**Enabling Data-Driven Research for Evidence-Based Interventions**

**Written Evidence to Ofcom: Call on researchers' access to information from regulated online services.**

Evidence submitted by the Minderoo Centre for Technology asnd Democracy at the University of Cambridge.

Dr Hugo Leal, Dr Stefanie Felsberger, Dr Timothy Charlton, and Professor Gina Neff

**24 January 2025**

**Headline:**

Data access depends on platforms' policies. While platforms once provided open APIs and/or data access programmes for researchers, most of these options are no longer available. This makes reliable and verifiable research into online safety matters practically impossible. A new framework must redress this imbalance in the name of public interest research and evidence-based policymaking.

**About MCTD:**

The Minderoo Centre for Technology and Democracy (MCTD) is an independent team of academic researchers at the University of Cambridge who are radically rethinking the power relationships between digital technologies, society, and the planet. Through our ambitious research agenda, we are enhancing public understanding of digital technologies and delivering positive changes to society's relationship with these technologies. We are also part of the EU Horizon 2020 funded project AI4TRUST: AI-based-technologies for trustworthy solutions against disinformation, a project of 11 partners across 17 countries building trustworthy, beyond state-of-the-art misinformation detection and mitigation tools to amplify the human efforts to counter disinformation. We welcome this opportunity to contribute to this call.

## **Question 1:** How, and to what extent, are persons carrying out independent research into online safety related issues currently able to obtain information from providers of regulated services to inform their research?

1. The landscape for conducting meaningful research into online safety issues has eroded over time. Despite years of striving for transparency and observability, platforms today operate in a state of opacity.

2. Independent research on online safety has relied on companies that provide access to data. As a result some platforms have been better studied than others, and research is often based on a single platform.[1] Previously, the focus for much academic research was Twitter (now X) because the platform's API provided a relatively easy gateway for researchers. Insights into disinformation from Twitter were often extrapolated to represent social media as a whole; this analysis was limited because it glossed over the characteristics, cultures, dynamics, user bases, and safety challenges that are unique to each platform.[2]

3. This history of social media research underscores how data access policies for researchers significantly shape the production of knowledge. Establishing standardised access policies for researchers across all relevant online service providers is essential for fostering an accurate and consistent understanding of online safety issues across existing platforms.

4. As researchers on information integrity, we are acutely aware of the role played by platforms in the spread of misinformation about vitally important issues, including vaccines, climate change, and elections.[3] There are also challenges for information integrity as rogue actors promote harmful content, and online polarisation increases, alongside the 'organic' spread of harmful narratives or material.

5. Research on online safety issues has become even more difficult as platforms have closed research programmes and interfaces that facilitate researchers' access to data. For example, Twitter discontinued the free Premium API for researchers in February 2023, and META shut down Crowdtangle in August 2024.

6. Today, many researchers rely on either insufficient or outdated data to produce outputs. Policy recommendations for key information integrity issues are often based on poor quality data. For instance, we have no way of knowing how algorithms, individuals, groups, companies, or states determine what we see, read, and believe online. Researchers can no longer analyse and map networks of harmful content, *independent of platforms' direct consent*. With no legal obligations on platforms to provide data access to

---

[1] Roger Rogers, "Digital Methods for Cross-Platform Analysis," in *The SAGE Handbook of Social Media*, ed. Jean Burgess, Alice Marwick, and Thomas Poell (Sage, 2017), 91-110.

[2] Warren Pearce, Suay M. Özkula, Amanda K. Greene, Lauren Teeling, Jennifer S. Bansard, Janna J. Omena, and Elaine T. Rabello, "Visual Cross-Platform Analysis: Digital Methods to Research Social Media Images," *Information, Communication & Society* 23, no. 2 (2020): 161-180.

[3] See for example: Steven Lloyd Wilson and Charles Wiysonge, "Social Media and Vaccine Hesitancy," *BMJ Global Health* 5, no. 10 (October 1, 2020): e004206, https://doi.org/10.1136/bmjgh-2020-004206; and Michela Del Vicario et al., "The Spreading of Misinformation Online," *Proceedings of the National Academy of Sciences* 113, no. 3 (2016): 554–59.

researchers, social media networks are opaque and potentially harmful operations—whether real or perceived—can thrive.[4]

7. Government and regulators should guarantee that researchers have legal rights to platform data to inspect how platforms' policies and business practices affect the circulation of information, misinformation and disinformation, and its impacts on society.

## **Question 1a:** What kinds of online safety research does the current level of access to information enable?

- What type of independent researchers are carrying out research into online safety matters?
- What topics/issues are they researching?

8. We are involved in the AI4TRUST project, an EU Horizon Research and Innovation Action programme aimed at identifying, analysing and countering the spread of misinformation and disinformation online. The AI4Trust consortium's work has been severely affected by narrowing avenues for researcher access to platform data.[5] At the time of our proposal, AI4TRUST could reasonably expect to have access to Twitter's Premium API for researchers, Facebook's Crowdtangle (for Facebook and Instagram), YouTube's Researcher Program, and publicly available news sources.

9. With the phasing out of established programmes and technical solutions for researcher data access, the AI4Trust project had to look for alternative sources of access. This has been a burdensome process. Our project submitted requests for data access to both Twitter/X and the Meta content library under the EU DSA. Both were rejected. TikTok was considered as an alternative data source, but our request there has stalled. Currently, our project has only managed to secure researcher access to YouTube's Global Data API through the YouTube research programme. We are also working to access Telegram's open API.

10. The rapidly declining ability of researchers to access platform data has been a major challenge to our research on information integrity. Platforms' restrictions on data access hamper both the development of innovative scientific methods

[4] See for example: Elizabeth Seger et al., "Tackling Threats to Informed Decision-Making in Democratic Societies: Promoting Epistemic Security in a Technologically-Advanced World" (The Alan Turing Institute, 2020); or Bernhard Rieder and Jeanette Hofmann, "Towards Platform Observability," *Internet Policy Review* 9, no. 4 (December 18, 2020), https://policyreview.info/articles/analysis/towards-platform-observability.
[5] For more information on the AI4TRUT project, see https://ai4trust.eu/.

to identify, map, and counter disinformation, including the training of new AI tools for the public good.

11. AI4Trust has built AI verification tools for online content. This was only possible because partners in the consortium shared access to online social media data that they had collected through their own activities. While these tools allow us to meet some of our information integrity goals, the AI models that we trained on these datasets will not be able to counter *evolving* types of misinformation and disinformation disseminated on social media.

12. The limited data access environment has fundamentally hindered the depth and breadth of our research capabilities in AI4TRUST. The final AI4Trust product will be much more limited in range and scope than originally proposed, and our work has been significantly delayed.

13. In our case, a legally binding data access framework as proposed in the UK's Online Safety Act would enable us to conduct groundbreaking research of vital importance on misinformation and disinformation. Without such access, we will not be able to produce meaningful research on the diffusion of misinformation and disinformation on our three main research themes: environment, migration and public health.

**Question 1b:** Are there types of information that independent researchers are currently unable to access that may be relevant to the study of online safety matters? If so, what are they and what kind of research would they facilitate?

14. In June 2020, an article in *Nature* explained how researchers had been using data from social media platforms to "revolutionise social science" and contribute to the betterment of society.[6] Less than three years later, the same journal reported how a decision to end data access threatened "to upend large social media studies".[7] The prospect of completely shutting down researchers' access to data threatens the ability to do science and make recommendations about digital society.

15. For instance, we have no scientific way of examining exactly how the racially motivated riots moved from social media platforms onto England's streets in Summer 2024. Researchers cannot quantify or qualify the dissemination of

[6] Heidi Ledford, "How Facebook, Twitter and Other Data Troves Are Revolutionizing Social Science," *Nature* 582, no. 7812 (June 17, 2020): 328–30, https://doi.org/10.1038/d41586-020-01747-1.
[7] Heidi Ledford, "Researchers Scramble as Twitter Plans to End Free Data Access," *Nature* 614, no. 7949 (February 14, 2023): 602–3, https://doi.org/10.1038/d41586-023-00460-z.

misinformation and disinformation about vaccines on social media platforms. Child well-being online cannot be rigorously assessed.

16. This type of research is especially needed in the current environment. Social media platforms increasingly rely on automated content moderation solutions that are often opaque and their efficacy must be independently verified. Companies, like Twitter/X, have also drastically reduced their trust and safety teams.[8] In January 2025, Meta announced the end of its cooperation with fact-checking partners in the United States.[9] Amidst rising concerns over the increasing spread of misinformation and disinformation across social media platforms and the proliferation of more sophisticated AI-generated content, independent cross-platform research is indispensable in order to understand how harmful information flows through/via different platforms.

## **Question 1c:** What data governance models are currently used to allow access to online services' information for researchers?

- This might include: open-access forms of information-sharing, such as publicly-accessible information libraries or databases; information-sharing models that rely on vetting or accreditation of individuals or organisations; and/or models that rely on the accreditation of the specific use cases for the information.
- Please provide relevant examples of these governance models used in the online services industry.

17. Under current data frameworks, platforms have become the sole gatekeepers of data access. This leads to a state of information asymmetry, in which private actors hold more information about the public than public interest institutions. This information asymmetry also means platform owners have better data relevant for social policy than states or citizens. As discussed, previous existing access regimes such as API have been limited or discontinued, exacerbating the problem.

18. Some believe industry-academic partnerships to be a possible solution. They create 'independence by permission' by which platforms control access to data and determine research priorities.[10] For example, during the 2010 US Congressional elections, Facebook and academic researchers collaborated to run a "randomized controlled trial of political mobilization messages" to

[8] Thomas Brewster, "Elon Musk Fired 80 Percent of Twitter/X Engineers Working on Trust and Safety," *Forbes*, January 10, 2024, https://www.forbes.com/sites/thomasbrewster/2024/01/10/elon-musk-fired-80-per-cent-of-twitter-x-engineers-working-on-trust-and-safety/.
[9] Liv McMahon, Zoe Kleinman, and Courtney Subramanian, "Facebook and Instagram get Rid of Fact Checkers ," *BBC News*, January 15, 2025, https://www.bbc.co.uk/news/articles/cly74mpy8klo.amp.
[10] Michael W. Wagner, "Independence by Permission," *Science* 381, no. 6656 (July 28, 2023): 388–91, https://doi.org/10.1126/science.adi2430.

understand whether they influence voter turnout.[11] There was a strong statistical relationship between seeing the messages on the platform and voting. Facebook's experiment directly and intentionally led 60,0000 voters to the polling stations.[12] In another example, Facebook tested the networked "contagion of emotional expression"[13] by dialling up and down the positive and negative content in the feeds of almost 700,000 users.[14] The Social Science One (SSO) partnership was established after the US 2016 elections. After 18 months of operation, the European Advisory Committee of SSO made a public statement denouncing the fact that Facebook had "not provided academics with anything approaching data access".[15] The 2020 Meta election study tested questions about the role of the company's platforms on the US Presidential election.

19. The challenge for these partnerships is that the research must take place on the companies' terms. The researcher evaluating the research process of the Election 2020 study concluded, "independence by permission is not independent at all [...] Scholarship is not wholly independent when the data are held by for-profit corporations, nor is it independent when those same corporations can limit the nature of what it studied".[16]

20. Companies continue studying their own platforms. While the bulk of their research is geared towards the monetisation of users' digital footprint, platforms have also been running large-scale social experiments.[17] Academic and Civil Society Organisation researchers can neither independently initiate nor verify research projects using the data unduly retained by the platforms. This means that research in the public benefit, such as that required for an online safety regime, can currently only happen with the companies' permission.

---

[11] Robert M. Bond et al., "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature* 489, no. 7415 (September 2012): 295–98, https://doi.org/10.1038/nature11421.
[12] Will Oremus, "One Facebook Banner Ad Caused 60,000 More People To Vote in the 2010 Elections," *Slate*, September 13, 2012, https://slate.com/technology/2012/09/facebook-voting-study-online-friends-influence-voter-turnout-in-elections.html.
[13] Lorenzo Coviello et al., "Detecting Emotional Contagion in Massive Social Networks," *PLOS ONE* 9, no. 3 (March 12, 2014): e90315, https://doi.org/10.1371/journal.pone.0090315.
[14] Kashmir Hill, "Facebook Manipulated 689,003 Users' Emotions For Science," *Forbes*, accessed January 18, 2025, https://www.forbes.com/sites/kashmirhill/2014/06/28/facebook-manipulated-689003-users-emotions-for-science/.
[15] "Public Statement from the Co-Chairs and European Advisory Committee of Social Science One," December 11, 2019, https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one.
[16] Michael W. Wagner, "Independence by Permission," *Science* 381, no. 6656 (July 28, 2023): 388–91, https://doi.org/10.1126/science.adi2430.
[17] Kevin Roose, Mike Isaac, and Sheera Frenkel, "Facebook Struggles to Balance Civility and Growth," *The New York Times*, November 24, 2020, sec. Technology, https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.

21. Some APIs continue in a limited way. For instance, Alphabet maintains the YouTube research programme as respective API access open to academics. Telegram's API can also be used to access data on public broadcasting and discussion channels. The case of Reddit also illustrates the changing nature of data access. Reddit limited access to its official API in 2023. A third-party dedicated archive of Reddit communities called Pushshift once worked with Reddit's API as a 'mirroring' and archival service. It allowed for large computational analysis over longitudinal data spanning many years. But Pushshift is no longer accessible to the general public, only to moderators of Reddit communities.[18] The change in Reddit's API policy coincided with high-value licensing agreements on the use of Reddit user data in AI model training.[19]

22. Current data governance is a preserve of private platforms. Most researchers we speak with acknowledge that platforms routinely reject all kinds of data requests. For example, to the best of our knowledge, there are no independent researchers with proper data access to TikTok, one of the fastest-growing social media platforms in the UK.

## **Question 2:** What are the challenges that currently constrain the sharing of information for the purpose of research into online safety related issues?

23. Platforms have closed technical avenues to data access and restricted legal avenues to access data by making some methods of access a violation of their Terms of Service. Platforms can and have banned researchers who have tried to use academic freedom rights to deliver public interest research. Platforms' tactics have included lawsuits against researchers and projects that are using demonstrably legal means to obtain data, such as crowdsourced data collection and web-scraped publicly accessible data.[20] *Sandvig v Barr*

---

[18] See for example: Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 830–839. https://doi.org/10.1609/icwsm.v14i1.7347; or https://www.reddit.com/r/reddit/comments/145bram/addressing_the_community_about_changes_to_our_api/.

[19] E.g. the Reddit-Google Partnership: https://blog.google/inside-google/company-announcements/expanded-reddit-partnership/

[20] See for example: Nick Robins-Early, "Judge Dismisses 'Vapid' Elon Musk Lawsuit against Group That Cataloged Racist Content on X," *The Guardian*, March 25, 2024, sec. Technology, https://www.theguardian.com/technology/2024/mar/25/elon-musk-hate-speech-lawsuit; and: Jeff Horwitz, "WSJ News Exclusive | Facebook Seeks Shutdown of NYU Research Project Into Political Ad Targeting," *Wall Street Journal*, October 23, 2020, https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533.

sought clarity on companies not prosecuting academic researchers for independent audits of their systems for discriminatory bias.[21]

## **Question 2a:** What are the legal challenges/risks to sharing information from online services with independent researchers?

24. First, platforms use their Terms of Service to define the rules of engagement, and there are many examples of platform-led legal action against both academic researchers and civil society organisations.[22] A clear data access framework is urgently needed to stop this.

25. Second, large online service providers spent years lobbying against the approval of a comprehensive personal data protection framework, such as the GDPR. But, at the same time companies have used privacy protection as a justification for blocking research into their policies, including into the handling of political ads and misinformation campaigns.[23] The same contradictory argument has been used to undermine important regulatory efforts, such as the EU Digital Services Act, the UK Online Safety Act, and provisions (primary or delegated) to mandate researchers' access. Social science and medical research have robust procedures and guardrails for privacy protection. Privacy concerns should not be used to bludgeon calls for evidence-based research in the public interest.

26. Third, in terms of the legal and ethical use of data by researchers: academics and researchers at NGOs have long dealt with these questions across a series of disciplines, including in medical and social sciences. In the UK, universities typically have robust systems in place to guarantee ethical compliance, at all levels (individual researcher, project, and departmental levels). Examples of academic approval procedures for personal data use frequently involve peer networks, line managers and three levels of Research Ethics Committees (Departmental, School, and University) depending on the complexity of the project and the associated risks for both the researchers and the data subjects.

27. Academic research has established mechanisms for ensuring accountability, informed consent, and beneficence. Most large online services have internal

---

[21] *Sandvig v Barr*. https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2016cv1368-67

[22] Shannon Bond, "Elon Musk Sues Disinformation Researchers, Claiming They Are Driving Away Advertisers," *NPR*, August 1, 2023, https://www.npr.org/2023/08/01/1191318468/elon-musk-sues-disinformation-researchers-claiming-they-are-driving-away-adverti.https://www.npr.org/2023/08/01/1191318468/elon-musk-sues-disinformation-researchers-claiming-they-are-driving-away-adverti.

[23] "Research Cannot Be the Justification for Compromising People's Privacy," *Meta* (blog), August 4, 2021, https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/.

research departments that carry out research without any public accountability, verification, traceability, and reproducibility. A data access regime for research independent of the platforms can build on existing models from medical and social research to address privacy and other concerns. These models and mechanisms include ethical review, training, and technical measures for ensuring individual-level and sensitive data are handled appropriately.

### **Question 2b:** What are the technical challenges relating to sharing information from online services with independent researchers?

- What are the challenges relating to the scale and complexity of the information involved?

28. Providing API access to vetted independent researchers is a fully tested model of access that platform companies themselves approved until only very recently. Vetted independent researchers could be invited to submit data management plans that guarantee the integrity of data sets and research, and protection of users' privacy. A third-party public institution, established with the input of UKRI, should assume the role of vetting institute and data intermediary.

### **Question 2c:** What are the security challenges relating to sharing information from online services with independent researchers?

- What are the security challenges relating to the potential sensitivity of information?
- What are the security protocols required to protect information from misuse?
- To what extent do you view security as a governance issue compared to a technical infrastructure issue?

29. Currently platforms are opaque to policymakers and the public. This opacity poses a potential national security threat that is difficult for governments to assess. Being able to independently assess the health and safety of online platforms is impossible under current conditions.

### **Question 2d:** What are the information quality challenges relating to online services sharing information with independent researchers?

30. Questions of data integrity, data completeness, data validity, etc., are particularly pertinent when platforms control data access. Social Science One (SSO) illustrates the challenges of incomplete data. The dataset Facebook shared to study the effects of social media on elections was flawed. Fabio Giglietto, a researcher on the Vera.ai project, found out the data contained serious errors, affecting the findings in an unknown number of academic

papers.[24] In his letter of resignation as co-chair of SSO, Stanford legal scholar Nate Persily wrote that they had "learned many lessons over that period about how difficult it is to dislodge data from social media companies for independent academic analysis".[25] Since then, the situation only got worse. A data access regime cannot rely on a company's good will for compliance. It must be underpinned by a robust and transparent framework that guarantees that platforms will share complete datasets subject to independent verification, validation, and quality controls.

## **Question 2e:** What are the financial costs to online services relating to online services sharing information with independent researchers?

31. As we have described, online services preserve, use, and often abuse user data. They possess vast data infrastructures that, as demonstrated by the Crowdtangle experience, can easily accommodate seamless data access frameworks. Although Crowdtangle was insufficient, incomplete, and fully controlled by META, the programme did provide academic researchers and journalists access to relevant data (e.g. public groups and pages) at modest operation costs. If regulation includes the creation of a public data infrastructure, such as the recently proposed National Data Library, a significant proportion of these costs would be shifted to this third-party entity.[26] Furthermore, important questions, such as data transferences, maintenance, preservation, and safety would become shared attributions of the new data controllers and processors, namely the public data infrastructure and the researchers.

32. Also in the context of costs to platforms, it is important to distinguish between very large platforms, and small and medium ones. For example, the EU's Digital Services Act places different requirements on 'very large online platforms' (VLOPs) and 'very large search engines' (VLOSE), as compared to small or medium platforms. This is based on user scale and the wider impact platforms have on society.[27] The UK Online Safety Act does not make this distinction, but with regards to researcher access policies, a distinction could

---

[24] Craig Timberg et al., "Facebook Made Big Mistake in Data It Provided to Researchers, Undermining Academic Work," *Washington Post*, September 10, 2021, https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/.
[25] "Some Changes at Social Science One," October 2, 2020, https://socialscience.one/blog/some-changes-social-science-one.
[26] Emma Gordon, "The New UK Government Wants a National Data Library: A Brilliant Aspiration, If Built on Solid Foundations," accessed January 21, 2025, https://www.adruk.org/news-publications/news-blogs/the-new-uk-government-wants-a-national-data-library-a-brilliant-aspiration-if-built-on-solid-foundations/.
[27] "DSA: Very Large Online Platforms and Search Engines | Shaping Europe's Digital Future," January 20, 2025, https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops.

be implemented.

33. The financial costs of one-size-fits-all data sharing framework would also fall unequally on different online services depending on their size. Such an approach to data access could either overwhelm smaller and medium platforms by asking too much of them, or fall short on asks to large platforms by failing to account for their greater capacity to share high-quality data.

**Question 2f:** What are the financial costs to researchers trying to make use of information shared by online services?

34. While not all research involves huge amounts of data, few institutions have the capacity for high-intensity data analysis (e.g. with HPC, AI, or storage). This sort of high-intensity data handling could represent the most obvious financial cost to researchers. Partnerships between research organisations or publicly available compute capacity are amongst solutions that could be explored once data access for online safety research has been established.

35. The indirect financial costs to UK Academia and the British economy should also be considered. The academic sector in the UK is world-leading. With the development of comprehensive data access frameworks in other areas, such as the EU, UK universities risk losing their competitive advantage. Universities, NGOs, and policymakers in the EU could benefit from this potential knowledge gap and gain a competitive edge. This will represent another blow to an industry that is at the forefront of the UK government's plans to reinvest in science and the knowledge economy over the coming years.

---

**Question 3:** How might greater access to information for the purpose of research into online safety issues be achieved?

**Question 3a:** What models, arrangements or frameworks exist for allowing researchers access to sensitive information beyond the online services industry? What are the benefits and risks of those models, and how might they apply to the online services context?

36. Under the UK GDPR, data controlled and processed by platforms are legally owned by their 'users'. Platforms use user data for operational and profit-seeking purposes. A system for researcher access to data could mandate full legal compliance with data access regulation, the reintroduction of robust

application programming interfaces, and the implementation of interoperable protocols that allow access to vetted independent researchers and the production of ethical research in the public good for online safety.

37. The absence of legal, safe, and ethical pathways for researcher access to platform data has given rise to ‘parallel data markets’. This is a direct consequence of the current model. These alternative data access models should be a source of great concern as they pose serious methodological and ethical problems. Frequently, both the business models of these private, for-profit data brokers, and their data collection practices, are completely opaque. There are no codebooks, and no information on sampling criteria that enable serious research to occur. Some also operate in gray zones with regard to privacy regulation.

38. As part of our work for AI4TRUST, we were approached by a self-described ‘non-profit initiative’ that offered access to data from specific platforms. The datasets in question were made up of publicly accessible data scraped from the internet. Web scraping publicly available data has been deemed legal in some jurisdictions but not others. After seeking legal advice from partners and third-parties, we decided to run some experiments with the data the company was willing to share.

39. The company pitches to their clients highly sought after cross-platform data access. Clients are offered API and/or dashboard access to publicly available data scraped from several platforms. As there is neither an overarching research project nor an underlying research design with accompanying detailed codebooks, sampling criteria, etc., the datasets are collected, appended, mixed, and sold in bulk.

40. After performing some due diligence, AI4TRUST researchers found out not only that the company’s data integrity, but also their business model, was highly questionable. Further research revealed that initiative was the non-profit arm of a company that had been involved in a lawsuit that alleged the company had accessed personal user data without users’ consent or knowledge. Additionally, the company was accused of running a botnet through a VPN without users’ consent. The company allegedly also worked with surveillance companies accused of violating human rights. It became clear that the initiative did not comply with the same ethical standards and legal requirements for data governance as our project.

41. In a context where accessing data for research is becoming increasingly difficult, similar initiatives and rogue data brokers are emerging and creating a parallel data market. If independent researchers are forced to access data from such data brokers, without full knowledge of their background or data

collection methods, it could lead to flawed analyses and misguided policy recommendations. This can undermine research integrity and evidence-based policymaking.

**Question 3b:** Are there any models or arrangements that exist in the online services industry already that might provide increased access to information for research purposes if applied more generally across the industry? If so, what are these and what are the benefits and disadvantages of these models/arrangements?

42. A good model for data access is how **NHS England** decides which researchers, charities, and medical companies can gain access to health data. Their robust framework prioritises security, transparency, and accountability, and ensures that all data is used to benefit health and social care.[28]

43. NHS England regularly publishes information on what data is available through Data Access Environment (DAE), a secure platform and interface that facilitates remote access to NHS's available data for authorised institutions and individuals, increasing data security by reducing the need for data to be transferred or stored elsewhere. The DAE also features tools that allow a wide range of research activities, such as public health or clinical research. Data is only accessible after a rigorous process of application, review, approval, and contractual agreements managed by the Data Access Request Service (DARS).

44. After approval, two different contracts provide safeguards over data use, upon which access to data is conditional. A Data Sharing Framework Contract specifies the legally binding terms and conditions, while the Data Sharing Agreement delineates the "specific use of data covering the data required, minimisation, purpose, funding, any commercial element, processing activities, outputs, benefits and transparency and legal basis".[29] The contracts ensure the NHS has control over what the data is used for and what action needs to be taken if contracts are broken.

45. While the data sharing contracts are active, the NHS also commissions independent audits confirming whether data recipients meet all obligations, including whether data is stored safely and securely and is used for the

[28] 'How NHS England Makes Decisions about Data Access', *NHS England Digital*, accessed 22 January 2025, https://digital.nhs.uk/services/data-access-request-service-dars/how-nhs-england-makes-decisions-about-data-access.
[29] ibid

purposes set out in the agreements. Audit reports are also publicly available, allowing public scrutiny of the process. Data must also be deleted when an agreement expires.

46. Access is restricted to charities and academic institutions, and commercial organisations only where they can clearly demonstrate their data use is tied to creating "benefit to health and social care" but not to sell products or services, to conduct market or advertising research.[30] This could provide a model for a similar online safety-focused regime.

47. Another model is being put forward under the **EU Digital Services Act**. Here, "systemic risks" are defined as a priority and provide the justification to establish that VLOPs and VLOSEs have mandatory duties to report on systemic risks (Article 34) and propose mitigation measures (Article 35). Moreover, the document also includes path-breaking provisions on researchers' data access and scrutiny (Article 40). Provisions on data access will be further clarified through a 'Delegated Act' (secondary legislation), which should be released in the first months of 2025.

48. Article 40 of the DSA provides a legislative frame of reference to guarantee that "vetted researchers" requesting data for research projects related to the "detection, identification, and understanding of systemic risks" must be given access to data, "including, where technically possible, in real time". Although restricting data access to research addressing 'systemic risks' might sound limiting, the scope of eligible research is broad. Projects examine a host of relevant topics, ranging from dissemination of illegal content to negative effects on electoral processes, public security, civic discourse, the exercise of fundamental rights, gender violence, public health, protection of minors, and more.

49. Providing independent researchers are 'vetted', an individual project looking at the spread of gender-based violence in one country and on a single platform, and a consortium looking at gender-based violence in multiple countries, across multiple platforms, will both be eligible for data access. Knowing that human, technical, and financial resources are unequally distributed, this framework seeks to adapt data access rights to research capacity differences. Article 40 tries to create a level playing field for all vetted researchers investigating the broad range of topics related to systemic risks.

50. The expression 'vetted researcher' has been the object of some debate. While the 'Delegated Act' will legislate on the specifics of Article 40, the DSA already provides a general definition. Vetted researchers (Art 40(8)) must be affiliated

[30] Ibid.

with a research institution. This definition covers academia and organisations (e.g. NGOs) thus providing an inclusive data access framework that acknowledges the great work being done by CSOs, particularly in the field of digital rights and risks. To become 'vetted', researchers must also meet specific criteria, including independence from commercial interests, transparent funding, technical capacity to fulfil data security and confidentiality requirements

51. Regarding the modalities of access, Article 40 contains provisions according to the nature of the risk and the purpose of the research. Paragraph 12 stipulates that all vetted researchers, including academics and those in CSOs, should be given access to all publicly accessible data in the platforms' online interfaces. For example, in the case of X, this comprises all public posts, whereas on Facebook this covers public posts, pages, and groups. Paragraph 4 includes a provision of access to data to investigate the "adequacy, efficiency and impacts of the risk mitigation measures". According to the EU's European Centre for Algorithmic Transparency (ECAT), which supports the European Commission in the implementation and enforcement of the DSA, paragraph 4 opens researchers' access to non-public data through the submission of a data access request to a Digital Services Coordinator (DSC).[31]

52. Data access requests concerning publicly accessible data must be provided "without undue delay" to vetted researchers without intermediation of the DSC. Although this regulation needs to be clarified by the 'Delegated Act', Paragraph 12 is being interpreted as stipulating an obligation to VLOPs and VLOSEs and a right to researchers. The obligation concerns the (re)creation of platform-based intermediary data access tools (e.g. Open APIs and/or CrowdTangle), whereas the right is the long-sought legal support to allow web scraping of publicly available data based on the WYSIWYG (what you see is what you get) principle. While Paragraph 12 is merely reiterating the general understanding and jurisprudence on web scraping, it does provide legal backing to researchers and protects us from the often frivolous legal action initiated by platforms.

53. Another matter that will be further regulated by the 'Delegated Act' is relative to interfaces, codebooks, data formats, etc. The DSA stipulates that platforms "shall facilitate and provide access to data […] through appropriate interfaces specified in the request, including online databases or application

[31] "DSA Data Access for Researchers - European Commission," accessed January 22, 2025, https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13_en.

programming interfaces" (Paragraph 8).

54. While the DSA's Article 40 and the 'Delegated Act' hold great promise, its implementation and enforcement bring some uncertainty to researchers working on systemic risks and online safety. Depending on the territorial regulation of the access model, it could also create unfair competitive advantages to individual researchers, academic institutions and CSOs based in the EU and/or working on 'systemic risks' confined to the space of the EU. Only a coherent legislative effort that cuts across national boundaries can aspire to regulate inherently global platforms. The approval of a similar data access framework in the UK could be a decisive step in the right direction.

**Question 3c:** What are some possible models for providing researchers with access to relevant information that may not exist or be widely used yet, but which might be implemented by industry?

55. Platform providers could collaborate with research institutions or central repositories of public-interest data, such as the UK Data Service or Smart Data Research UK, to host data collection services of public social media platforms. Hosting and managing API access to vetted researchers could then be managed by the public entity, reducing the cost and staffing burden on the platform providers, who otherwise retain monetisation privileges.

**Question 3d**: What are the advantages and disadvantages of this approach?

- These may include elements pertaining to financial, legal, security, technical or feasibility issues

**Question 3e:** What role could third party organisations, such as regulatory bodies, civil society or public sector organisations have in facilitating researcher access to online safety information?

56. A data access framework based on a third-party model must guarantee the third party has full control over the process thus assuring accountability, independence, and transparency. This must avoid reproducing the flaws of previous third-party models, such as the ones either wholly owned by platforms (such as Crowdtangle) or disproportionately controlled by social media platforms (such as Social Science One). The failure of SSO and the demise of Crowdtangle show that a data access regime for independent assessment and oversight of social media platforms that acts in the best interest of the public must be independent of platform control over the questions to be asked.

57. A public data and information clearinghouse supported by the technical and human resources could be a way forward. The UK already has underway the technical work on the foundations of what such a clearinghouse might look like through UKRI and the National Data Library.

**Question 3g:** What categories of information should online service providers give researchers access for the study of online safety matters? Why would this information be valuable for the study of online safety matters?

58. Creating an online safety regime in the UK will require some other types of research question that, on first glance, might not be directly focused on safety. For example, studies on the dynamics of information sharing across different domains, platforms, and modes will be important for understanding misinformation.[32] Research on online contagion, for instance, can be about safety but also could be used to gain insights about the spread of online innovations, health behaviours, collective decision-making, etc.[33] Although not directly related to 'online safety', it is easy to understand how combining studies on health care behaviours and collective decision-making could provide vital information to design better health information campaigns. As researchers, we think once a trusted mechanism is in place for researcher access to data that includes reasonable ethical standards, public accountability and transparency, and requirements to make research public, then the type of questions asked becomes less important than the principle of beneficence that is applied.

59. An overly restrictive definition of 'online safety' could hamper good research in the public interest. For example, it is important to consider both 'societal harms' (e.g. the spread of harmful narratives undermining democratic processes) and 'individual harms' (e.g. self-harm content or other material that could lead to direct harm of an individual). We have previously argued that a 'network' approach to studying mis/disinformation networks is more effective at understanding how harms spread and evolve online than any 'case-by-case' or single-platform study, but this would require broader data access frameworks to allow researchers to map issues across platforms, languages, and contexts.

---

[32] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg, "Differences in the Mechanics of Information Diffusion across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter," in *Proceedings of the 20th International Conference on World Wide Web* (WWW '11: 20th International World Wide Web Conference, Hyderabad India: ACM, 2011), 695–704, https://doi.org/10.1145/1963405.1963503.

[33] Damon Centola, *How Behavior Spreads: The Science of Complex Contagions* (Princeton, New Jersey Oxford: Princeton University Press, 2020).

[End].