Written evidence submitted by Lujain Ibrahim, Dr Luc Rocher, Dr Ana Valdivia, University of Oxford

May 2023

Introduction

1. We write as researchers in the area of technology, privacy and security, and human-computer interaction. Lujain Ibrahim is a doctoral student at the Oxford Internet Institute (OII). Dr Luc Rocher is a lecturer at the OII and the director of the DPhil programme in Social Data Science. Dr Ana Valdivia is a lecturer of AI, Governance, and Policy at the OII. Dr Rocher's research on the limitation of anonymisation practices has been referenced by many relevant bodies, such as DG CONNECT, DG FISMA, DG COMM, JRC, OECD, World Bank, WEF, AEPD, US FTC, EPA, in US legal cases, and led to changes to the UK's Data Protection Bill. Dr Valdivia has given evidence based on her research in algorithmic governance and AI to international bodies and national organisations, including the UK Parliament, Spanish government, Ministry of Education of El Salvador, and the European Parliament.

2. Our evidence below focuses primarily on the conditions of data access programs, as well as the specific need for research on platforms' algorithms and algorithm-user interactions.

3. Platforms play a pivotal role in influencing the health of our information systems, mediating our access to goods, services, and news among other essential aspects of modern life[1]. Until now, limited access to platform data has been a major issue, hindering academia and civil society's inquiries into platforms and platform algorithms' functions, benefits, and harms.

4. Platforms have shared data with academic researchers in the past, e.g., through tools such as Twitter Decahose, Facebook Social Science One, and Meta CrowdTangle. However, these tools always provide a very limited view of platforms' functioning. In 2021, academics discovered systematic gaps in Crowdtangle transparency data that Meta was providing to academics and regulators.[2,3] Scraping techniques, often used by academics and independent researchers to gather further data, cannot provide information on the underlying algorithms that platforms use. Studying the algorithms that make important

[1] Bak-Coleman, J.B., Alfano, M., Barfuss, W., Bergstrom, C.T., Centeno, M.A., Couzin, I.D., Donges, J.F., Galesic, M., Gersick, A.S., Jacquet, J. and Kao, A.B., 2021. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, *118*(27), e2025764118.

[2] Bobrowsky, M, 2021. Facebook Disables Access for NYU Research Into Political-Ad Targeting. WSJ.

[3] Matias, J.N., 2023. Humans and algorithms work together—so study them together. *Nature*, *617*(7960), pp.248-251.

determinations (on, e.g., content moderation and recommendations) is currently extremely challenging.

Conditions of Data Access Programs

5. The European Data Protection Supervisor (EDPS) wrote in its Preliminary Opinion on Data Protection and Scientific Research that “data protection obligations should not be misappropriated as a means for powerful players to escape transparency and accountability”.[4] In this section, we address the increasing use of privacy-enhancing technologies, and call for caution in their deployment for vetted data access.
6. Privacy-enhancing technologies (PETs) are now increasingly used by public bodies (e.g., US Census Bureau, UK NHS and ONS) and by companies (e.g., Apple, Google, Uber) to make sensitive human data “anonymously” available to researchers. Modern PETs range from “differential privacy” (recently employed by the US Census) to generative deep learning models applied to social graphs, timeseries, and images to produce synthetic data. However, researchers have expressed strong concerns that the validity of research findings may be altered by privacy-preserving techniques,[5] distorting statistical inferences and increasing disparities in outcomes for racial minorities.[6]
7. Researchers have raised the alarm that modern and hailed approaches, such as ‘synthetic data’ mechanisms, generally offer the same trade-offs as traditional anonymisation techniques in practice.[7,8] Upon scrutiny, modern mechanisms such as ‘differential privacy’ used by platforms such as Google and Apple were found to provide weaker guarantees than initially communicated.[9,10] Yet setting the ‘privacy cursor’ higher (by adding noise to data or removing outliers) could reduce the quality of data in unknown but potentially harmful ways.
8. The lack of an anonymisation silver bullet implies that VLOPs would be required to make important determinations between protecting re-identification and safeguarding usability for researchers, when using anonymisation techniques.
9. Instead, we advocate for the use of existing ‘security-by-design’ solutions implemented notably in biomedical research, where secure and trusted platforms allow for safe and collaborative research.[11] The development of trusted infrastructure, held by universities or

[4] EDPS, 2021. A Preliminary Opinion on data protection and scientific research. https://edps.europa.eu/sites/default/files/publication/20-01-06_opinion_research_en.pdf

[5] Hauer, M.E. and Santos-Lozada, A.R., 2021. Differential privacy in the 2020 census will distort COVID-19 rates. *Socius*, *7*, p.2378023121994014.

[6] Santos-Lozada, A.R., Howard, J.T. and Verdery, A.M., 2020. How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences*, *117*(24), pp.13405-13412.

[7] Stadler, T. and Troncoso, C., 2022. Why the search for a privacy-preserving data sharing mechanism is failing. *Nature Computational Science*, *2*(4), pp.208-210.

[8] Annamalai, M.S.M.S., Gadotti, A. and Rocher, L., 2023. A Linear Reconstruction Approach for Attribute Inference Attacks against Synthetic Data. *arXiv preprint arXiv:2301.10053*.

[9] Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12. *J. Tang, A. Korolova*, X. Bai, X. Wang, and X. Wang. CoRR (2017).

[10] Houssiau, F., Rocher, L. and de Montjoye, Y.A., 2022. On the difficulty of achieving Differential Privacy in practice: user-level guarantees in aggregate location data. *Nature Communications*, *13*(1), p.29.

[11] Goldacre, B., 2022. Better, broader, safer: using health data for research and analysis. DHSC https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis

an independent body, would be sufficient if strict confidentiality and privacy rules determine how the analyses done on trusted infrastructure can be shared, in particular to regulators and the broader scientific community. Trusted research infrastructure provides a high level of security and confidentiality while allowing researchers to undertake a wide range of research tasks—that they might not succeed on coarse, aggregated, or anonymised data.

10. Finally, we believe that expanding data access in a responsible manner to researchers outside of the Europe Union is critical. The vast majority of users of social media platforms are outside of Europe, with some of the most represented countries being in the so-called Global South (e.g., India, Indonesia, Brazil, etc.).[12] Additionally, some of the largest platform failures, on the content moderation front as well as other fronts, have unfolded in the Global South (e.g., facilitating the Myanmar genocide,[13] and inciting Islamophobic violence in India[14]). Simultaneously, research on the functions and failures of platforms in those regions remains underrepresented and unexplored. Enabling data access through secure infrastructure in the European Union is likely to benefit independent researchers and civil society organisations. We could envision that independent researchers without access to such resources could benefit from the infrastructure of, e.g., universities with the technical capacity. This is notably the model used for academic high-performance computing where researchers can access pools of servers across Europe.

Category of Data Required to Study Platform Algorithms

11. Algorithms have been an important function of many online platforms. More recently, they have taken up an even greater role with the rise of algorithmically-curated feeds and automated content moderation. On a growing number of platforms, recommender algorithms are augmenting and replacing traditional forms of human content moderation (through algorithmic amplification and demotion). To study the impact of these automated technologies, data is needed beyond publicly-available content and engagement data (e.g., likes, shares, views, reactions). Yet the tools, APIs, datasets currently shared by platforms to vetted researchers are, in their present state, not sufficient to understand how these algorithms function.[15]
12. For instance, the deployment of algorithms by platforms create complex systems that are characterised by human-algorithm feedback loops. Outcomes are determined by a combination of platform decisions, algorithms, as well as (collective and individual) human behaviour. The difficulty in delineating which input of this complex system leads to which outcome (e.g., account being banned, content being moderated) has presented a

[12] Number of social network users worldwide in 2022, by region, Statista (2022)
[13] Milmo, D., 2021. Rohingya sue Facebook for £150bn over Myanmar genocide, The Guardian.
[14] Basu, S., 2019. Manufacturing Islamophobia on WhatsApp in India, The Diplomat.
[15] Albert, J., 2023. Platforms' promises to researchers: first reports missing the baseline. Algorithm Watch. https://algorithmwatch.org/en/platforms-promises-to-researchers/

serious problem for both the study of these systems as well as the delegation of responsibility for unwanted/harmful outcomes.

13. Another example is the use of human controls over recommendation and search results, such as YouTube's feedback buttons like "Dislike" and "Don't Recommend Channel". Research, user studies, and large-scale citizen science audits (including the audit of 22,722 people's feeds conducted by the Mozilla Foundation[16]) have shown that users believe controls are not only buried and difficult to find, but they are also functionally ineffective at preventing unwanted recommendations. As platforms continue to point to the existence of these tools as evidence for sufficient human control over platform outcomes, there is a need to further scrutinise and investigate such claims to protect subjects and marginalised communities.
14. The two examples highlight the need for advanced data access beyond public user-generated content and metadata. We believe that sharing data on the nature, usage, and effectiveness of human-algorithm interactions—in particular user controls pertaining to recommender & moderation systems—would be highly beneficial to vetted researchers. In addition to engagement data and recommendation data, additional data needed to study user controls may for instance include:
    a. Documentation of implemented user controls and the outcomes they lead to when used;
    b. Usage data of user controls & feedback signals;
    c. Data and metadata for content flagged by users through controls, and the resulting automated decisions.

[16] Ricks, B and Jesse McCrosky, J., 2022. Does This Button Work? Investigating YouTube's ineffective user controls. Mozilla. https://foundation.mozilla.org/en/research/library/user-controls/report/