Briefing on data access in the DSA

On 13 March 2024, 16:30-18:00 CET, Women in AI Austria hosted a panel discussion with Jennifer Adams (Comms Policy Collaborative at University of Vienna), Susanne Lackner (KommAustria), Thomas Lohninger (epicenter.works and EDRi), and Rania Wazir (leiwand.ai and Women in AI Austria), moderated by Ana María Jaramillo (Complexity Science Hub Vienna).[1] Natalie Ségur-Cabanac provided an introduction to the provisions of the DSA to allow the audience to follow the contributions. During the discussion, the panellists shared their perspective on and experiences with data access in the context of the DSA, the importance of data access for informing regulatory responses, and why collaboration is key to making data access an effective tool for monitoring the effects of algorithmic systems.

In this briefing, we provide a short recap of some of the themes discussed during the event as well as a background note to situate the discussion.

What you will find in this briefing

[1] https://www.linkedin.com/events/accessdenied-challengesontheque7166356628814872576/

Contributions by

Jennifer Adams, Valerie Hafez, Ana María Jaramillo, Susanne Lackner, Thomas Lohninger, Natalie Ségur-Cabanac and Rania Wazir

Discussion

With our post-event briefings, we like to make transparent where the ideas contained in this text come from. Jennifer's interventions are formatted *green*, Susanne's comments are *purple*, Thomas' quotes are *orange* and Rania's contributions are *blue*.

One key element of the data access provisions in the DSA is that they only apply to VLOPs/VLOSEs. This restriction is based on the necessity of proportionality – because these online platforms are so large, *these obligations are manageable for them*. In contrast, the data access rights offered by the DSA *are not limited to researchers in the EU*,[2] which may help bring in external expertise both on the effects of algorithmic systems and on the effects of the DSA. Finally, the data access provisions of the DSA relate not only to risks, but also to *risk management measures,* and research on these measures *will help to understand their effectiveness*. And even though the data access obligations have not entered into force yet as the delegated act has thus far not been published, preparatory work is underway to ensure that once the legal basis arrives, templates are available and it is clear which documents need to be provided to support the application for data access.

As noted during the discussion, data access for studying algorithmic systems is closely related to the question of control over data. *Without the DSA, only VLOPs/VLOSEs themselves have data* – alongside *a few hand-picked researchers* – to study the effects of algorithmic systems. Ideally, data access should be *democratic* because *data on social media platforms is a public good*, and the *DSCs support very broad access*, as in their view, data access plays a key role in the governance model of the DSA. Through its monitoring provisions, the DSA also – for the very first time – *provides civil society and academia with a voice* and a counterweight to business interests in the context of digital platforms. For this reason alone, *publicly available data should remain accessible*. Therefore, a legal obligation (Art. 40 (12) DSA) was introduced to ensure accessibility. As we could see from the audience questions, also regulatory and supervisory authorities have been experiencing difficulties in requesting access to data to conduct their monitoring activities, which greatly limits the possibilities for evidence-based policy making.

However, data is not simply of interchangeable quality. If you set out to study the effects of algorithmic bias, for instance, *this is only possible with access to sensitive data*. Data access for conducting research of public interest is therefore necessary *but obviously there are privacy issues as well* which need to be dealt with carefully. At a later stage, however, it is often the terms of use of VLOPs/VLOSEs which prohibit sharing data curated for research purposes; this practice in fact *undermines the validity of the scientific process*, as it renders basic verification by peers impossible. Additionally, there is anecdotal evidence that the enclosure of data which the DSA seeks to prevent is carried forward by other means, such as intimidation via liability forms or the prohibition of scraping. This capture of data is in part explained by the rise of LLMs, which require large amounts of interactional data; therefore, the owners (i.e. the

[2] This assessment is shared by some DSCs, but it is yet to be determined whether the European Commission shares this view.

VLOPs/VLOSEs hosting the data) try to reduce its accessibility for public purposes while reaping the benefits for their own products. The old question of digital rights – essentially, *how do you use public data while protecting private data* – therefore has to be rethought again for LLMs, preferably in a democratic manner. Likewise, the issues of GDPR as a basis for processing data need to be reconsidered when the data used is both private data and a public good, as pointed out by an attendee.

In addition to the legal concerns around data access, there are concerns around the technical means to actually make use of this access. Before GDPR came into force, *specialised tools existed as open source and freely available solutions* to help researchers scrape and handle online data for research purposes. With the restriction of access, these tools fell into disuse, which makes it difficult for researchers to be independent from the API access offered by platforms. As a member of the audience pointed out, there are very few incentives in academia to develop and maintain such tools over time; there are *some initiatives to counteract this with public funding for free and open source software projects and the communities that make them possible*,[3] but they are few and far between. The API access in turn is highly volatile, with *limitations to download size and number of downloads as well as constant changes to APIs* endangering the research efforts throughout the project. Crucially, this type of knowledge is *not common for regulators*, who do not have the required expertise either from social sciences or technically. This is why DSCs and proto-DSCs are in the process of both recruiting experts such as data scientists and are aiming at establishing a structured dialogue with academia and civil society.

Systemic risks such as hate speech, disinformation and effects on elections are under investigation by researchers from academia and civil society alike, who *hold up mirrors to platforms* on their neglect of public care. Independent research therefore helps to challenge narratives about issues, for instance in the case of hate speech, where a study conducted by Facebook found the prevalence of hate speech to amount to ca. 0.01%,[4] but a subsequent study by Amnesty International set the prevalence to ca. 1%.[5] This is due to Facebook measuring the prevalence of hate speech based on overall content, whereas Amnesty International selected particular political topics, as the prevalence of hate speech significantly increases on issues related to gender, poverty, Covid, rights to citizenship or rights to protest. In addition, Facebook published its analysis as a percentage of the overall viewed content, while Amnesty International expressed its findings in the percentage of available content. These fine differences are important because the chilling effect of hate speech can affect particular groups, such as women, more than others. *Sometimes we censor, sometimes we self-censor, and sometimes there are mechanisms which enable censoring – and civil society is key to monitoring these*.

In addition, the speed at which society works enacted by the engagement architectures of

[3] https://www.sovereigntechfund.de/
[4] https://transparency.fb.com/reports/community-standards-enforcement/hate-speech/facebook/
[5] https://www.amnesty.it/campagne/contrasto-allhate-speech-online/

VLOPs/VLOSEs has consequences not only for how society becomes aligned with algorithmic time, but also for the knowing-spaces which can effectively monitor and call attention to developments. For this reason, it is paramount to *open the playing field and see what algorithms are doing* but also to not *set the bar so low* as to settle for anything better than "another Myanmar".[6] In this context, *interdisciplinarity is key* and also *collaboration between researchers, civil society and regulators* to *work together to get a good picture of what is happening or what is not happening* and to *bring facts to the table*. On this note, Susanne pointed to an event planned by the Austrian DSC for late May 2024 to share knowledge, exchange experiences and provide feedback on the processes foreseen by the DSA – we will certainly keep an eye out for it!

[6] See: https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Background note on data access

The Digital Services Act is a comprehensive framework for digital intermediaries, replacing the earlier eCommerce Directive, with a range of rules for different types of intermediaries. Our focus rests on the data access obligations introduced by the DSA for very large online platforms (VLOPs) and very large online search engines (VLOSEs).[7]

VLOPs and VLOSEs are defined as services which reach at least 45 million active recipients in the EU on a monthly basis and therefore could pose systemic risks – defined as dissemination of illegal content including disinformation, adverse effects on fundamental rights, on electoral processes and on gender-based violence or mental health. Systemic risks in part stem from the use of AI systems in these digital spaces.

Generally, providers of VLOPs and VLOSEs are required to assess and monitor systemic risks themselves and develop adequate countermeasures. However, the DSA also introduces the obligation to grant access to their data for studying systemic risks. Digital Services Coordinators (DSCs) can make use of the data access obligations, but so can the research community, which primarily extends to vetted researchers from academic institutions but also encompasses researchers from civil society organisations. These provisions are spelled out in Article 40(4-8) DSA for vetted researchers and Article 40(12) DSA for civil society organisations.[8]

Two aspects about data access are particularly interesting from our point of view. Firstly, these obligations enable wider monitoring of proprietary AI systems than any other tool. This means that the AI systems which affect almost everyone who is online may potentially become the best-studied AI systems – with the exception of open source AI systems. Considering the real and wide-ranging harms that have been enabled by recommendation algorithms, the transparency enabled by independent monitoring is a long-overdue step in the right direction.

Secondly, the obligations of VLOPs and VLOSEs to grant access to data differentiate strictly between researchers with an academic affiliation and those without. The former may access a wider variety and more detailed data than the latter, because civil society organisations are explicitly restricted to accessing publicly accessible data.

Recital 98 describes publicly accessible data as including *aggregated interactions with content from public pages, public groups, or public figures, including impression and engagement data such as the number of reactions, shares, comments from recipients of the service*. Prior to the DSA, access to such data was available mainly through APIs (accessible with a developer account) subject to frequent changes and inconsistencies, with in part cumbersome application and download processes, and the possibility of applications for access to data being rejected

[7] For a brief overview of the entire package, we recommend the European Parliament's briefing: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/689357/EPRS_BRI(2021)689357_EN.pdf
[8] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065

both for academic and non-academic research purposes.[9]

With the approval or rejection of requests to data access for academic research purposes now moving to the DSCs,[10] it is likely that applications will be processed more consistently and possibly for multiple VLOPs/VLOSEs at once. However, the process for researchers from civil society organisations is not yet clear, although it is certainly a step in the right direction that they now in principle have backing by legal provisions.

The divide between data access granted to academic and non-academic researchers is not fully understandable from our perspective. Much of the research that informed the notion of systemic risks in the DSA originated from non-academic researchers and civil society organisations and they have played a key role in communicating their findings to inform policy. Academic research is also incentivised to follow the funding, which means that the issues investigated tend to echo those of the institutions offering grants and may not align fully with civil society concerns. This is particularly worrisome because significant funding stems from VLOPs/VLOSEs and these companies may yet have an outsized influence on setting research agendas.

What we would therefore like to see is strong cooperation between academic institutions and civil society organisations in researching systemic risks of AI systems deployed on digital platforms – especially because the workload to ensure that we understand their basic workings is so high that we need to get all hands on deck. We are not alone in this wish for enhanced cooperation, as similar events organised e.g. by the Viennese Science, Research and Technology Fund (WWTF) show.[11]

[9] For a detailed discussion of these hurdles, see: https://www.wahlbeobachtung.org/wp-content/uploads/2021/04/guidelines_api_final.pdf

[10] The DSC of the country of establishment of the academic institution to which a researcher is affiliated can conduct an initial assessment of whether the status of "vetted researcher" is warranted and send its initial assessment as well as the application to the DSC in the country of establishment of the VLOP/VLOSE to which the data access request is addressed. The DSC in the country of establishment of the VLOP/VLOSE will make a final decision on whether academic researchers should be granted the status of "vetted researchers".

[11] https://www.wwtf.at/veranstaltung-der-eu-digital-services-act-chancen-fuer-den-forschungsstandort-wien/

Profiles of panellists

Jennifer Adams

Jennifer Adams (she/her) is an international expert and consultant on gender, media and human rights-centered digital policy. Her specialization is in the areas of program development, and policy research for feminist and inclusive governance, online and off. She is passionate about issues that lie at the intersection of human rights and global innovation, with a particular focus on the safety and participation of womxn and marginalized communities in digital development initiatives. As one of the leads of the Comms Policy Collaborative at the University of Vienna, she facilitates policy-oriented research on digital spaces and contributed to a range of projects through the CPC's partnership networks – including the premier partnership with the European Platform for Regulatory Authorities (EPRA) – on youth digital safety, VSP regulation, dis/misinformation, and Media and Information Literacy.

Susanne Lackner

Susanne Lackner is the Vice-Chair of the Austrian regulatory authority for media, KommAustria, which has recently been designated as the Digital Services Coordinator for Austria. She represented KommAustria as a member of the Board of the European Regulators Group for Audiovisual Media Services in 2023 and previously also worked for the European Commission in audiovisual media policy. Since 2021, her responsibilities at KommAustria have encompassed communications platforms and she now leads the efforts surrounding the enforcement of the Digital Services Act.

Thomas Lohninger

Thomas Lohninger is Executive Director of the digital rights NGO epicenter.works in Vienna, Austria. The Center of Internet and Society of the Stanford Law School holds him as a non-residential Fellow and he was Senior Fellow of the Mozilla Foundation working on Net Neutrality in the European Union. He worked on the European Net Neutrality regulation as Policy Advisor for European Digital Rights (EDRi) and is serving on the board of EDRi since 2019.

Rania Wazir

Rania Wazir is co-founder and CTO of leiwand.ai. She is a mathematician with degrees from Stanford (B.Sc.) and Brown University (Ph.D.). For the past five years, she has been active as freelance data scientist, focusing on Trustworthy AI and Natural Language Processing (NLP), notably for projects for Wahlbeobachtung.org, Amnesty International Italy, the Organisation for Security and Cooperation in Europea (OSCE) or the Fundamental Rights Agency (FRA). Rania is Austrian delegate to international standards organizations working on AI (CEN/CLC JTC21 and ISO/IEC JTC1/SC42), where she is actively engaged in the development of new AI standards. She is also a member of the advisory committee on the ethics of AI for the Austrian UNESCO commission.