/ Position by AlgorithmWatch

Call for Evidence: Data access rules must empower researchers where platforms won't

May 2023

Long before AlgorithmWatch was forced by Facebook to shut down our Instagram monitoring project in 2021, we urgently advocated for a DSA which empowers watchdogs working in the public interest with the legal means to investigate systemic risks stemming from online platforms—risks like negative effects on electoral processes, on public debates, or on public health. That is why **we welcome Article 40 of the DSA as a potentially groundbreaking framework** which promises to enable vetted public interest researchers to access the data of Very Large Online Platforms and Very Large Online Search Engines with more than 45 million active users in the EU.

Platforms' past and present actions make clear that they cannot be relied upon to voluntarily share data with independent researchers in a consistent or meaningful way. This is evidenced, for example, by the failures of Social Science One, the adversarial actions of Facebook against good-faith research by the NYU Ad Observatory and AlgorithmWatch, as well as Twitter's recent move to impose prohibitive fees on its API for studying publicly available data. An analysis of major platforms' first implementation reports under the EU's revamped Code of Practice on Disinformation further shows they have done little to "empower the research community" despite formal commitments made in June 2022.

The DSA's data access rules will thus require robust enforcement to ensure that platforms are compelled to share meaningful data with vetted researchers meeting the appropriate criteria and employing appropriate safeguards to protect user data. The precursor to this will be a strong framework to ensure adequate standards for data sharing—both in terms of platforms' data transparency and access infrastructures as well as in the vetting of researchers and research proposals. The European Commission's delegated act on Article 40 of the DSA is an important opportunity to clarify such technical and procedural standards to ensure that public interest researchers are indeed empowered to scrutinize real and potential risks which follow from the design, functioning, and use of platforms' algorithmic systems.

Policy recommendations

Our response to the Call for Evidence centers on six main areas to help inform a strengthened data access framework:

1) Empowering a broad base of vetted researchers, 2) Governance documentation as data, 3) Vetting exemption requests, 4) Defining and ensuring reliable access to publicly available data, 5) Protecting independent research from platform abuse, and 6) Independent advisory mechanisms to support data access applications & vetting.

We conclude by calling for an inclusive and iterative process to further develop data access structures and guidelines going forward.

1) Empowering a broad base of vetted researchers

Article 40 presents a tremendous opportunity to empower researchers from academia and civil society to access platform data, so long as applicants fulfill a set of necessary criteria to demonstrate they are indeed working in the public interest within the DSA framework and employing necessary data security safeguards. As discussed in our sixth recommendation, an independent, third-party intermediary body could provide meaningful guidance to Digital Services Coordinators tasked with vetting such researchers, and help ensure that vetting and data sharing is done in a timely fashion.

A forthcoming Delegated Act should further clarify conditions under which a consortium of researchers may gain access to platform data, which may include independent journalists and researchers not based in the EU. Such consortia could, for example, include partnerships between an EU-based civil society organization and independent journalists, or comprise both European and non-European-based academics. Broadly speaking, these vetting procedures must be robust in order to prevent abuse from bad-faith actors, yet not so burdensome so as to deter researchers, especially from civil society, from asserting their rights under Article 40.

2) Can we see the emails? Governance documentation as data

Article 40(4) of the DSA clarifies that purposes for vetted research include not only helping to detect, identify, and understand systemic risks in the EU, but to assess the "adequacy, efficiency, and impacts of the risk mitigation measures" taken by platforms in accordance with their DSA obligations. In order for researchers to fulfill these research imperatives, **it is crucial that Article 40 not be interpreted too narrowly in terms of the types of data**

which can be made available to vetted researchers. Researchers should be permitted to gain access to a broad range of internal and procedural documentation in order to more comprehensively assess how consequential decisions on risk assessment and mitigation were reached, and whether these decisions were indeed adequate, efficient, or impactful.

Until now, the public has only had indirect access to such documentation thanks to whistleblowers like Frances Haugen, who leaked internal Facebook documents revealing that, time and again, Facebook researchers identified risks stemming from the platform that were ignored or downplayed by company leadership—for example, that Facebook's revamped newsfeed algorithm was amplifying misinformation and violent content, or that high-profile users were being given carte blanche to violate community standards.

Under the DSA framework, platform accountability should not have to rely on whistleblowers to expose such (mis)conduct. Regulators, auditors, and independent researchers should have access to governance documentation which may include (without being limited to) reports from internal reviews, results of A/B testing, user surveys and feedback, internal presentations, memos and emails, so long as such disclosures are made with full respect for data protection requirements.

Access to such documentation would help researchers to scrutinize, for example, whether companies indeed undertook proper risk assessments prior to executing design choices, how potential tradeoffs between business and safety were accounted for in risk mitigation, and the speed or manner in which risks flagged internally were responded to.

3) Data access exemptions need clear (and strict) boundaries

Platforms will no doubt resist complying with a whole range of data access requests on the grounds that providing such data would expose "confidential information" or "trade secrets". **We expect regulators to not allow platforms to abuse these exemptions to routinely deny data access requests, as this would severely undermine the central aims of Article 40.**

In order to achieve this, it is incumbent on regulators with privileged access to platform data to investigate the full range of available data, including governance documentation, and to set appropriately strict boundaries around exemptions from data sharing. These boundaries should, among other things, be guided by existing law (e.g. data protection)—however, data protection must not serve as an automatic excuse for platforms to deny access requests. **Platforms should rather be required to demonstrate that the provision of particular data would force them to violate other legal requirements when invoking exemptions for sharing data.**

Recital 64 in the DSA clarifies that the law must be interpreted such that platforms do not unduly invoke "trade secrets" as an excuse to deny access to data to vetted researchers. We assert that the strict vetting criteria that researchers must meet to gain access to platform data (independence from commercial interests, transparent funding, high data security requirements, etc.) ensure that researchers handle legitimately sensitive information responsibly. Just as researchers will be strictly vetted, so should platforms' attempts to amend data access requests be subject to a strict vetting process.

4) Defining and ensuring reliable access to “publicly available data”

Despite their imperfections, Facebook’s CrowdTangle and Twitter’s public API were once helpful in making publicly available platform data accessible to a broad range of researchers, civil society watchdogs, and journalists. The current disordered state of these APIs shows how quickly things can change for the worse in a world of self-regulation, as unreliable access to data for public interest researchers has seriously hindered systematic research on platforms and their impacts on society.

The largest online platforms and search engines are obligated under DSA Article 40(12) to develop better systems for sharing publicly accessible data, and they have formally committed to doing so under the EU’s Code of Practice on Disinformation. It is nevertheless clear that they are taking steps in the opposite direction, moving to restrict the sharing of publicly available data or “delivering” it on unworkable terms. This unfortunate status quo highlights the need to establish clear standards for these APIs in terms of the data they make accessible, how and when access is provisioned, and to whom.

AlgorithmWatch is part of a group of civil society experts actively discussing the practical implementation of the DSA’s public data sharing scheme. This cohort combines technical expertise and years of experience in platform monitoring, data protection, and human rights. **Together, we put forward a set of five recommendations to the European Commission and to the designated platforms directly as they move to implement Article 40(12) of the DSA:**

1. Public data should be complete, comprehensive, and include historical data
2. Data must be verifiable, for which multiple access methods are needed
3. Permissioned access must come on fair and reasonable terms
4. Platforms must not hinder independent public interest research
5. Data sharing should include a diversity of researchers

The full recommendations can be read in our open letter.

5) Scraping in the public interest isn't a crime — but obstructing it should be

Among our recommendations for implementing Article 40(12) is that platforms must not hinder independent, public interest research. In addition to the need for reliable data sharing, and given the challenges and potential loopholes referred to above, we believe **independent data collection methods that use scraping have been—and will continue to be—essential for public scrutiny of platforms' algorithmic systems.**

Such data collection may be carried out with the help of sock-puppet accounts, crawlers, or data donations from real-world volunteers using browser plugins, for example.[1] These methods allows researchers to "scrape" the platform and assemble their own datasets based on direct observation rather than relying on data which is provided by the platform.

Independent audits in this vein have a proven track record in driving public awareness toward algorithmic risks. In our Instagram Monitoring project, for example, we collected data donations which showed that Instagram was prioritizing far-right political content in Germany in the run-up to the German elections. In a different study, the non-profit AI Forensics (formerly Tracking Exposed) used sock-puppet accounts to show that TikTok was promoting wartime content that was supposedly banned to users in Russia.

Such independent audits alone cannot capture the full story of how algorithmic recommender systems work (they may, for example, be limited to publicly available data or suffer from selection bias). Yet they also have certain advantages to direct data access by enabling more exploratory research, and by illuminating what users actually "see" on the platform.

While some data scraping activities are tolerated by platforms, we have also seen platforms weaponize their Terms of Service to stop this kind of research by simply blocking relevant plugins or activities through technical measures or legal threats. Especially the latter can have a chilling effect on researchers trying to hold them accountable. **The DSA's data access framework must therefore ensure that researchers conducting privacy-compliant research in the public interest are better protected from platform power.** DSA Article 40(12) should be understood as providing a safe harbour for independent research which addresses systemic risks using publicly available data, and should at a minimum require non-interference into public interest, GDPR-compliant research efforts in line with the rest of the provision. In cases where platforms suspect an independent research project violates the GDPR, for example, the burden of proof should fall on the platform to substantiate such claims.

[1] For an overview of auditing methodologies for Very Large Online Platforms and Search Engines, see work from Ada Lovelace Institute and Stiftung Neue Verantwortung.

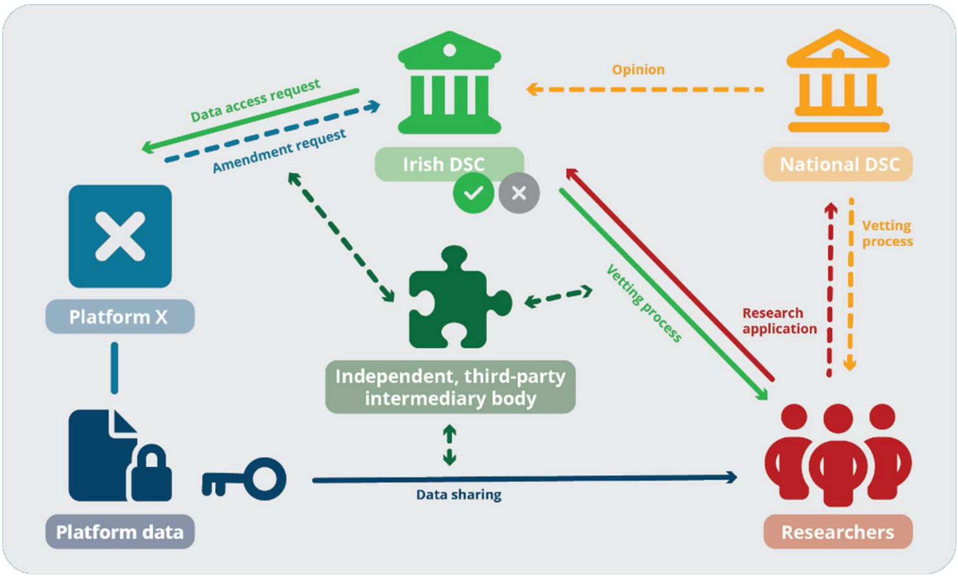
6) An independent, third-party intermediary body to facilitate platform-to-researcher data sharing

To help overcome platforms' strong incentives to oppose, avoid and obstruct transparency demands, **we believe an independent institution is essential to serve as an intermediary in the DSA's research access regime.** Such an independent intermediary is provided for in Article 40(13) of the DSA as an "independent advisory mechanism" and reiterated in Commitment 27 of the Code of Practice on Disinformation, in which major platforms commit to "developing, funding, and cooperating with an independent, third-party body that can vet researchers and research proposals."

Our basis for recommending the creation of such an intermediary body follows from AlgorithmWatch's Governing Platforms project, which draws on best practices from other sectors to operationalize an effective research access framework in platform governance. **In our report, we suggest that this independent institution (or institutions) be regarded as a "transparency facilitator" and empowered to serve a variety of important functions,** such as reviewing the quality of reported data by periodically auditing disclosing parties; pre-processing and pseudonymizing data (in a transparent way) before making it accessible to researchers; and maintaining relevant access infrastructures, such as public databases, virtual machines, and discussion fora.

A similar set of functions for such an intermediary is detailed in the European Digital Media Observatory (EDMO)'s report on researcher access to platform data, which affirms our assertion that platforms may share data with independent researchers in a way that protects users' rights in compliance with the General Data Protection Regulation (GDPR), and would be strongly supported by such an independent intermediary body.

In AlgorithmWatch's user guide to the DSA's data access rules, we visualized the blueprint for Article 40 to provide a basic overview of the vetting process.



The above illustration shows how important such an independent intermediary body could be to the overall data access regime—particularly in the DSA's early stages, given the complex vetting task facing some national regulators who may have limited capacities, resources, or expertise in this domain. In order to be effective, however, **this intermediary body must be equipped with a relevant and clear mandate within the overall DSA framework, adequate resources, sufficient expertise, and genuine independence.** It is essential that platforms' commitment to fund such an intermediary under the Code of Practice on Disinformation does not impede on its independence in any way for the sake of research integrity.

Moving forward: an inclusive and iterative process to further develop data access structures and guidelines

There are still many open questions with regard to implementing the DSA's ambitious data access framework. We acknowledge the European Commission for its work on a Delegated Act that seeks to bridge this gap and clarify, with inputs from the research community, standards that would help ensure platforms actually produce useful data—and that this data is made accessible to vetted public interest researchers in a privacy-protecting manner.

In order to put Article 40 into practice, however, and to accomplish its central aim of enabling greater public scrutiny of platform risks, the inputs of the research community must not be limited to the consultation on the delegated act. We advocate for the institutionalization of an inclusive, iterative process at various levels, including regular exchanges between researchers from academia and civil society with DSCs and with the European Commission. To that end, it is crucial that regulators, intermediaries, and platform providers offer proactive support for researchers to ensure they have the resources and tools necessary to actually conduct research.

Finally, platforms must be compelled to fulfill their commitments to the research community under the Code of Practice on Disinformation—commitments which will be enforceable under the DSA once it formally becomes a Code of Conduct—and to provide far more meaningful reporting of their progress which actually accounts for feedback from affected researchers and may be independently verified. This includes commitments to provide vetted researchers with reliable access to public data; to help establish and cooperate with an independent, third-party body that can vet researchers and research proposals; and to support good faith research into Disinformation that involves their services.