinent laws, including copyright, but also is undertaken under valid lawful basis consistent with the UK data protection framework.

Our position

Only with the necessary transparency across the AI supply chain will those who accept responsibility as controllers of data when working to create specific and defined new copyright (i.e. the creators and authors of creative works) have any practical basis to understand how their "walled" and editorially overseen work ("Producer/Publisher Works) may be used by third party AI developers who are interested in scraping identifying or personal or sensitive data that is supposed to be under the control of the creator, producer or publisher.

Processor: Transparency mechanisms are needed to help legitimate data processors delineate practices for any processing of data linked to general-practice AI development and ingestion/training, including personal data undergoes processing.

Compliance would mandate that developers ensure their processing activities are not only in compliance with pertinent laws, including copyright, but also is undertaken for agreed and recognised data processing and retention purposes that are valid, lawful, and consistent with the UK data protection framework.

Controllers: Transparency over how and when their work is ingested and used is crucial. For example, the original controller to prepare and apply an appropriate risk assessment, transparency and disclosure of DPIA's prepared and linked to AI LLM development must be an easily available and practical option for controllers before their data is mined or reproduced by the AI developer.

"Joint controllership of data" along the lines which the Consultation appears to recommend as a "solution to complexity" must exist within a level playing field and true transparency as the foundation to ensure the basis of future use of the works and caveats are recognised by both sides.

A lead in transparency requirements under oversight from the ICO must be an important part any future governance arrangement.

Conclusion

AI developers are routinely processing personal data which is meant to be controlled by a third party without express permission from that third party.

This risks the third-party controller being (unwittingly) in breach of data protection law requirements and leads to infringement of both data protection laws and (in many cases) infringement of copyright which exists in the data and works being copied by an AI developer.

Steps need to be taken to practically support the principle that, without the express permission relating to every individual purpose of data processing, there is no lawful basis for AI developers to process or store any personal data.

AI developers are ultimately responsible for ensuring transparency and diligent record keeping of ingested materials as well as ensuring data protection principles are considered from the outset to protect personal data and respect existing UK copyright and data protection law frameworks.

In order for effective regulatory oversight to be re-established, urgent steps are needed to require Transparency Templates to be completed and follow up commercial discussions to be enabled linked to the future development of general-purpose AI models. This will include copyright permissions and licensing structures for the benefit of the wider UK economy.

We therefore hope that the ICO will recognise that there is a vital transparency role to be played in supporting effective future development of general-purpose AI models within the UK which are transparently developed to allow for the trusted data and content required to be licensed and use effectively for the benefit of consumers and users generally.