CITRIS AND THE BANATAO INSTITUTE | CITRIS POLICY LAB

UNIVERSITY OF CALIFORNIA | BERKELEY DAVIS MERCED SANTA CRUZ

May 28, 2023

Director General Roberto Viola
Directorate-General for Communications Networks, Content and Technology European Commission
1049 Bruxelles/Brussel, Belgium

Re: Delegated Regulation on data access provided for in the Digital Services Act

Roberto Viola and DG Connect:

Thank you for soliciting feedback on Article 40 of the Digital Services Act (DSA).

We are researchers and practitioners affiliated with UC Berkeley, with expertise in platform research and development, safety, security, policy, and ethics.

While Article 40 of the DSA holds great promise to improve independent researchers' access to platform data for public interest research and oversight purposes, we have identified limitations that should be addressed. In addressing these limitations, the DSA will more likely be able to achieve its goal of identifying and mitigating systemic risks.

We offer the following submission for your consideration.

Thank you,

Brandie Nonnecke, PhD
Assoc. Research Professor, Goldman School of Public Policy; Director, CITRIS Policy Lab & Our Better Web, UC Berkeley*
nonnecke@berkeley.edu

Jonathan Stray
Senior Scientist
Center for Human-compatible Artificial Intelligence (CHAI), UC Berkeley*

Luca Belli, PhD
Founder, Sator Labs;
Tech Policy Fellow, UC Berkeley*

Jared Lewis, MPP
Tech Policy Fellow, CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley
Head of Policy, dentsu Good*

***** *Affiliation for identification purposes only. The views expressed are those of the authors and not of their institutional affiliations.***

Executive Summary

In order for the Digital Services Act (DSA) to achieve its goal to facilitate research that identifies and addresses *new* "systemic risks" on very large online platforms and very large online search engines, the following actions should be taken:

1. **An independent, intermediary body should be established to assist the DSCs in vetting researchers and research projects.** The DSCs may lack the expertise and capacity to adequately review researchers and research projects in a timely manner. Thus, the independent body can provide robust peer review while streamlining and quickening the review process for the DSCs and assist in ensuring compliance with relevant laws and regulations (e.g., GDPR). The European Digital Media Observatory (EDMO) has launched a working group to establish such a body.

2. **Establish requirements for high-value API access.** The DSA currently prioritizes deductive rather than inductive research by requiring researchers to define a research question/hypothesis and data needed *a priori*. This process prioritizes deductive research that affirms a known systemic risk is present rather than facilitating inductive research that explores data and/or conducts experiments to identify unknown or *new* systemic risks. A high-value API that contains data relevant to identifying systemic risks should be made available to vetted researchers.

3. **Support platform/auditor collaborations to conduct experiments.** Article 37, Independent Audit, should be interpreted to include a right for external auditors to perform systemic-risk relevant interventional experiments (e.g., A/B tests). Many policy questions around systemic risks are fundamentally causal questions (i.e., counterfactual questions) and require close collaboration between auditors and platform data scientists and developers, and must be subject to human experimentation ethics, privacy considerations, and the possibility of business disruption.

4. **Enable remote query execution.** Rather than supporting development of "clean rooms", which may require travel to a secure location, vetted researchers should be able to develop a query that returns only aggregate, privacy protecting results. The query would be submitted to platform data scientists for review and execution and would be overseen by the independent body (see #1 above).

Comments in Response to Questions Posed

1. Data access needs:
    a. What types of data, metadata, data governance documentation and other information about data and how it is used can be useful to DSC's for the purpose of monitoring and assessing compliance and for vetted researchers for conducting research related to systemic risks and mitigation measures?
        i. The current process outlined in the DSA prioritizes deductive rather than inductive research. Because researchers must specify a research proposal (i.e., research question / hypothesis) in advance *and* the data needed, the research supported through the DSA primarily affirms that a known systemic risk is present. It does not enable identification of *new* systemic risks through exploratory analysis. Deductive research involves posing a research question/hypothesis and gathering data. Inductive research involves more of an exploratory analysis of data, identifying *new* systemic risks that were previously unknown. In order to support inductive research, API access is critical. The API should provide access to a variety of data that are relevant to identifying "systemic risks." We recommend that an independent entity be established to provide guidance on what types of data should be made available by each VLOP via the API.[1]
    b. What sort of analysis and research might DSC's and vetted researchers conduct for the purposes of monitoring and assessing compliance and conducting research related to systemic risks and mitigation measures?
        i. Many policy questions around systemic risks are fundamentally causal questions (i.e., counterfactual questions). Thus, "what is the effect of a current recommender system on mental health" is logically equivalent to the question "how would mental health outcomes change if the recommender design were different?" It is well understood in the scientific community that answering such questions accurately requires experiments, not just observational analyses of platform data. Otherwise, it is not possible to distinguish between, for example, "recommenders are making people depressed" and "depressed people use recommenders more." We believe Article 37, Independent Audit, should be interpreted to include a right for external auditors to perform systemic-risk relevant interventional experiments (e.g., A/B tests). This is a complex proposition; and one which will require close collaboration between auditors and platform data scientists and developers, and must be subject to human experimentation ethics, privacy considerations, and the possibility of business disruption. However, there is no other reliable way to answer basic causal questions

[1] Nonnecke, Brandie and Camille Carlton. 2022. "EU and US legislation seek to open up digital platform data." Science 375, no. 6581 (February): 610-612. DOI: 10.1126/science.abl8537

such as: How does social media use contribute to polarization? How do different features of a recommender system manipulate users' behaviors?

2. Data access application and procedure:
    a. Digital Services Coordinators (DSCs) in the Member States will play a key role in assessing researchers' applications and they will act as intermediaries with the platforms. How should the application process be designed in practice? How can the vetting process ensure efficient exchanges between researchers and platform providers?
        i. Under the status quo, platforms are likely to be involved in some parts of vetting researchers and research projects (DSA 40.12). This will call into question researcher and research independence. To address this, an independent, intermediary body should be established to assist the DSCs in vetting researchers and research projects. The DSCs may lack the expertise and capacity to adequately review researchers and research projects. Thus, the independent body can provide robust peer review while streamlining and quickening the review process for the DSCs. The European Digital Media Observatory (EDMO) has launched a working group to establish such a body.
    b. Article 40(8) exhaustively defines criteria for vetting researchers. How can a consistent assessment across DSCs be ensured, while still taking into consideration the specificities of each request?
        i. The independent body mentioned previously can serve this function. Vetting of researchers and research projects should be done through peer review.
            1. Based on our previous experience we believe that designing a productive research project based on platform data is typically a much more extensive and challenging proposition than currently anticipated, likely requiring close consultation with platform data scientists as to availability and interpretation of existing data, as well as the detailed operation of current systems, strategies for compliance with existing privacy regulations, etc. The formal request and response process in DSA 40.4-6 will be a cumbersome method to arrive at a feasible and impactful research design. Therefore, we expect most researcher data requests, at least initially, will represent poor quality research designs. To prevent good research from getting stalled behind poorly conceived requests, and to streamline the review process, proposals should first be prioritized by timeliness and scientific impact. For example, research proposals seeking to identify systemic risks to a near-term election should go through expedited review, while proposals which are underdeveloped, do not contain specific queries to be executed

on platform data that is known to exist, only marginally relevant to systemic risks, or methodologically unsound should not delay high-impact work.

ii. Regulatory effectiveness may be strengthened if the independent body and DSCs can vet research institutions as opposed to individual researchers.

1. Under this framework the institutions might be authorized to conduct research under specific research domains of systemic risks (i.e., Environment, Equity, and Disinformation).
2. A network of research institutions might be established and be required to adhere to a range of intellectual property protections that may include: physical data storage, prohibition of exporting data outside of network storage facilities, collaborative research that supports institution focus areas.
3. This approach also encourages institutions to pose research questions and incentivizes researchers to support these questions in research.
4. Regulating institutional research networks also supports the aims to identify systemic risks and supports efforts in continuous regulatory oversight.

c. What additional provisions or specifications could be useful to help balance the new data access rights and the protection of users' and business' rights, e.g. related to data protection, confidential information, including trade secrets, and security?

i. Delivering data to researchers directly will always be risky. Few academic institutions are equipped to defend against a determined adversary, or to vet everyone who has access (recall that it was an academic who provided Facebook data to Cambridge Analytica.) Some have proposed "clean room" access where data cannot be copied or removed, but this is cumbersome, possibly requiring physical travel to a secure location. Instead, we believe the default approach should be remote query execution. That is, a vetted researcher would develop a query that returns only aggregate, privacy protecting results then submit this query, as source code, for review and execution by platform data scientists. This approach has several advantages:

1. It reduces the need for communication with and involvement of platform data scientists, since query development can take place on synthetic data
2. It allows for complex aggregate analysis (e.g., sophisticated content classifiers, longitudinal individual-level outcomes)
3. It provides a single point of precise privacy, confidentiality, and security review, i.e. the submitted query code

            4. It completely removes the need to transmit unaggregated data to researchers

    d. What kind of safeguards can be put in place to assure that data gathered under Article 40 is used for the purposes envisaged and to minimize the risk of abuses?
        i. The independent body detailed above should also provide guidance and oversight over independent researchers' and research institutions' responsible access, use, and storage of data. These processes should be vetted with the DSCs to ensure compliance with relevant laws and regulations (e.g., GDPR). The independent body should work with researchers and research institutions to document data access, use, and storage and provide summary reports to the DSCs.
    e. Article 40(13) introduces the possibility of an independent advisory mechanism to support the management of data access requests and vetting of researchers. What would be the added value of such a mechanism?
        i. We strongly encourage the establishment of an independent body (see comments above). In establishing such a body, the following will occur:
            1. Increase speed and validity of researcher/research institution and research proposal review
            2. Reduce redundancies in applications as one non-EU researcher could partner with several EU-based researchers and submit their proposal to several DSCs. By having a single entity to assist in reviewing all proposals, duplications/redundancies will be identified.
            3. Reduces redundancies of same research question(s) and data requests. There is potential for several researchers to propose similar proposals (e.g., political ads placed before an election). By having a single independent body, these proposals can be identified and researchers may be encouraged to share data and/or collaborate.

3. Data access formats and involvement of researchers:
    a. What technical specifications could be considered for data access interfaces, which takes into account security, data protection, ease of use, accessibility, and responsiveness (e.g., APIs, data vaults and other machine-readable data exchange formats)?
        i. Public APIs should contain only material that is already widely public (ala article 40(12) and data aggregated at a coarse enough level of detail (temporally, and across users, regions, demographics etc.) to protect privacy
        ii. A high-value API that provides more data relevant to investigating "systemic risks" should be made available to vetted researchers/institutions on an ongoing basis

iii. External queries and analyses to be “thrown over the wall” provide a more general, powerful, and privacy-preserving mechanism than either API access or clean room access. This process should be facilitated and encouraged. Platforms would have to audit researcher submitted code to ensure the returned results are sufficiently aggregated so as not to infringe user privacy.

iv. Platforms should consider publishing synthetic datasets which match the schema and perhaps macro statistical properties of internal datasets.[2] These would be useful for two purposes:

1. to allow researchers to design and code data queries and analysis, which can later be executed internally on the real data (“over the wall”).
2. to do basic descriptive statistical analysis, where the synthetic data is designed to reproduce particular features of the distribution of real data..

b. What capacity building measures could be considered for the research community to take advantage of the opportunities provided by Article 40?

i. The independent body detailed above can provide guidance on data security mechanisms, data storage, appropriate analysis of data, and more.

ii. The independent body, in collaboration with the DSCs, can also provide researchers and platforms with much-needed clarification of GDPR compliance. For example, if a researcher gains access to data and then a user requests deletion, is the data still compliant?

c. Would it be desirable and feasible to establish a common and precise language for DSCs, vetted researchers, VLOPs and VLOSEs to use when communicating about data access, e.g. by formulating a standard data dictionary and/or business glossary? How might this be implemented?

i. The independent body can assist in developing a standard data dictionary and/or business glossary. However, this has some limitations. The benefits of defining a cross-platform data dictionary has its limitations because platforms vary greatly in their internal data structures (e.g., what elements and reactions constitute a “post”). While some standardization of terms can be operationalized across VLOPs, all VLOPs exhibit different data characteristics and this should be represented in any dictionary/glossary.

4. Access to publicly available data:

a. Not only vetted researchers will have greater opportunities for accessing data, all researchers meeting the conditions set out in Article 40(12) will be able to get

[2] Shiffman, N. (April 26, 2023). “Tools for platform research: Lessons from the medical research industry.” *Tech Policy Press*, https://techpolicy.press/tools-for-platform-research-lessons-from-the-medical-research-industry/

direct access to publicly available data. What processes and mechanisms could be put in place to facilitate this access in your view?

i. There should be an API that is widely available to researchers, including “unvetted” researchers. In recent years, VLOPs have become more restrictive in public API access. There should be a public API that provides basic data for free to researchers.
ii. Vetted researchers should be able to gain access to an API that provides more data relevant to systemic risks than the publicly available API.