OpenUK Response to ICO Consultation on AI and Generative Models due 1 March 2024
https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-series-on-generative-ai-and-data-protection/

1. Nature of Open Source

In the call for evidence, you refer to an "Open Source" approach as follows:

> *"If copies or extensive details (eg model weights, starting code, etc) of the underlying generative AI models are made available by the initial developer to third parties, developers are expected to have much less control over how the model will be used downstream. In these cases (sometimes referred to as an 'open-source' approach), customers typically run their own instance of the generative AI model."*

Understanding what is meant by "open source" and an "open source approach" both in terms of the component pieces of AI and the shades of openness which are apparent in the AI market place is a critical issue in this landscape. We are currently seeing global regulators and policy makers struggle with this. Open source software has a long established Open Source Definition and the Open Source Initiative is the guardian of this definition and approves open source licences. Key to these are the lack of licensing restrictions around who can use the code and its purpose enabling a free flow. This is a far cry from for example the Llama community licence with its commercial restrictions and Acceptable Use Policy. Open data also has long established licences although there is no definition that is widely used.

However there are frequent references to Open Source which do not meet the definition for software and which are also intended to cover data, i.e. it is being used unclearly as a generic term. As each level of openness has different impacts and risks it is critical that these are understood whatever the use case or naming and that clarity is created in these meanings. This has also been noted in the House of Lords report on the LLM Enquiry published 2 February.

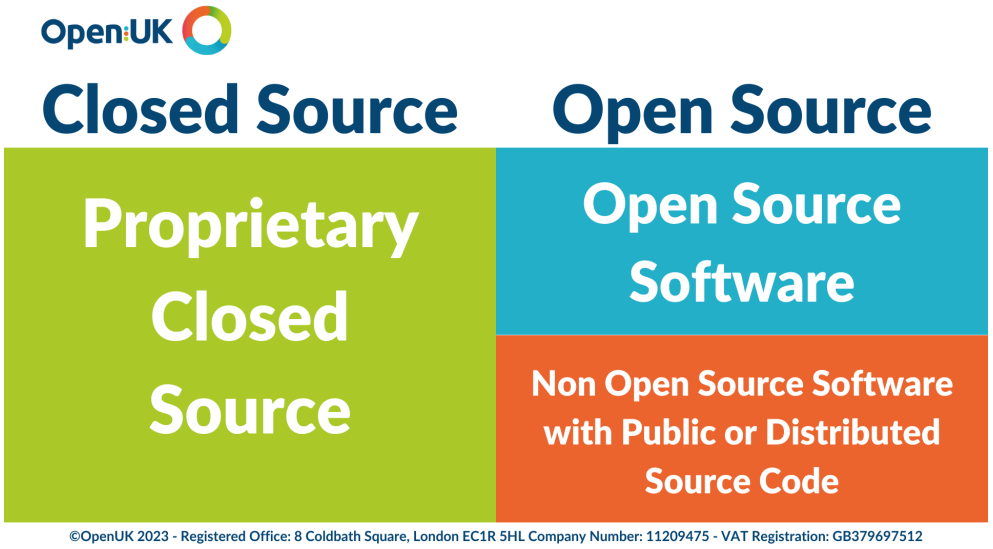
The paragraph in the call for evidence merely describes the provision of the software and data to third parties. The use of an "open source approach" is in fact much broader than any current definition of Open Source (and indeed could include models licensed on proprietary terms) but also fails to match a generic usage of the term as a catch all.

The ICO's usage of the term in fact specifically confuses 'Open Source Software ', with 'Source Available Software' or 'Public Source Software' which are terms used where software has source code shared but does not meet the standard of the Open Source Definition, generally as there is some restriction as to users or the use cases for the code, which would not meet definitions 5 and 6 of the Open Source Definition.

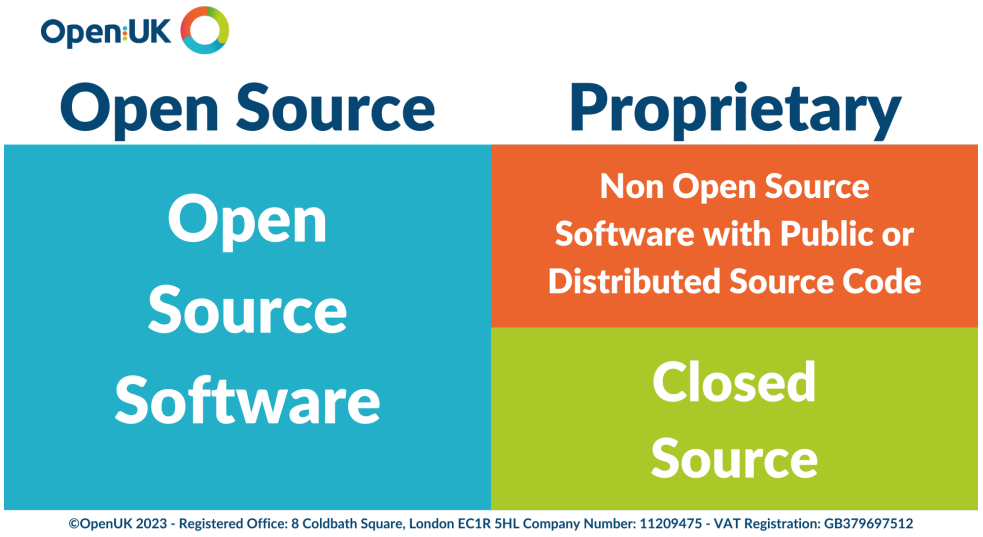
These are in fact one of the other shades of openness which generally means that there are restrictions in the licensing which may have a commercial impact and certainly alter the risk

profile and fails to distinguish that open source is the provision of software on licence terms which satisfy the Open Source Definition.

The conflation of the two comes from a common misconception that the opposite of Open Source is closed source, which would utilise this model, which appears to be the one being adopted by regulators but which would be viewed as generally incorrect by the Open Source Community. This mis categorisation is frequently referred to as “Open Washing” and when the licences used for Public or Distributed Source are categorised in this way they are referred to as “fauxpen”.



The way that Open Source Software has been considered and largely accepted within Open Source Development follows this approach and this has a very different impact.



What this demonstrates very clearly is the key issue in language and understanding around openness in AI and the very different impacts that misunderstanding around the terminology can bring. Also it highlights the difference between any kind of openness that meets the Open Source Definiton's key terms that anyone can use the code (in that case) for any purpose, versus other shades of openness that are more restrictive and the impact of such restrictions.

We recommend that the ICO takes steps to create clear definitions around open source, and other levels of openness in software and the non software and model categories, eg data, to which the term is currently mis-applied. We also recommend that these align with other UK policy plans across DSIT, the Home Office, CMA and other regulators.

To avoid confusion OpenUK has referred to Open Innovation in AI and the House of Lords LLM Report refers to Open Access AI.

Additionally we consider that the ICO statement fails to grasp the impact of transparency in building trust and control which can be achieved through open models or the risks of closed AI components.

Without the transparency of openness there is no clarity on the AI components and models being used or the data upon which they are trained. This may therefore lead to discrimination and the perpetuation of false and misleading information and to untrackable privacy and confidentiality breaches. There is also a risk of single points of failure and lock-in from closed AI.

2. ICO's conclusions in the call for evidence are anti-Open Source

It is inconsistent with Open Source licensing to put in place downstream controls on use, as freedom of 'field of endeavour' is a core tenet of Open Source Software licensing and forms Definition 6 of the Open Source Definition.

Indeed it is impossible to impose such controls in an Open Source Software context due to the free flow previously explained and the fact that such controls would destroy that free flow.

This fundamental freedom is something that millions of users rely upon. Bearing in mind the scale of open source adoption in infrastructure and enterprise, reported in 2023 as 96% of software stacks using it and 76% (according to Synopsis 2023 study) of these a regulatory attempt to restrict this would be hugely problematic.

Therefore, the approaches to risk mitigation set out in the call for evidence as being necessary to support lawfulness under legitimate interest following a balancing test, being essentially methods of controlling downstream use, are incompatible with Open Source Software licensing across the globe.

The net effect of this would be to render any Open Source Software licensing to be unlawful, and will be a serious disadvantage to innovation and the development of AI in the UK and sit contrary to the UK's Open Source First public sector policy, established over a decade ago as the first in the world.

Indeed, this may be seen as asymmetrically punishing those who want to work openly and collaboratively giving others the benefit of their code.

Open development of AI is in the public interest

The societal benefits from AI are clear and the societal benefits of openness in AI are also clear, from the democratisation of technology and enabling innovation through to the trust that can be created via transparency.

AI development that is open - whatever the shade of openness - is in our view clearly in the public interest and in line with the stated policy aims of the UK Government to retain and enhance its current position in the development and business of AI and to ensure that regulation does not adversely impact innovation.

AI development that is open has the additional societal benefit of increasing the number and diversity of AI developers in the UK, supporting inclusiveness, innovation, and national competitiveness which has been essential to the UK's current position in AI.

Open source development increases the rate of innovation and increases transparency, allowing for greater scrutiny and robust development.

This is particularly important to the democratisation of technology and ensuring that our AI future does not end up in the hands of a few companies.

Web scraping to support AI development is in the public interest

Web scraping has become necessary for the functioning and open internet today and were it necessary to consider whether a web page contained any personal data, the internet would not function effectively today.

Opportunity to mitigate risk to data subjects prior to scraping

Web scraped data is obtained from publicly available sources, made available by platforms or publishers that have sourced the material and have responsibility for establishing lawful basis for such publication. When data is publicly available and accessible online, a deliberate choice has been made to share that information with the public domain. It is the responsibility of those publishers to put appropriate measures in place to ensure purpose limitation, and it is reasonable for those accessing that data (in accordance with the terms on which it is published) to be able to rely on the publisher having taken such steps. It is always open to the publisher/platform to include a prohibition on web scraping in its terms of use, and include technical measures to prevent scraping. Web publishers often employ technical means to protect their content and maintain control over how it is accessed and

used, and have done so for many years. One common method is the utilisation of a robots.txt file, a text file placed on a website's server that instructs web crawlers and scrapers which pages or sections of the site they are allowed to access. By configuring the robots.txt file, web publishers can specify rules for web crawlers, such as allowing access to certain parts of the site while restricting access to others. While robots.txt can be an effective tool for controlling access to content, it relies on the cooperation of web crawlers to adhere to the preferences outlined in the file.

Web publishers may employ further technical means to more robustly control access to their data by putting content behind a paywall, requiring users to provide login credentials or pay a fee to access premium content. To provide further protection against unauthorised access by circumventing through techniques such as credential sharing or automated account creation, web publishers often employ further technical measures, such as rate limiting, CAPTCHA challenges, and user-agent detection.

Risk occurs at point of use

Risk in relation to personal data processing when it comes to AI is less about personal data included in a training dataset (which is generally not recorded within the model and not reliably reproducible at the output), and more relevant to the data provided to a trained AI at the point of inference (i.e., the point of end use).

Rather than imposing impossible requirements on an open source developer, and seeking to control risks at the point of development - which are inherently unknowable - we believe it is more appropriate to apply regulatory control at the point of use. A business user is best placed to assess the risk and has chosen the environment in which it is to be used. They are ultimately and have historically been the point of risk and bear any associated liability.

The business user must exercise an appropriate level of discernment in their business decision making including their choice of digital and technology tools. Where these include open source software, our view is that the business user ought to be responsible for the curation - the technical hygiene and good governance of the code used in their environment and products.

The responsibility for ensuring that the use and behaviour of the AI model at point of use ought also in our view, again as with any technology tool to sit with the user, who will make judgements about the data to be processed, and the purpose and means for such processing. They are also subject to regulation appropriate to the chosen use case.

It is statistically possible for an AI model to ‘invent’ personal data, even if it wasn’t trained on any. Therefore even where an end user does not introduce any personal data, we consider that the user is still responsible for using the AI model in a lawful way.

Any such personal data accidentally produced by a model, whether trained on personal data or not, would constitute incidental processing which the user would be best placed to safeguard.

Web scraping and Open Source licensing can be lawful and satisfy the three-part test for legitimate/public interest

Lawfulness

In relation to areas of law other than data protection in the UK, web scraping is considered to be lawful assuming it respects intellectual property rights and contractual terms. We see lawfulness other than in a purely data protection context as out of scope of this call for evidence and therefore, beyond a statement that any processing must not breach any other laws. We believe the ICO's guidance should be limited to lawfulness under data protection law.

The purpose of the processing is legitimate, and the processing is necessary:

In order to develop AI models it is necessary to train those models on publicly available data which includes personal data. By aggregating and organising this data, web scraping permits the development of AI models by developers who otherwise would not have access to the scale of datasets required. Not allowing developers to use web scraping in the development of AI would have the effect of limiting AI development to large enterprises that already have such datasets available or have the resources to procure them.

The substantial public interest should be taken into account, and reinforces the fact that any adverse impact on data subjects as a consequence of the processing would need to be very material for it to outweigh the overriding public interest.

The data subjects' interests do not override the interest being pursued:

The data being scraped is already in the public domain, and its use in training AI is unlikely to result in an adverse impact on data subjects - particularly where personal data is not contained in the model or carried through into the outputs.

Legitimate interest may not be the best answer

Image generation GenAI is quite capable of producing images of individuals that it has seen frequently enough, and presumably this would constitute Special Category Personal Data, for which legitimate interest would not apply.

Downstream controls are not the only option

We believe there is a more nuanced approach to risk mitigation, whereby the risk to data subjects can be reduced using alternative approaches and not merely by downstream controls. For example, if the AI system can be shown not to 'carry through' personal data, or

the inclusion of personal data in the training dataset is purely incidental or not a material part of the training data, then the risk to data subjects will be minimal.

Conclusion

As identified in the call for evidence, the risk to data subjects needs to be balanced against the interests of the controller and the public interest. We consider it inappropriate to require the developer to be required to impose downstream controls, which would have a fundamental chilling effect on any development of AI that is open source or open innovation in the UK. We urge the ICO to consider what controls are to be imposed and where these might be imposed in a more proportionate way.

OpenUK

OpenUK is the UK organisation for the business of open technology with a purpose of UK leadership and global collaboration in open technology - open source software, open hardware and open data and includes standards and AI. It convenes the conversations on the key issues of the day in open technology and operates on 3 pillars: Community, Legal and Policy and Learning.

OpenUK is a not-for-profit company limited by guarantee registered in England number 11209475, VAT Registration: GB379697512, registered office at 8 Coldbath Square, London EC1R 5HL It is not pay to play organisation and is funded primarily by enterprise sponsorship and donation.

https://openuk.uk/

Contact:

CEO and Chief Policy Officer, Amanda Brock, amanda.brock@openuk.uk

Chief Legal Officer, Christopher Eastham, christopher.eastham@openuk.uk