The Dialogue™
INFORM ENGAGE IDEATE

The Dialogue's Comments to the U.K.'s ICOs Call for Evidence: The lawful basis for web scraping to train generative AI models

MARCH 6TH, 2024

Written Comments

The Dialogue's Comments to the U.K.'s ICOs Call for Evidence: The lawful basis for web scraping to train generative AI models

Authors:

Vaishnavi Sharma
Pranav Bhaskar Tiwari
Kamesh Shekar

The Dialogue™ is a public policy think tank with a vision to drive a progressive narrative in India's policy discourse. Founded in 2017, we believe in facilitating well-researched policy debates at various levels to help develop a more informed citizenry, on areas around technology and development issues. The Dialogue™ has been ranked as the world's Top 10 think tanks to watch out for, by the Think Tank and Civil Societies Programme (TTCSP), University of Pennsylvania in their 2020 and 2021 rankings.

For More Information, Visit:
https://thedialogue.co

Introduction

The Dialogue recently responded to the U.K.s Call for Evidence on lawful basis for web scraping to train generative AI models.[1]

Reproduced below are our observations regarding the same; we have reproduced two substantive comments from the Call for Evidence.

1. One of the questions we wish to address in this document pertained to whether we, as an organisation, agree with the analysis presented in the Call for Evidence?

2. One of the other questions we wish to address in this document pertained to whether we, as an organisation, agree with the analysis presented in the Call for Evidence, specifically regarding whether the legitimate interests test could be met if technical and organisational measures to limit the use of the Gen AI model are in place.

---

[1] U.K. Information Commissioner's Office, Generative AI first call for evidence: The lawful basis for web scraping to train generative AI models, https://ico.org.uk/GenAILawfulBasis.

1. One of the questions we wish to address in this document pertained to whether we, as an organisation, agree with the analysis presented in the Call for Evidence?

At the outset, we see issues with implementing and operationalising the three-part test to generative AI, especially concerning issues where the training must be based on large datasets.

Since generative AI has the potential to have aggravated and disproportionate impacts, it is necessary for digital and fundamental rights to be (i) clearly stated and accorded to individuals and (ii) adequate grievance redressal mechanisms to be established and appropriately communicated to the public at large. Further, there is a need for further clarification on who can provide these rights within the generative AI supply chain since not all players would be suitably situated to do so.

Our past work as a basis for answering this Call

We address the Call for Evidence with relevant research of our own, where we conducted primary research and engaged with relevant AI stakeholders.

- Research Paper: Principles for Enabling Responsible AI Innovations in India: An Ecosystem Approach.[2]
- Research Paper: Towards Trustworthy AI: Sectoral Guidelines for Responsible Adoption.[3]
- Open Loop Programme.[4]

1.1. On the three-part test

We are from India; the "three-part test" is also referred to as the "proportionality test" in India - as fleshed out by the Indian Supreme Court (nine-judge bench) in its landmark case of *K. Puttaswamy v. Union of India* in 2017. While the proportionality test has always existed in some sense in Indian jurisprudence (for instance, certain prongs of the test), the test has started to become more concretised in recent years.[5] For instance, an argument can be made that the "reasonable" restrictions under Article 19(1) of the Indian Constitution may not mandate a proportionality test (even though recent jurisprudence builds on "reasonableness" to enumerate the proportionality test).

Since the proportionality test (the "three-part test", as the Call mentions) is evolving and becoming more concretised across jurisdictions, not all jurisdictions have the (a) same/similar definitions of the proportionality test and (b) some principles of the test may normatively carry more weight as opposed to others. This essentially means that (a) not all stakeholders are mandated to follow the same or (b) stakeholders only summarily address their compliance with the three-part test. Since data controllers will be the principal entities carrying out the three-part test under Article 6(1), a written test and

---

[2] Shekar, K., Sahiba, J., Saxena, G. and Birla, B. (2023, September). Principles for Enabling Responsible AI Innovations in India: An Ecosystem Approach. The Dialogue [hereinafter "Responsible AI Research Paper"].

[3] Vedashree, R., Sahiba, J., Agarwal, B. & Shekar, K. (2024, February). Towards Trustworthy AI: Sectoral Guidelines for Responsible Adoption. The Dialogue.

[4] The Dialogue and Open Loop, Policy Prototyping Project to Implement the AI Principles of Human Centricity, https://thedialogue.co/policy-prototyping-project-to-implement-the-ai-principles-of-human-centricity/.

[5] Aparna Chandra, *Proportionality in India: A Bridge to Nowhere?*, 3(2) University of Oxford Human Rights Hub Journal (2020).

elaborate guidance on the same is necessary. Further, it would also enable regulators to make an argument for case-by-case assessments in generative AI uses.

1.2.   Our observations w.r.t. the Call

A.   Evidence and identify a valid and clear interest

- With multiple soft-touch approaches to AI regulations, there is a lack of consistency in representational practises for evidence purposes while gathering and processing data.
- While it is difficult to adhere with absolute certainty to certain data practices like (data minimisation, data de-identification, purpose limitation, and data retention practises), the three-part test may compel stakeholders to state their data and privacy protection practices, even with caveats. Such a mandate may allow users to understand the full scope of the sorts of data and privacy protection safeguards the controller, and third-party entities are following in the data information cycle.
- While stakeholders, including individuals, are concerned about the opaqueness of generative AI decisions, they are also concerned about dense explanations given for AI decisions that may not be comprehensible to the public. Any explanation as regards the legitimate aims must be accessible to the audiences, especially keeping in mind the context of the direct and indirect audience. In contexts such as Indians, such discourse on the accessibility of legitimate aims and valid interests becomes critical. To operationalise the principle of transparency and explainability at the stage of collecting and processing data in the AI lifecycle, it is important to ensure that the processes and methodologies used to gather and process data are clear, understandable, and well-documented. AI Developers should document the data collection methods, such as the sources, sampling techniques, any potential biases or limitations associated with the data, and data processing techniques, including data cleaning, filtering, and feature selection processes. Furthermore, developers should offer explanations of the decision-making processes and the factors influencing the outcomes, thereby enhancing the understandability of the AI system. We have explored this in our "Responsible AI" research paper.[6]

B.   Balancing test - *particularly* when they *do not or cannot* exercise meaningful control over the use of the model.

- Stakeholders (esp. developers) do not have clear guidance on evolving jurisprudence on digital and fundamental rights; this may translate into carrying out activities without appropriate safeguards, disproportionately impacting individuals living at the margin. This poses three key issues: (i) they lack the expertise and resources to identify and assess the human rights implicated in the deployment of their product. (ii) they do not have a definite and/or consistent stance on how to proceed about balancing the rights of the individuals at risk vis-a-vis any potential benefits. (iii) there is a lack of a concrete feedback loop.
- In our conversations with AI start-ups, we have found that some principles can complement some principles, and some may contradict them (like data minimisation may contradict the objective of reducing discrimination and bias, especially while operationalising an AI solution).[7]

---

[6] *See* Responsible AI Research Paper.
[7] *Id.*

A three-part test may require data and privacy protection mandates to be followed across data information lifecycles.

Further, every principle may mean or carry varying importance at different stages of the AI lifecycle. For instance, the human-in-the-loop principle during the planning and design stage requires you to consult with the relevant stakeholders, and while deploying the same, human oversight may be required to undo/redo/reassess certain decisions the AI has made after the planning and design stage and maybe most involved. While human intervention is required at both stages, the designated individuals have varying degrees of control over the data. Consequently, there needs to be more clarity in understanding the degree of rights and interests at stake at each stage of the lifecycle in the supply chain. A balancing test at each stage and with each specific entity involved would give a clearer picture of the potential harms.[8]

C. Demonstrate how the interest they have identified will be realised and how the risks to individuals will be meaningfully mitigated, including their access to their information rights.

The legitimate interests test would mandate the entities to undertake a structured risk classification and management framework based on the systemic risks the specific entity within the supply chain poses, a much more evolved and appropriate framework to identify risks, as opposed to solely relying on industry practices on data minimisation, purpose limitation, consent and notice, etc.

Such a framework is already used in other sectors that use sensitive personal information, like the financial domain, where entities must adhere to a risk management framework (risk identification, risk measurement and assessment, risk mitigation, risk monitoring, etc.).

Furthermore, risks can occur at any stage of the processing in the supply chain; any harm may have a systemic effect on the entire supply chain. Certain entities in the supply chain may have different risk perspectives at different stages of the lifecycle, whereby the entity that designed the product envisions different risks and harms at the design phase than the entity which deploys it.[9]

Accordingly, risk management is important for generative AI as it would assess the likelihood and severity of harms and materially inform in identifying mitigation strategies, ultimately reducing the risks. It would also provide for a more structured approach and do away with ad-hoc practices, and it would ultimately aid in identifying reasonably foreseeable harms in separate stages of processing data.

---

[8] *Id.*
[9] *Id.*

1.3. Our remarks specifically with regards to training generative AI using copyrighted material

Fair Use Analysis in Generative AI Training Using Copyrighted Data

1. Recommendation:

Acknowledge the transformative use of copyrighted material in Generative AI training as a legitimate exercise of the fair use doctrine. This perspective supports innovation while upholding the integrity of intellectual property rights.

The AI solution could give out copyrighted content as a response, however, it is important to model a display of disclaimers and source information. Irrespective of this, if users of the tech are misusing it despite the disclaimer, then appropriate measures are to be taken. Also, it is important to bring out the difference between the foundational model and generative AI. The same entity may not develop these two, so appropriate measures have to be suggested to different players.

2. Analysis

The application and training of generative AI within the bounds of copyright law can be aptly described under the fair use exception, employing a method known as non-expressive copying. Non-expressive copying refers to the use of data not for its original, creative content but for its functional value in training AI models. This approach diverges significantly from traditional forms of replication, which often focus on reproducing the expressive elements of copyrighted material. Legal precedents such as *Sega v. Accolade*[10] and *Kelly v. Arriba*[11] have ruled on the legitimacy of this transformative approach, highlighting its crucial role in the advancement of technology.

The nature of copyrighted work, often an amalgamation of factual reporting and creative expression, necessitates a sophisticated interpretation. This interpretation leans more towards the functional use of data rather than its original, expressive intent. Despite the extensive use of data by generative AI, these models are distinct in that they do not store specific content. Instead, they assimilate and learn from overarching patterns, setting them apart from direct forms of copying.

While significant, the potential market impact of generative AI is generally consistent with the principles of fair use. This is exemplified in landmark cases like *Google v. Oracle*,[12] where the Court found that while Google did copy Java API code, this act differed from traditional copying as it was aimed at creating a new, transformative platform, and not for expressing the same content, thus qualifying as a fair use exception in copyright law. As generative AI continues to evolve, it becomes increasingly important for models to cite their sources, particularly when relying on copyrighted data from web sources. This practice enhances transparency and aligns with ethical standards, ensuring respect for the original creators and their economic rights. By balancing innovation with intellectual property protection, this approach promotes a progressive and equitable landscape in the realm of copyright law and AI development.

---

[10] Sega Enterprises Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992).
[11] Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003).
[12] Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183 (2021). *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258 (2023).

2. The other question we wish to address in this document pertained to whether we, as an organisation, agree with the analysis presented in the Call for Evidence, specifically regarding whether the legitimate interests test could be met if technical and organisational measures to limit the use of the Gen AI model are in place.

We have reservations as regards this observation. It would be more appropriate to move beyond the three-part test and instead look at generating an environment for generative AI whereby all the entities engaged in the ecosystem act responsibly (and not use the three-part test as a litmus test). Regulators must operationalise various market mechanisms, apart from regulating AI, to build healthy relationships and cooperation with AI developers and AI deployers with a limited disposal capacity.[13]

---

[13] *See* Responsible AI Research Paper.