What would you think are the key issues for scholars and journalists to know right now about access to data?

One of the most important subjects of research at the moment is our online world, and researchers and journalists are trying to educate the public about what our new online sphere portends for society. Is public discourse online being manipulated or distorted by the platforms that host and curate conversations online, or by malicious third parties who try to hijack conversations or hijack the machinery of content moderation to advance their agenda at the expense of the public interest?

But it is extremely difficult to study those questions because you need data to study them, and the platforms, for their part, tend to be pretty stingy in the data they make available to researchers and journalists. And so there are many, many journalists and researchers now who try to study the platforms independently, by acquiring data through their own means, either by recruiting volunteers to their studies to enable research, or by collecting data directly themselves for their studies.

The problem is that the platforms tend to view those kinds of activities as violations of their terms of service; they tend to view them as illegal. And the platforms, and I’m mainly thinking of Facebook here, have threatened researchers and journalists who engage in these public interest investigations with cease-and-desist letters, threats of litigation, with the effect of shutting down some of these individual research projects or causing them to change in ways that make them less useful to the public.
Maybe even more troubling is the fact that these threats cast a broader chill on the research community. There are many researchers who don’t take up these sorts of investigations because they don’t want to have to worry about a corporate behemoth like Facebook sending them a legal threat and a legal demand that they don’t have the resources or the institutional backing to fight. So that, to my mind, is one of the most important areas where we need research and study, but where corporate terms of use have created an enormous blind spot for humanity.

**Have you been keeping up with any of the legislation that’s been put forward about platform transparency and safe harbors for research?**

Yeah, absolutely. We’ve been following the legislation closely. We actually have drafted some model legislation ourselves, and the model legislation that we drafted was incorporated recently, almost verbatim, into one of the draft bills that’s been under discussion. There are a variety of ways you can tackle the problem of platform transparency, and the way that we were focused on was trying to enable independent research, because we view independent research as especially important. Even in a world where you have data provided by platforms to researchers to study what’s happening on their sites, you still need independent research to verify that that data actually reflects what’s going on, on the platforms. And you need to know that this data isn’t modified in any way, intentionally or not, in a way that makes the research less valuable or less reliable.

In 2018, we proposed to Facebook that it adopt a safe harbor to its terms of service that would enable researchers and journalists to study the platform and engage in digital investigations without worrying about liability for violating terms of service. We negotiated with the platform for 18 months, and they eventually rejected this idea, and so we converted it into a legislative safe harbor. If Facebook wasn’t going to agree to this amendment to its terms of service, maybe Congress or some other legislature should just mandate it. And so, we drafted a legislative safe harbor that would immunize researchers and academics and journalists from liability for studying the platforms independently, so long as they respected user privacy. And we shared a draft of this with staffers on the Hill, and it ultimately made its way into the bill proposed by Senators Coons, Klobuchar, and Portman, in the Platform Accountability and Transparency Act. That bill includes a number of other provisions. It includes a provision that Professor Nate Persily worked on that would mandate access by vetted researchers to data held by the platforms. And that would also require the platforms to publish, affirmatively, certain categories of data, including advertising information, information about advertisements on the platform. So, we’ve been closely following legislation, and that bill, or others like it, might be a significant improvement. Whether there’s political will to pass those laws is anybody’s guess.

**Is there anything else you want to add?**

One of the biggest challenges to platform transparency is how to achieve it in a way that respects user privacy, that accounts for the fact that much of the data necessary to study the platforms is extremely sensitive. Some of it is private, and even the data that is public can be very sensitive when aggregated and analyzed using modern analytic techniques.
I think the reality is that it’s essential that we figure out how to allow for the sharing of this sort of data in a way that respects user privacy, and we’ve done it in other contexts. It’s not a new phenomenon that society might need to rely on private or sensitive data to advance scientific understanding. We do that in the biological sciences all the time. We have an entire legislative infrastructure designed to protect patient privacy while enabling medical research; it’s called HIPAA. It’s our medical privacy law in the United States, and it allows for this research, even though the data that researchers collect is some of the most sensitive data that we have about ourselves. And the platforms have shielded themselves so far from that same sort of scrutiny by cloaking themselves in the very important value of user privacy, but without recognizing that we need to be able to balance user privacy against the need for research.

**Do you think a model like HIPAA would work in the case of protecting user data?**

I think it could, yeah, absolutely. And our legislative safe harbor adopts that sort of model, that it would impose strict limitations on what kind of data could be collected and what purposes it could be used for. Professor Persily’s proposal [within the Platform Accountability and Transparency Act] would include pretty strict limits on how that is used, and if I’m remembering correctly, it would even include pretty significant punishments for misuse of data obtained pursuant to his bill.

I think we should study those models, because this problem of needing access to private data for the public good is not a new one. It’s one that we’ve tackled before, but this is a new context, so we have to modify our old frameworks to account for the new context. But it is absolutely possible, and it is just ironic that the platforms have managed to use user privacy as a shield from those sorts of efforts, especially because independent research is responsible for so much of what we know about the platform’s own abuses of user privacy. They are one of the major threats to user privacy, and so it just can’t be that the interest in user privacy should prevent researchers and journalists from having access to data that we need.