Policy Memos

Canadian Commission on Democratic Expression

Learning Session 1: What data should be shared and with whom?

Thursday, Oct. 7, 2021 | 1:00 p.m. – 2:30 p.m. ET (UTC -4:00)
Virtual event via Zoom

Abstract of session

This session seeks to address a fundamental democratic challenge of our online ecosystem: opacity. Digital infrastructure is comprised of automated systems and algorithms, economic models, and vast amounts of data which structure many integral parts of our public life. This includes what content we see online, whether we consume reliable information and news, whether we are targeted with hate speech and abuse, or with deliberately deceitful content designed to influence our vote and public opinion, and if our personal information is used for financial gain. These processes remain hidden by "black box" systems and industry confidentiality.

Not only does opacity undermine fairness and due process for those directly impacted, but it also poses grave challenges when attempting to regulate known harms such as disinformation campaigns through advertising or the use of discriminatory algorithms in high-stakes public sectors. While many platforms report on their activities with regards to specific content, those reports are entirely self-regulated and cannot be independently verified. Platforms also largely resist releasing their data or opening their algorithms for inspection. The lack of transparency makes it near impossible for governments to verify if the companies are enforcing their own policies or complying with the law. Moreover, limited rules around data preservation mean that that there is little oversight around what data is stored in the first place. Unsurprisingly, governments and civil society actors around the world are calling for open data access for researchers, governments and the public, and online platforms' preservation of data with human rights implications.

Policy questions:

Are existing voluntary transparency mechanisms sufficient and effective?
What data should be shared with regulators, researchers, and the public?
What information (technical and non-technical) should a regulator ask for in order to understand how the design of our digital infrastructure shapes the nature of our public sphere?

6 October 2021

Policy Brief on Platform Data Access

Digital platforms such as Facebook, Google, Instagram, TikTok, Twitter, and YouTube shape and structure our daily lives—the information we consume, our social interactions, and much more—in profound ways. They have become essential to social, economic, and political life. And yet we know relatively little about the precise impacts—either good or bad—these platforms have on individuals, social groups, and society as a whole. This is because the most basic element necessary for observing and analyzing such impacts, platform data, remain within walled gardens, accessible only to the companies themselves. Though civil society organizations, journalists, and academic researchers have diligently worked with what limited information they are given or can themselves collect, ultimately vital data remain inaccessible. This in turn allows the platforms to obfuscate, releasing "transparency reports" and internal research findings on their own terms and without context, doling out bits of data when and how it suits them, and persistently critiquing independent researchers' findings for being based on incomplete information—information the platforms refuse to provide.

Independent researchers have tried numerous approaches to resolving this dilemma on their own or with voluntary cooperation from the platforms. None have worked. The platforms frequently "break" independent researchers' data collection efforts through technical means or simply shut down those researchers' accounts. And the largest academic-industry research partnership, Social Science One, has been beset by delays, broken promises, and massive errors in the data provided to researchers. Ultimately, the substantial power imbalance between independent researchers and digital platforms means that researchers have little recourse when the platforms interfere with their work.

To rectify this situation, I believe we need new regulations that take three interrelated steps: (1) mandates access to platform data for vetted independent researchers, (2) establishes standards and protocols for privacy-protecting, ethical, and responsible access to this data, and (3) allows verification and validation of the accessed data via a system of independent audits. I briefly discuss each below.

(1) Mandatory data access
    Mandates for data access should be tied to specific objectives (e.g., the performance of independent risk assessments, the evaluation of platforms' impacts in particular areas of policy focus) but broad enough to allow the independent researchers to define what precise forms of data will be required to carry out the research in question. It is difficult, if not impossible, to define what data will be needed until a specific research task is defined.

(2) Standards and protocols for responsible data access
    Before data access mandates can be successfully implemented, technical standards and various logistical protocols must be developed and agreed. The platforms, for example, each hold a variety of data with different structural properties. Whether transferring data directly to

researchers or designing systems that allow researchers to access data within the platforms' infrastructure (e.g., within so-called data "clean rooms"), complex technical questions must be resolved to ensure that data can be accessed in ways that are both efficient and privacy-protecting. In addition, important questions regarding what parties (companies *and* researchers) should be covered by any applicable framework, need to be addressed, and processes must be developed for authorizing participation, monitoring compliance, assigning and distributing liability, and enforcing when violations occur. In short, significant work is required to ensure that any measures that mandate data access can be implemented in practice and subsequently enforced.

(3) Verification and validation of data
Finally, any data provided by the platforms for the purposes of independent analysis must be verified as complete (or if sampled, representative), error-free, and appropriate for the intended research. This will require a two-step process, whereby (a) the platforms carefully document how the data were selected and prepared and (b) random audits are imposed that permit examination of (i) relevant internal documentation, (ii) the computational code (or "pipelines") used to prepare, transform, and deliver the data, and (iii) tests run using synthetic data generated by the auditor.

Rebekah Tromble
Director, Institute for Data, Democracy & Politics
Associate Professor, School of Media & Public Affairs
George Washington University
Washington DC, USA

Unlocking and Protecting Evidence-Based Digital Policy

J. Nathan Matias, Citizens and Technology Lab

How can we protect people from digital harms and advance existing social policies when so much of people's lives are mediated by digital technologies? The science of prevention can provide a trustworthy evidence-base for industry accountability and policy evaluation. But evidence-based governance will continue to be nearly-impossible without policies that support and protect industry-independent research.

Prevention vs Response: Current digital policies tend to focus on content moderation regimes that respond to harms rather than prevent them (Bradford et al 2019; Gillespie 2017). Governments and technology platforms should also prioritize interventions that prevent harms, not just wait to act until after they happen (Ko et al 2017).

Evidence-Based Policy: Current transparency metrics reward platforms for responding more quickly to more and more harms rather than reducing those harms. Prevention-focused research can estimate the harms prevented by a given policy (Matias 2019). Governments and platforms could choose policies on the basis of which ones would prevent more harms.

Industry-Independent Research: Since companies are incentivized to withhold evidence when it benefits them, society needs high-quality, independent research that can inform policy decisions and hold platforms accountable (Matias 2020).

Policy recommendations:

- Compel platforms to cooperate with open, independent research, including research that evaluates harm prevention policies
- Prohibit platforms from creating policies or taking actions that forbid responsible independent research
- Expand the capacity of academia, civil society, and journalism to conduct industry-independent research

References

- Bradford, B., Grisel, F., Meares, T. L., Owens, E., Pineda, B. L., Shapiro, J. N., ... & Peterman, D. E. (2019). Report of the Facebook Data Transparency Advisory Group. Yale Justice Collaboratory.
- Gillespie, T. (2017). Governance of and by platforms. *SAGE handbook of social media*, 254-278.
- Ko, A., Mou, M., Matias, J.N. (2017) The Obligation to Experiment: Tech companies should test the effects of their products on our safety and civil liberties. MIT Media Lab.
- Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116*(20), 9785-9789.
- Matias, J.N. (2020) Why We Need Industry-Independent Research on Tech & Society. Citizens and Technology Lab

About Dr. J. Nathan Matias

Dr. J. Nathan Matias organizes citizen behavioral science for a safer, fairer, more understanding internet. Dr. Matias is an assistant professor in the Cornell University Department of Communication and founder of the Citizens and Technology Lab.

CAT Lab has worked with communities of tens of millions of people to test ideas for preventing harassment, broadening gender diversity on social media, responding to human/algorithmic misinformation, managing political conflict, and auditing social technologies.

Before Cornell, Matias was an associate research scholar at Princeton University at the Center for Information Technology Policy. Matias completed his Ph.D. at the MIT Media Lab and spent several years as a fellow at Harvard's Berkman Klein Center for Internet and Society.