Written Testimony of Alan Mislove
Professor and Interim Dean
Khoury College of Computer Sciences, Northeastern University

Before the
United States House Committee on Science, Space, and Technology
Subcommittee on Investigations & Oversight

Hearing entitled
The Disinformation Black Box: Researching Social Media Data

Tuesday, September 28, 2021
via Zoom

Chairman Foster, Ranking Member Obernolte, and distinguished members of the Subcommittee, thank you for the opportunity to appear before you today to discuss accessing and analyzing data from social media companies.

My name is Alan Mislove. I am a Professor and Interim Dean of the Khoury College of Computer Sciences at Northeastern University. My research is on *algorithmic auditing*; I develop methodologies that allow me to study large online platforms—such as those operated by social media companies—to better understand how they work, how they may be abused, and what impacts they are having on end users. I conduct my research independently: Without companies' permission, and without insider access to data. Put simply, I have no more access to these platforms than any of you do. This is a significant challenge: It is difficult to develop the techniques that enable my work, especially because companies are resistant to external accountability and we work in a legal environment makes such research carry non-trivial risk.[1]

With that context, as Congress is considering legislation in this area, I am well-positioned to provide input on what currently can be measured about social media platforms, and what is needed going forward to ensure researchers fully understand the impact that these platforms are having on society at large. In particular, there are four messages I wanted to convey in response to the questions the subcommittee posed:

1. Independent research is *critical* to uncovering and addressing platforms' societal impact

Today, social media platforms mediate an increasingly large fraction of online communication, ranging from interpersonal communication to political messaging, from news dissemination to access to life opportunities. Even in the best of worlds, understanding how these platforms are impacting end users and society is too big a task for the platforms themselves. In reality, despite the critical role they play in our society, social media platforms are frequently non-transparent about the content that is shared on their platforms, as well as the advertising systems that fund their operations.

Despite these challenges, my group and collaborators have been successful at studying a variety of such platforms, identifying alarming behaviors, and working with platforms to make improvements. We

---

[1] *Sandvig v. Sessions — Challenge To CFAA Prohibition On Uncovering Racial Discrimination Online.* `https://www.aclu.org/cases/sandvig-v-sessions-challenge-cfaa-prohibition-uncovering-racial-discrimination-online`.

have shown how popular e-commerce platforms often use techniques such as price discrimination to maximize profit at users' expense[2]; how online "gig economy" hiring platforms can rank candidates in ways that disadvantage women[3]; how ride-sharing services were calculating "surge" prices[4] and how they are impacting legacy taxi services[5]; and how social media platforms' algorithms can produce large distortions in the racial and gender makeup of who sees ads, often without the advertiser's input or even awareness.[6] In the coarse of doing this research, we identified bugs in Uber's surge pricing algorithm[4] and privately disclosed to Facebook multiple ways in which malicious advertisers could leak private user data.[7] In both cases, the platforms fixed the issues we identified.

The upshot is that independent, third-party research is critical to fully understand how social media platforms are impacting end users and society as a whole.

2. Today, independent researchers' access to data is *woefully* inadequate

Today, the kinds of questions that researchers are able investigate are severely hampered by researchers' limited access to platforms, as well as the platforms' choices about what to make available. Typically, studies are based on running experiments directly on the platform and collecting data, on recruiting cohorts of real users who agree to share their data, or on analyzing the aggregate data that the platforms provide.

When we run our own experiments, because every platform is different, we must typically spend significant time to understand what kind of data we can get out of the platform, and whether that data is useful scientifically. For example, when studying Facebook's advertising platform, we ran our own ads and measured how they are being delivered using Facebook's interface that tells advertisers how their ads are performing. For our experiments alone, we spent over a year understanding how the ad platform can be used to measure the properties of the underlying relevance algorithm, and spent over $25,000 actually running ads.

This approach to studying platforms requires deep computing expertise to develop, is expensive in both money and time, and is not scalable to address the impact that platforms are having. Worse, the answers we can find with it and conclusions we can draw are often relatively small in scope, relative to the impact that platforms are having on society. For example, we can often only comment on how *our own content* is treated, making it difficult to understand what is happening to others'. This means we can often show that the ad platforms' algorithms have certain properties, but it is much more challenging to understand the degree to which those properties impact real-world ads and end users.

When we recruit cohorts of real users, we often ask users to install software such as browser extensions or mobile applications in order to collect data automatically. These users are typically compensated, and

[2]Aniko Hannak et al. "Measuring Price Discrimination And Steering On E-commerce Web Sites". In: *ACM Internet Measurement Conference*. Vancouver, Canada, Nov. 2014.

[3]Aniko Hannak et al. "Bias In Online Freelance Marketplaces: Evidence From Taskrabbit And Fiverr". In: *ACM Conference on Computer Supported Cooperative Work*. Portland, Oregon, USA, Feb. 2017.

[4]Le Chen, Alan Mislove, and Christo Wilson. "Peeking Beneath The Hood Of Uber". In: *ACM Internet Measurement Conference*. Tokyo, Japan, Oct. 2015.

[5]Shan Jiang et al. "On Ridesharing Competition And Accessibility: Evidence From Uber, Lyft, And Taxi". In: *International World Wide Web Conference*. Lyon, France, Apr. 2018.

[6]Muhammad Ali et al. "Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead To Biased Outcomes". In: *ACM Conference on Computer Supported Cooperative Work*. Austin, Texas, USA, Nov. 2019.

[7]Giridhari Venkatadri et al. "Privacy Risks With Facebook's PII-based Targeting: Auditing A Data Broker's Advertising Interface". In: *IEEE Symposium on Security and Privacy*. San Francisco, California, USA, May 2018.

we go to great lengths to ensure that they understand the experiment and that their privacy is protected. However, this methodological approach has a number of drawbacks: it is difficult to recruit a diverse and representative population of users, it is expensive to do so at large scale, it is technically challenging to collect data when users are on mobile devices, and platforms often use technical and legal means to attempt to prevent data collection in this manner despite these projects having user consent.[8,9,10]

When we rely on the aggregate data that platforms themselves provide, researchers face challenges of first obtaining access to the data (which is not always possible), determining whether the data is scientifically useful, and putting faith in platforms that the aggregation was done correctly. For example, Facebook's Ad Library ostensibly shows, for political ads, "a range of how much they spent, and the reach of the ad across multiple demographics".[11] However, this data is limited in significant ways: the amount spent on an individual ad is provided only in coarse-grained ranges, the demographic breakdown of who actually saw ads is provided only at a superficial level, and no targeting information is revealed. The situation is worse for non-political ads, where none of this information is available at all. Similarly, Facebook's Open Research and Transparency (FORT) initiative provides targeting information for political ads from a particular three-month period, but omits information on ads with fewer than 100 impressions[12] (likely a large fraction of the ads during that period, making it difficult if not impossible to answer most scientific questions). Worse, this data set is not public, requires approval from Facebook to be able to access, and cannot be shared with other reseachers.

When relying on platform-provided aggregate data, researchers must further trust that the platform correctly aggregated the data. This trust may be misplaced, as evidenced by a recent incident in which Facebook's data provided to the Social Science One initiative left out data from *half* of U.S. users[13], calling many previously-published studies that relied on the data into question. Finally, platforms have often under-invested in even the simple data transparency tools they do provide. For example, the Facebook Ad Library has been shown to have numerous reliability issues[14] that has made it difficult to use it as a basis for scientific work. Similarly, another popular data tool, CrowdTangle, is reportedly in the process of being broken up by Facebook.[15]

3. Major platforms are actively attacking independent researchers' ability to do their work

Researchers today are effectively relying on platforms' "good will" to allow studies to be run at all—a situation that is becoming less and less tenable as platforms become more entrenched. In some cases, platforms refuse to many available data that would enable necessary research. To wit, Facebook recently criticized[16] a study on misinformation by saying it focused on who engages with content and not who sees

---

[8] *Facebook Seeks Shutdown of NYU Research Project Into Political Ad Targeting*. https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533.

[9] Engadget. *Researchers shut down Instagram study following backlash from Facebook*. https://www.engadget.com/algorithmwatch-facebook-shutdown-184729493.html.

[10] The Markup. *Facebook Rolls Out News Feed Change That Blocks Watchdogs from Gathering Data*. https://themarkup.org/citizen-browser/2021/09/21/facebook-rolls-out-news-feed-change-that-blocks-watchdogs-from-gathering-data.

[11] *What is the Facebook Ad Library and how do I search it?* https://www.facebook.com/help/259468828226154?helpref=search.

[12] *Facebook Open Research & Transparency*. https://fort.fb.com.

[13] The Washington Post. *Facebook made big mistake in data it provided to researchers, undermining academic work*. https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/.

[14] The New York Times. *Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised*. https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html.

[15] The New York Times. *Inside Facebook's Data Wars*. https://www.nytimes.com/2021/07/14/technology/facebook-data.html.

[16] The Washington Post. *Misinformation on Facebook got six times more clicks than factual news during the 2020 election, study says*. https://www.washingtonpost.com/technology/2021/09/03/facebook-misinformation-nyu-study/.

it—but that's only true because Facebook does not make such impression data available to researchers. While research may sometimes be going on inside companies—where researchers presumably do have access to such data—public relations concerns are paramount and can prevent important findings from being shared. For example, Facebook reportedly recently blocked the publication of a report[17] that indicated the top-performing story on their platform was, in fact, COVID-19 misinformation.

In other cases, social media platforms are actively hostile to third-party research, and sometimes even going as far as blocking researchers and threatening legal action. For example, just in the past few weeks we have observed Facebook block the accounts of New York University researchers[8] and threaten the European group Algorithm Watch[9] with litigation, both for running independent, ethical research in the public interest.

4. Mandating transparency requires nuance, but is feasible and urgent

Social media platforms sit inside broader social-technical systems, and the data made available to researchers must be comprehensive enough to recognize the complexity of such systems. However, addressing these challenges is feasible, and is becoming increasingly urgent.

To understand why comprehensive data is needed, consider the recent debate over ad *targeting options*, which can enable advertisers to limited their ads shown to heavily biased—and potentially discriminatory—groups of users. One might think limiting advertisers' ability to use various targeting options would address this concern; however, our work has demonstrated that doing so can actually cause ads to be shown *more* biased groups. The reason is that not all users targeted by the advertiser actually see the ad: Platforms typically have *relevance algorithms* that select the subset of the targeted users to whom content is most relevant. Thus, by removing targeting options, the relevance algorithms tend to have greater leeway in choosing which users actually see the ad as the targeted groups tend to be larger.[18] In this example, for researchers and policy to fully understand the impact of relevance algorithms—as well as to develop effective policy mitigations—platforms need to provide data on *both* the ad's targeted audience as well as the actual delivery audience.

Different kinds of content require different kinds of data. It is important to note that social media platforms typically allow sharing of a variety of different types of content. For example, Facebook alone allows "organic" posts (which can contain mixes of text, images, videos, etc); aggregate content including pages, events, and groups (themselves containing users, posts, and other data); ads (each containing creative content, targeting information, external links); and many others. Because each kind of content has different features and attributes, a one-size-fits-all approach to making data available is unworkable. Instead, when platforms are required to make data available, the kind of data that is released should be tailored to the particular type of content.

Transparency over both "what" and "who" is crucial to understand platforms' impact. Specifically, existing platform-provided transparency mechanisms have focused primarily on making data available on the (popular) content that is being shared, while there has been much less emphasis on data about *who* this content is being shown to. Making aggregate data on the demographics of who is being shown

---

[17]The Washington Post. *Facebook says post that cast doubt on covid-19 vaccine was most popular on the platform from January through March.* https://www.washingtonpost.com/technology/2021/08/21/facebook-coronavirus-vaccine/.

[18]For example, when we targeted an ad for jobs in the janitorial industry to an equally gender- and race-balanced audience on Facebook, it was delivered disproportionately to Black women (compared to jobs in the lumber industry, which were delivered disproportionately to white men).

content is critically important, as it is necessary to be able to understand platforms' impact on end users. Unfortunately, without the requirement to do so, platforms have been resistant to releasing such data: for political ads only, Facebook reveals demographic breakdowns for only the users who saw the ad, whereas Google reveals demographic breakdowns for only the targeted users.

Finally, while ethical concerns exist over social media platform data, these have successfully been addressed historically, and existing approaches could be used directly when sharing this information. Moreover, scientific communities are used to addressing the ethical implications of scientific research, and have processes in place to ensure protection of human subjects and that research meets community ethical standards.

To conclude, social media platforms do not currently have the proper incentives to allow research on their platforms, and have been observed to be actively hostile to important, ethical research that is in the public interest. At the same time that such platforms' power and influence is reaching new heights, our ability as independent researchers to understand the impact that they are having is being reduced each day. Thus, I and other researchers need Congress's help to enable researchers to have sufficient access to data from social media platforms in order to ensure that the benefits of these platforms do not come at a cost that is too high for society to bear. In particular, proposed legislation such as the Algorithmic Justice and Online Platform Transparency Act of 2021 and the Social Media Disclosure And Transparency of Advertisements (DATA) Act of 2021 both take meaningful steps towards ensuring researchers continue to have sufficient access to such data.

Thank you again for giving me the opportunity to appear before you at today's hearing. I look forward to your questions.