POLICY FORUM

DATA ACCESS

EU and US legislation seek to open up digital platform data

Constraints on data access must be addressed to facilitate research

By **Brandie Nonnecke**[1] *and* **Camille Carlton**[2]

Despite the potential societal benefits of granting independent researchers access to digital platform data, such as promotion of transparency and accountability, online platform companies have few legal obligations to do so and potentially stronger business incentives not to. Without legally binding mechanisms that provide greater clarity on what and how data can be shared with independent researchers in privacy-preserving ways, platforms are unlikely to share the breadth of data necessary for robust scientific inquiry and public oversight (*1*). Here, we discuss two notable, legislative efforts aimed at opening up platform data: the Digital Services Act (DSA), recently approved by the European Parliament (*2*), and the Platform Accountability and Transparency Act (PATA), recently proposed by several US senators (*3*). Although the legislation could support researchers' access to data, they could also fall short in many ways, highlighting the complex challenges in mandating data access for independent research and oversight.

As large platforms take on increasingly influential roles in our online social, economic, and political interactions, there is a growing demand for transparency and accountability through mandated data disclosures. Research insights from platform data can help, for example, to understand unintended harms of platform use on vulnerable populations, such as children and marginalized communities; identify coordinated foreign influence campaigns targeting elections; and support public health initiatives, such as documenting the spread of antivaccine mis- and disinformation (*4*).

The "Facebook Papers," leaked by whistleblower Frances Haugen, gave unprecedented insight into that platform's opaque practices (*5*). But reliance on whistleblowers and leaked data is untenable. Researchers need lawful access to platform data appropriately scoped to advance scientific knowledge and evidence-based policy-making. Yet, how to do this responsibly and in compliance with relevant data privacy laws and regulations remains debated (*6*).

Platforms have made data available to independent researchers through public application programming interfaces (APIs); however, because platforms' sharing of data with independent researchers has been primarily voluntary, data access has often been unreliable, inconsistent, and incompatible with research needs (*6*). For example, research questions that require data unavailable through an API, such as impression data and demographics of those exposed to disinformation campaigns, must rely on research partnerships with a platform or circumvention methods such as web scraping or requesting data directly from users (*7*). These methods are often not ideal because they pose ethical and legal concerns and may result in collection of data that is limited in terms of scale, quality, and precision (*8*).

COMPETING INCENTIVES

Despite the societal benefits, researchers' access to platform data is becoming more difficult. A tension exists between private incentives to retain data for financial, reputational, and privacy reasons and public incentives to access data for scientific research and oversight (*1*). For example, platforms have pushed back against mandatory data disclosures, emphasizing legal requirements to ensure user privacy and protection of proprietary information (*6*). Lawmakers have countered that data access is justified because large platforms wield substantial and increasingly monopolistic control over information online, which poses risks to individual rights and collective well-being [see Recitals 53 to 58 of the DSA (*2*)]. For lawmakers, access to data for empirical research is seen as a necessary step in ensuring transparency and accountability.

Platforms' hesitancy to share data with researchers is not wholly unwarranted. A Facebook-approved research partnership with Cambridge Analytica resulted in scandal, a $5 billion fine by the US Federal Trade Commission (FTC) for privacy violations, and new requirements in an FTC consent order to implement comprehensive data privacy and security safeguards (*9*).

In the face of steep fines and unwanted oversight, platforms often justify restrictions on data sharing with independent researchers by claiming that there is a lack of clarity in data privacy legislation and regulatory obligations (*10*). One of the primary pieces of legislation invoked is the General Data Protection Regulation (GDPR) by the European Union (EU). Platforms claim that the GDPR lacks clear standards for data anonymization and pseudonymization and approved cross-border transfers of personal data (*10*).

Difficulties faced in the Social Science One initiative are a quintessential example. Despite Facebook's partnership with Social Science One, an initiative that seeks

[1]University of California, Berkeley, CA, USA.
[2]Center for Humane Technology, San Francisco, CA, USA. Email: nonnecke@berkeley.edu

to facilitate academic researchers' access to Facebook data for public-interest research, it took nearly 20 months for researchers to gain access to the data the platform had promised (*6*). Facebook attributed its delay to the need to establish data privacy mechanisms to maintain GDPR compliance. Researchers used the data for 2 years before Facebook admitted that the data had a serious flaw: It was only representative of approximately half of US users, those who had a detectable political leaning (*11*). Researchers' trust in future data they receive has been seriously diminished (*11*).

Researchers trying to collect data outside of API access and formal partnerships have faced fierce opposition. In August 2021, Facebook disabled accounts of researchers behind the NYU Ad Observatory, claiming that the researchers had inappropriately scraped users' data on targeted advertising (*8*). Although Facebook justified its actions as necessary to ensure data privacy compliance mandated in the FTC consent order, the FTC issued a letter rebuking Facebook's actions and reemphasized that the order "does not bar Facebook from creating exceptions for good-faith research in the public interest. Indeed, the FTC supports efforts to shed light on opaque business practices, especially around surveillance-based advertising" (*12*).

LEGAL MANDATES FOR DATA ACCESS

Platforms' resistance to share data demonstrates the need for clarity in how data privacy legislation and regulatory mandates should be interpreted (*6*, *11*). The European Data Protection Supervisor in January 2020 reemphasized that the GDPR seeks to support data access for research and provided additional guidance on data governance for research purposes (*13*). The European Commission established a "Code of Practice on Disinformation" to support researchers' access to data, and the European Digital Media Observatory is establishing a framework for GDPR-compliant data access (*14*). Although these efforts are critical, platforms likely will not make data available without binding legal mechanisms. Legislation proposed in the EU and United States that seeks to mandate platform data access for research and oversight could thus be transformative.

The Digital Services Act

In light of the role played by "very large online platforms" (VLOPs; having at least 45 million active users in the EU) in "facilitating the public debate and economic transactions," the DSA would compel VLOPs to conduct assessments of systemic risks stemming from their services, such as dissemination of illegal content; impacts on fundamental rights; and the "intentional and, oftentimes, coordinated manipulation of the platform's service, with a foreseeable impact on health, civic discourse, electoral processes, public security, and protection of minors" (*2*). Once identified, platforms must implement appropriate risk mitigation strategies (*2*). These assessments and mitigation strategies would be auditable and may require platforms to make data available to a "Digital Services Coordinator," an independent authority established in each member state; to the European Commission; or to "vetted researchers" to support transparency, accountability, and compliance with relevant laws and regulations (*2*). The Act defines "vetted researchers" as individuals with an affiliation with an academic institution, independence from commercial interests, proven subject or methodological expertise, and the ability to comply with data security and confidentiality requirements (*2*).

The DSA requires platforms to make three categories of data available through online databases or APIs (*2*): (i) Data necessary to assess risks and possible harms brought about by the platform's systems; (ii) data on the accuracy, functioning, and testing of algorithmic systems for content moderation, recommender systems, or advertising systems; or (iii) data on processes and outputs of content moderation or of internal complaint-handling systems. In response to increased awareness of the risks of targeted advertising, Article 63 explicitly requires VLOPs to create a public digital ad repository that must include the ad's content; the entity behind the ad; whether it was targeted and, if it was, the parameters used for targeting; and the total number of recipients.

The DSA also stipulates that the European Commission, in consultation with the "European Board for Digital Services" established by the DSA, is tasked with adopting derivative acts that will determine the "technical conditions" for GDPR-compliant data sharing (*2*). A primary challenge will be to determine how data should be constructed and shared with researchers in ways that are GDPR-compliant while maintaining enough detail to make data useful for research. The DSA provides protections to platforms from having to share data that may pose a security or financial risk, such as trade secrets. Transparency in how VLOPs use this protection will be important to mitigate exploitation.

Platform Accountability and Transparency Act

PATA is the most comprehensive law proposed in the United States to require large platforms (having more than 25 million monthly users) to make data available to support scientific research and oversight (*3*). Proposed in December 2021, the Act compels platforms to make data available to "qualified researchers" through a process intermediated by the National Science Foundation (NSF) and the Platform Accountability and Transparency Office (PATO) to be established within the FTC (*3*).

The Act defines a "qualified researcher" as university-affiliated and establishes a process by which researchers are granted access to platform data (*3*). First, all projects must receive ethics approval from the institutional review board at the researcher's affiliated institution, followed by approval from the NSF. Then, in collaboration with the researchers, the NSF determines what platform data and information is necessary to carry out the research. Last, the research project is referred to PATO, which brokers data access between the platform and the researcher. PATO is also responsible for establishing privacy and cybersecurity safeguards for platform data and information provided to researcher(s) (*3*).

The Act establishes provisions to mitigate potential harms of data sharing and better ensure platforms' compliance. For example, researchers must comply with privacy and cybersecurity provisions and may only use the data for the specified research project. Those who intentionally violate the privacy and cybersecurity provisions will be subject to civil and criminal enforcement (*3*). Platforms that fail to comply with the Act may face financial penalties and lose immunity protections for user-generated content granted to them in Section 230 of the Communications Decency Act (*3*).

To support broader access to platform data, the FTC is granted authority to "require platforms to report on or disclose data, metrics, or other information" that will "assist the public, journalists, researchers, the Commission, or other government agencies" in assessing the impact of platforms on consumers, institutions, and society; promoting the advancement of scientific and other research; and ensuring compliance with federal law (*3*). When possible, the data and information must be made publicly available through a format that is "accessible and understandable to the public," such as a searchable database or API (*3*). The FTC may also require platforms to disclose the following on an ongoing basis: content that has been "sufficiently disseminated" (an unclear metric to be clarified by the FTC); content originating or spread by major public accounts (having at least 25,000 followers or at least 100,000 monthly viewers); and statistically representative samples of public content (*3*). The data and content must be accompanied by supporting information, such as dissemination and engagement data, audience characteristics, and whether the content was recommended or amplified by the platform's

algorithms (*3*). Like the DSA, PATA also compels platforms to report data and information on targeted advertising, such as the use and characteristics of algorithms used and content moderation tactics.

PATA also establishes a process to balance the private interests of platforms with the interests of the public. Researchers are required to submit a prepublication version of the research to PATO, and platforms are given authority to object to publication or release of any analysis that they believe does not comply with federal, state, or local privacy laws or risks disseminating confidential business information or trade secrets. Researchers are given the opportunity to amend their research or appeal. If the platform still objects, PATO makes the final determination. Platforms may thus be limited in their ability to block the release of unfavorable research findings under the guise of legal compliance or protecting trade secrets.

RECOMMENDATIONS

Although DSA and PATA are promising, several potential constraints merit further attention.

Broaden the scope

Both Acts have restrictions in scope that, if broadened, would better enable research on a wider breadth of topics from varied epistemological approaches. The DSA and PATA support data access for research but differ in their underlying intentions for doing so. The DSA primarily focuses on data access to support research that informs oversight and compliance with relevant laws and regulations, whereas PATA also seeks to enable general research. By scoping data access to only that which is necessary to ensure oversight and compliance with specific laws and regulations, the data to be made available may have limited usefulness for broader research. Legislation should compel platforms to disclose data for the purpose of advancing scientific knowledge. Research supports greater transparency and empirical understanding of platforms' effects, which is necessary for accountability and oversight.

Transparency is needed in how "vetted/qualified researchers" are selected, and mechanisms should be in place to ensure that certain institutions and disciplines are not disproportionately favored. We recommend expanding the types of researchers who are qualified to access platform data to include nonacademic researchers, such as journalists and others who aim to inform the public about critical matters. These individuals could also be "vetted/qualified" through a formal review process. Unlike the DSA, PATA proactively expands data access beyond academic researchers by creating a safe harbor for journalists' and other researchers' collection of platform data through web scraping; voluntary donation by users, including through browser extensions and plug-ins; and creation of research accounts. Without this protection, research and public awareness may be stifled out of fear of retaliation.

Infrastructure and intermediaries

Increasing the availability of platform data will have limited effects if there is not also equity-driven investment in research infrastructure. Research institutions such as the NSF or the European Research Council should increase support for research infrastructure to better ensure that a greater diversity of researchers across institutions and disciplines are equipped to store and analyze data in compliance with the legislation. Otherwise, the requirement that researchers must have the capacity to comply with data security and confidentiality requirements will likely favor larger institutions and disciplines with greater research infrastructure (such as computing power and cybersecurity).

To counter infrastructure constraints, lawmakers should also look to the data intermediary model to enable diverse researchers' access to secure, interoperable, cross-border data. Trusted intermediaries may be able to support development of robust data-sharing ecosystems by working with platforms, researchers, and users to establish data-sharing models that advance public-interest research while ensuring compliance with relevant laws and human rights norms.

Address methodological challenges

Platform data is often not collected with the intent that it will be used in scientific research. Its structure may be at odds with the goals of scientific inquiry (*4*). For example, a sample of network data may have characteristics that are wholly different from the entire network itself (*15*). Moreover, researchers are completely reliant on platforms to provide datasets free from omissions or edits. Any datasets made available to researchers should provide metadata and other contextual information, including how the data was generated and collected, effects of the platform's design (such as recommender system) on the data; how sampling was conducted; and, if applicable, how the data was cleaned, transformed, or modified before being shared. In doing so, data and research insights will be of higher quality and accuracy, better ensuring that any resulting oversight is appropriately scoped and effective—a benefit for both the public and the platforms.

The legislation will require robust anonymization efforts, such as the use of differential privacy, which may diminish the data's value for researchers (*4*). To address these challenges, partnerships between lawmakers, platforms, and independent researchers are necessary to better ensure that data are generated, collected, and made available in ways that are of high value for scientific inquiry and public accountability while maintaining legally compliant data privacy and security.

DEMAND FOR TRANSPARENCY

The European Parliament adopted the DSA during its January 2022 plenary. The Act will now go through subsequent negotiations between the EU Parliament, EU Council, and the EU Commission. PATA's fate is less certain. Although the bill has garnered bipartisan support, it has yet to undergo formal congressional debate. PATA builds on several bills that aimed to make platforms more transparent and accountable. However, it addresses many of the previous bills' shortcomings by providing a comprehensive strategy for how to compel platform data access for research and oversight. As such, PATA's chances of becoming law may be stronger. The emergence of these legislative efforts suggests that as large platforms increasingly play a centralized role in our social, economic, and political interactions, the demand for transparency and accountability mechanisms will continue to grow. ■

REFERENCES AND NOTES

1. D. M. J. Lazer *et al.*, *Science* **369**, 1060 (2020).
2. European Commission, "The Digital Services Act: Ensuring a Safe and Accountable Online Environment" (2021).
3. "Platform Accountability and Transparency Act" (2021); www.coons.senate.gov/download/text-pata-117.
4. D. Lazer *et al.*, *Nature* **595**, 189 (2021).
5. "Facebook Files: 5 things leaked documents reveal," *BBC News*, 24 September 2021.
6. N. Persily, J. A. Tucker, in *Social Media and Democracy: The State of the Field and Prospects for Reform*, N. Persily, J. A. Tucker, Eds. (Cambridge Univ. Press, 2020), p. 313.
7. I. V. Pasquetto *et al.*, *Harvard Kennedy School Misinform. Rev.* 10.37016/mr-2020-49 (2020).
8. M. Bobrowsky, "Facebook Disables Access for NYU Research Into Political-Ad Targeting," *The Wall Street Journal*, 4 August 2021.
9. J. J. Simons, N. J. Philips, C. S. Wilson, "Statement of Chairman Joe Simons and Commissioners Noah Joshua Phillips and Christine S. Wilson In re Facebook, Inc.," Federal Trade Commission (2019).
10. M. Vermeulen, "The keys to the kingdom," Knight First Amendment Institute at Columbia University (2021).
11. D. Alba, "Facebook sent flawed data to misinformation researchers," *The New York Times* 12 September 2021.
12. S. Levine, "Letter from acting director of the Bureau of Consumer Protection Samuel Levine to Facebook," Federal Trade Commission (2021).
13. European Data Protection Supervisor, "A Preliminary Opinion on Data Protection and Scientific Research," European Data Protection Supervisor (2020).
14. European Commission, "European Digital Media Observatory (EDMO)" (2021).
15. A. Gelman, *J. Surv. Stat. Methodol.* **5**, 22 (2016).

ACKNOWLEDGMENTS

We thank anonymous reviewers for comments. We thank C. Crittenden, D. Lazer, D. Mulligan, N. Persily, J. Reinhardt, and V. Vaidyanath for their initial reviews and guidance.

10.1126/science.abl8537

EU and US legislation seek to open up digital platform data

Brandie NonneckeCamille Carlton

Science, 375 (6581), • DOI: 10.1126/science.abl8537