Statement of Yoel Roth, PhD
Former Head of Trust & Safety
Twitter, Inc.

Hearing on "Protecting Speech from Government Interference and Social Media Bias, Part 1: Twitter's Role in Suppressing the Biden Laptop Story"

Before the House of Representatives Committee on Oversight and Accountability

February 8, 2023

Thank you, Chairman Comer, Ranking Member Raskin, and Members of the Committee, for the opportunity to speak with you here today.

In nearly 8 years at Twitter, I worked in and led a division called Trust & Safety. Trust & Safety's core duty is content moderation: Labeling or removing Tweets that violate Twitter's terms of service, and suspending or banning users who repeatedly break the rules. This work is sometimes dismissed merely as censorship — but it represents a key way that Twitter and other companies live up to their responsibility to keep the users of their products safe.

Much of this work is uncontroversial and obviously good: For example, taking down accounts that engage in child sexual exploitation or promote terrorism. Although this content represents a tiny fraction of the overall volume of conversation on social media, the dangers it poses means platforms like Twitter have a responsibility to find it and remove it promptly. The scale of this work is considerable: In the second half of 2021, Twitter removed over 33,000 accounts for promoting terrorism or violent extremism, more than 100,000 for promoting the sale of illegal goods and services, and nearly 600,000 for engaging in child sexual exploitation.[1]

If all trust and safety work was just about universally abhorrent content, there would be no controversy. We would all agree that content moderation is unambiguously good. The gray area of this work is when trust and safety teams have to make decisions about so-called "lawful but awful" material: Content that may be legal in many jurisdictions, but isn't something most people would want to see or experience.[2] Think of things like posting someone else's home address without their permission; or sharing graphic videos of animal abuse; or bullying someone for a disability or for how they look.

---

[1] "Rules Enforcement, July - December 2021," Twitter Transparency Center, Twitter, last modified July 28, 2022, https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec.

[2] Daphne Keller, "Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users," *The University of Chicago Law Review Online*, June 28, 2022. https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech/.

A free-speech absolutist might say, "Yes, that kind of content is unpleasant, but it's not against the law. What right do you have to remove it?" The answer is that, as businesses, social media platforms must be appealing to their own users, if they hope to survive.[3] Consistently, in its own research, Twitter found that users were unhappy with the platform's approach to content moderation — and that this dissatisfaction drove away both users and advertisers. In other studies, we found that some portions of Twitter's user base stopped using Twitter for the opposite reason: Because they felt Twitter was too overbearing and censored too much. In short, the company had to look for a middle ground solution that would appeal to both sides of this spectrum, while still addressing the very real harms that can come from social media.

A key harm we talked a lot about on the Trust & Safety team was chilling effects: The idea that one person's unrestricted freedom of speech could have the consequence of stifling someone else's — the proverbial "Heckler's Veto" of First Amendment law. Again and again, we saw the "lawful but awful" speech of a small number of abusive users drive away countless others. Unrestricted free speech, paradoxically, results in less speech, not more.[4] Trust & Safety's job is to try to find a balance between one person's free speech, and the impacts of their free speech on the ability of others to participate. Getting this right is difficult, but essential to the success of a platform like Twitter.

But the importance of trust and safety work goes far beyond whether or not Twitter succeeds as a private business. There are broader national security implications for this work, too. In 2016, we saw significant interference in an American election by the Russian government, through social media platforms such as Twitter.[5] I led the team at Twitter that uncovered that interference.[6] I still remember the rage I felt when I saw accounts with names like "Pamela Moore" and "Crystal Johnson" — accounts purporting to be real Americans, from Wisconsin and New York, but with phone numbers tracing back to St Petersburg, Russia.[7] These accounts were operated by agents of a foreign government, and their mission was to stoke culture war issues on social media to try

[3] Yoel Roth, "I Was the Head of Trust and Safety at Twitter. This Is What Could Become of It," *New York Times*, November 18, 2022, https://www.nytimes.com/2022/11/18/opinion/twitter-yoel-roth-elon-musk.html.
[4] Danielle Citron, *Hate crimes in cyberspace* (Harvard University Press, 2014). Mary Anne Franks, "Fearless speech," *First Amendment Law Review* 17 (Symposium, 2018).
[5] Kathleen Hall Jamieson, *Cyberwar: How Russian hackers and trolls helped elect a president: What we don't, can't, and do know* (Oxford University Press, 2018).
[6] Twitter, "Update on Twitter's review of the 2016 US election," *Twitter Blog*, January 19, 2018, https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html
[7] U.S. Department of Justice, *Report On The Investigation Into Russian Interference In The 2016 Presidential Election, Volume I of II*, March 2019, https://s3.documentcloud.org/documents/5955240/Full-Mueller-Report.pdf.

to further divide Americans.[8] My team and I exposed and banned hundreds of thousands of these accounts, from Russia, but also from Iran, China, and beyond.[9]

And the kinds of attacks that we saw go far beyond fake Americans with Russian phone numbers. The actors targeting American elections are well-funded and increasingly sophisticated, and have continuously gotten better at covering their tracks. In 2016, Russian military intelligence carried out a sophisticated hack and leak campaign targeting the US elections; the details of it weren't declassified until long after election day[10], after the damage had been done. As Congress investigated what happened, a clear finding was that tech platforms and law enforcement had failed to appropriately work together to address these threats.[11] Twitter, and other companies, were widely — and I think rightfully — criticized for their inaction. We were told in no uncertain terms, by the public and by Congress, that we had a responsibility to do a better job protecting future elections.

In an effort to get ahead of these kinds of threats, Twitter and other tech companies worked to build closer information-sharing relationships with law enforcement such as the FBI.[12] In the recent reporting known as the Twitter Files, there was an attempt to portray interactions between Twitter and other social media platforms and the FBI as politically driven interference.[13] My experience of these interactions was different. Across the FBI, DHS, and other agencies, the professionals responsible for combating malign foreign interference in elections did so with integrity, and the utmost care and respect for the laws of this country — including the First Amendment.

Which brings us to Hunter Biden's laptop and the *New York Post*. In 2020, the Trust & Safety team noticed activity related to the laptop popping up on Twitter, and that activity, at first glance, bore a lot of similarities to the 2016 Russian hack and leak operation. Twitter had to decide what to do. The only information we had to go on to make this decision was what had been publicly

[8] Ahmed Al-Rawi and Anis Rahman, "Manufacturing rage: The Russian Internet Research Agency's political astroturfing on social media," *First Monday* 25, no. 9, http://dx.doi.org/10.5210/fm.v25i9.10801.
[9] "Moderation Research," Twitter Transparency Center, Twitter, accessed February 4, 2023, https://transparency.twitter.com/en/reports/moderation-research.html.
[10] U.S. Intelligence Community, *Assessing Russian Activities and Intentions in Recent US Elections*, January 6, 2017, https://www.intelligence.senate.gov/sites/default/files/documents/ICA_2017_01.pdf.
[11] U.S. Congress, Senate, Select Committee on Intelligence, *Russian Active Measures and Interference In the 2016 U.S. Election*, 116th Cong., 2d sess., S. Rep. 116-290, https://www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures.
[12] Salvador Rodriguez, "The FBI visits Facebook to talk about 2020 election security, with Google, Microsoft and Twitter joining," *CNBC*, September 4, 2019, https://www.cnbc.com/2019/09/04/facebook-twitter-google-are-meeting-with-us-officials-to-discuss-2020-election-security.html.
[13] Matt Taibbi, "Capsule Summaries of all Twitter Files Threads to Date, With Links and a Glossary," *Racket*, January 4, 2023, https://www.racket.news/p/capsule-summaries-of-all-twitter.

reported. And in that moment, with limited information, Twitter made a mistake: Under the Distribution of Hacked Material Policy[14], the company decided to require the *New York Post* to delete several Tweets linking to stories about the laptop, and prevent links to those stories from being shared across the service. This policy was not meant to be a tool to censor news: It was written to prohibit hacking groups from using Twitter to launder stolen documents — the same activity the Russian government had engaged in in 2016. And in this instance, the company's initial assessment was that the activity bore enough similarities to the 2016 hack and leak that it warranted enforcement.

I've been clear that, in my judgment at the time, Twitter should not have taken action to block the *New York Post*'s reporting. I recommended to Twitter leadership that we take a milder step while we tried to learn more: My recommendation was that we prevent the articles from being actively recommended or amplified by Twitter's algorithms, rather than blocking them altogether. However, in an effort to be consistent with the specifics of the Hacked Materials Policy, which didn't provide for the milder step I recommended, Twitter decided to follow a strict interpretation of the policy, and removed the *New York Post*'s Tweets. Just 24 hours after doing so, the company acknowledged its error[15], and ultimately changed its policies as a result.[16] But this isn't a case where I was right, and others were wrong: The decisions here aren't straightforward, and hindsight is 20/20. It isn't obvious what the right response is to a suspected, but not confirmed, cyberattack by another government on a presidential election. Twitter erred in this case because we wanted to avoid repeating the mistakes of 2016.

And so the basic job of trust and safety remains to try to strike this balance: Between the harms of restricting too much speech, and the dangers of doing too little. Some of the decisions we had to make, like taking down images of animal abuse, were obvious; others, like how to address various forms of misinformation about COVID, are less clear. But someone has to make a call. Companies like Twitter — teams like mine — have to exercise judgment about where to draw the line on this content and implement that judgment consistently at the scale of hundreds of millions of unique posts per day.

I'll be the first to admit that we didn't always get it right. Individual content moderation decisions will always be contentious, and reasonable minds can differ about whether a specific choice was right or wrong. I'm sure we'll talk about some of those choices here today. But what

[14] "Distribution of Hacked Material Policy," Help Center, Twitter, archived version last modified March 2019, https://web.archive.org/web/20190717143909/https://help.twitter.com/en/rules-and-policies/hacked-materials
[15] Kate Conger and Mike Isaac, "In Reversal, Twitter Is No Longer Blocking New York Post Article," *New York Times*, October 16, 2020, https://www.nytimes.com/2020/10/16/technology/twitter-new-york-post.html.
[16] "Distribution of Hacked Materials Policy," Help Center, Twitter, last modified October 2020, https://help.twitter.com/en/rules-and-policies/hacked-materials.

we tried to do at Twitter — across every decision — was to create a rules-based system of governance that would make clear what's allowed, or not, on Twitter, and why.

Transparency is at the heart of this work, and it's where I think Twitter — and all of social media — can and must do better.

Trust is built on understanding — and right now, the vast majority of people don't understand how or why content moderation decisions are made. Much of the knowledge about how platforms like Twitter make decisions is known only to the tiny number of people working at the companies themselves. This is particularly problematic for key decisions, like those impacting elections, where a company's actions are of immense public concern to millions of voters.

During my tenure at Twitter, we started down a path of increased transparency by beginning to pull back the curtains on these decisions. In 2018, we took the unprecedented step of publishing comprehensive archives of Russian election interference during and after the 2016 elections.[17] We released similar data about dozens of other campaigns, spanning hundreds of millions of Tweets and terabytes of media, unearthing government-backed troll farms around the world.[18] Through newer programs like the Twitter Moderation Research Consortium, we aimed to expand this even further, sharing data about key policy decisions in areas like misinformation with hundreds of researchers.[19] I'm concerned by recent reports that suggest this program has been canceled, with no staff left at Twitter to oversee it.[20]

Twitter's relationship with government employees would benefit from similar levels of transparency. While the Twitter Files show a lot of discussions between Twitter employees and political staff on both sides of the aisle, some key context is missing: We were careful to keep the teams involved in those interactions cordoned off from the implementation of our rules. Twitter's government relations staff — the people you see in the Twitter Files answering emails from campaign staff and members of Congress — did not have any kind of decision-making authority over policy enforcement. But how would anyone know that, especially when other big

[17] Vijaya Gadde and Yoel Roth, "Enabling further research of information operations on Twitter," *Twitter Blog*, October 17, 2018, https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.
[18] Twitter, "Disclosing state-linked information operations we've removed," *Twitter Blog*, December 2, 2021, https://blog.twitter.com/en_us/topics/company/2021/disclosing-state-linked-information-operations-we-ve-removed.
[19] Yoel Roth, "The Twitter Moderation Research Consortium is now open to researchers," *Twitter Blog*, September 22, 2022, https://blog.twitter.com/en_us/topics/company/2022/twitter-moderation-research-consortium-open-researchers.
[20] Sheila Dang, "Twitter research group stall complicates compliance with new EU law," *Reuters*, January 28, 2023, https://www.reuters.com/technology/twitter-research-group-stall-complicates-compliance-with-new-eu-law-2023-01-27/.

companies blur the lines between these functions?[21] Twitter set up its teams to promote impartiality; but in the absence of transparency, people reasonably assume that there's opportunity for abuse.

Legislation and regulation can help here. The bipartisan Platform Accountability and Transparency[22] and Digital Services Oversight and Safety[23] Acts, for example, would require platforms to provide data to independent researchers, and empower the FTC to compel them to do so. And, Chairman Comer, your proposal in the Protecting Speech from Government Interference Act to restrict how government employees may pressure social media companies to moderate content would be an important step forward in establishing clear boundaries between government and the private sector.[24] Understanding what platforms are doing and why — and what the influences on them are — is the cornerstone of reestablishing public trust in social media.

It is often said that humor can be an effective antidote to stress. On some of the more challenging days, I've joked that, in trust and safety, there are no good options; just a bunch of bad ones, and your job is to try to pick what the least bad one is. While I was Head of Trust & Safety at Twitter, I strove to do this work with impartiality and a commitment to the fair enforcement of Twitter's written rules. Each day at Twitter, my team and I worked to build trust with the platform's millions of users around the world: By proactively addressing the harms that can come from social media; and by doing that work in a principled, consistent, and transparent way. But whether it's me, or Elon Musk, or another future policymaker, *someone* will have to make choices about the governance of online spaces. Those decisions shouldn't be made behind closed doors, or based on personal whims. I hope that we can work together to find ways to bring greater trust and transparency to social media, and I look forward to answering the Committee's questions about any of these topics to the best of my ability.

---

[21] Emily Birnbaum, "Facebook staff complained for years about their lobbyists' power," *Politico*, October 25, 2021, https://www.politico.com/news/2021/10/25/facebook-fatal-flaw-technologists-lobbyists-516927.
[22] Editorial Board, "A small step toward solving our social media woes," *The Washington Post*, January 17, 2022, https://www.washingtonpost.com/opinions/2022/01/17/legislative-step-toward-solving-our-social-media-woes/.
[23] Justin Hendrix, "Reps. Trahan, Schiff, & Casten Introduce Digital Services Oversight and Safety Act," *Tech Policy Press*, February 23, 2022, https://techpolicy.press/reps-trahan-schiff-casten-introduce-digital-services-oversight-and-safety-act/.
[24] Committee on Oversight and Accountability, "Comer, McMorris Rodgers, Jordan Introduce Bill to Stop Biden Administration from Pressuring Social Media Companies to Censor Americans," Press Release, January 12, 2023, https://oversight.house.gov/release/comer-mcmorris-rodgers-jordan-introduce-bill-to-stop-biden-administration-from-pressuring-social-media-companies-to-censor-americans-2/.