September 24, 2020

The Honorable Jan Schakowsky
Chairwoman
Subcommittee on Consumer
Protection & Commerce
Committee on Energy & Commerce
U.S. House of Representatives
Washington, DC 20515

The Honorable Cathy McMorris Rodgers
Ranking Member
Subcommittee on Consumer
Protection & Commerce
Committee on Energy & Commerce
U.S. House of Representatives
Washington, DC 20515

Dear Chairwoman Schakowsky, Ranking Member McMorris Rodgers, and Members of the Subcommittee:

My name is Julian Sanchez, and I'm a senior fellow at the Cato Institute who focuses on issues at the intersection of technology and civil liberties—above all, privacy and freedom of expression. I'm grateful to the committee for the opportunity to share my views on this important topic.

New communications technologies—especially when they enable horizontal connections between individuals—are inherently disruptive. In 16th century Europe, the advent of movable type printing fragmented a once-unified Christendom into a dizzying array of varied—and often violently opposed—sects. In the 1980s, one popular revolution against authoritarian rule in the Philippines was spurred on by broadcast radio—and decades later, another was enabled by mobile phones and text messaging. Whenever a technology reduces the friction of transmitting ideas or connecting people to each other, the predictable result is that some previously marginal ideas, identities, and groups will be empowered. While this is typically a good thing on net, the effect works just as well for ideas and groups that had previously been marginal for excellent reasons.

Periods of transition from lower to higher connectivity are particularly fraught. Urbanization and trade in Europe's early modern period brought with them, among their myriad benefits, cyclical outbreaks of plague, as pathogens that might once have burned out harmlessly found conditions amenable to rapid spread and mutation. Eventually, of course, populations adapt. Individuals—those who survive—adapt by developing immunities through exposure, which once a critical mass is reached yield herd immunity, leaving pathogens with too few susceptible hosts to spread effectively. Communities adapt by developing hygienic practices and urban architectures designed to divorce the benefits of human connectivity from the pathogens hoping to come along for the ride.

Our own transitional era is host to no shortage of ideological pathogens, from violent and fanatical religious movements to bizarre conspiracy theories such as QAnon. And the social media platforms on which these pathogens spread find themselves in the unenviable position of attempting, by trial and error, to discover how one builds a functional sewer system.

The major social media platforms all engage, to varying degrees, in extensive monitoring of user-posted content via, a combination of human and automated review, with the aim of restricting a wide array of speech those platforms deem objectionable. This typically includes such categories as pornographic content, individual harassment, and—more germane to this hearing's subject—the promotion of extremist violence, specific classes of misinformation or disinformation, and hateful speech directed at specific groups on the basis of race, gender, religion, or sexuality. In response to public criticism, these platforms have in recent years taken steps to crack down more aggressively on hateful and extremist speech, investing in larger teams of human moderators and more sophisticated algorithmic tools designed to automatically flag such content.[1] More recently—and more contentiously—those efforts have expanded to encompass various forms of disinformation, though more often with an aim of flagging and fact-checking the content and removing it.

Here it's necessary to make an obvious but important point: *All* the major platforms' policies go far further in restricting hateful or violent speech than would be permissible under our Constitution via state action, and involve more proactive "correction" or flagging of false speech than it would be desirable for a liberal democratic state to engage in even where legally permissible.

The First Amendment protects hate speech. The Supreme Court has upheld the constitutional right of American neo-Nazis to march in public brandishing swastikas[2], and of a hate group to picket outside the funerals of veterans displaying incredibly vile homophobic and anti-military slogans.[3]

While direct threats and speech that is both intended and likely to incite "imminent" violence fall outside the ambit of the First Amendment, Supreme Court precedent distinguishes such speech from "the mere abstract teaching ... of the moral propriety or even moral necessity for a resort to force and violence,"[4] which remains protected. Unsurprisingly, in light of this case law, a recent Congressional Research Service report found that "laws that criminalize the dissemination of the pure advocacy of terrorism, without more, would likely be deemed unconstitutional."[5]

Happily—at least, as far as most users of social media are concerned—the First Amendment does not bind private firms like YouTube, Twitter, or Facebook, leaving them

---

[1] See, e.g., Kent Walker "Four steps we're taking today to fight terrorism online" Google (June 18, 2017) https://www.blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/ ; Monika Bickert and Brian Fishman "Hard Questions: What Are We Doing to Stay Ahead of Terrorists?" Facebook (November 8, 2018) https://newsroom.fb.com/news/2018/11/staying-ahead-of-terrorists/ ; "Terrorism and violent extremism policy" Twitter (March 2019) https://help.twitter.com/en/rules-and-policies/violent-groups

[2] *National Socialist Party of America v. Village of Skokie*, 432 U.S. 43 (1977)

[3] *Snyder v. Phelps*, 562 U.S. 443 (2011)

[4] *U.S. v. Brandenburg,* 395 U.S. 444 (1969)

[5] Kathleen Anne Ruane, "The Advocacy of Terrorism on the Internet: Freedom of Speech Issues and the Material Support Statutes" Congressional Research Service Report T44646 (September 8, 2016) https://fas.org/sgp/crs/terror/R44626.pdf

with a much freer hand to restrict offensive content that our Constitution forbids the law from reaching. The Supreme Court reaffirmed that principle just last year, in a case involving a public access cable channel in New York. Yet as the Court noted in that decision, this applies only when private determinations to restrict content are truly private. They may be subject to First Amendment challenge if the private entity in question is functioning as a "state actor"—which can occur "when the government compels the private entity to take a particular action" or "when the government acts jointly with the private entity."[6]

Perversely, then any governmental intervention to encourage more aggressive removal of hateful, extremist, or simply false content risks producing the opposite of the intended result. Content moderation decisions that are clearly lawful as an exercise of purely private discretion could be recast as government censorship, opening the door to legal challenge. Should the courts determine that legislative or regulatory mandates had rendered First Amendment standards applicable to online platforms, the ultimate result would inevitably be far *more* hateful, extremist, and uncorrected false speech on those platforms.

Bracketing legal considerations for the moment, it is also important to recognize that the ability of algorithmic tools to accurately identify pathogenic content is not as great as is commonly supposed. I focus on content promoting terrorism or terrorist groups, since it seems like it ought to be the clearest cut both descriptively (knowing whether a post is factually false is typically more demanding and labor intensive than determining whether it praises Al Qaeda) and normatively (almost nobody thinks such content might in some ways be beneficial). Last year, Facebook boasted that its automated filter detected 99.5 percent of the terrorist-related content the company removed before it was posted, with the remainder flagged by users.[7] Many press reports subtly misconstrued this claim. The *New York Times*, for example, wrote that Facebook's "A.I. found 99.5 percent of terrorist content on the site."[8] That, of course, is a very different proposition: Facebook's claim concerned the ratio of content removed after being flagged as terror-related by automated tools versus human reporting, which should be unsurprising given that software can process vast amounts of content far more quickly than human brains. It is *not* the claim that software filters successfully detected 99.5 percent of all terror-related content uploaded to the site—which would be impossible since, by definition, content not detected by either mechanism is omitted from the calculus. Nor does it tell us much about

---

[6] *Manhattan Community Access Corp. v. Halleck*, 17–1702 (2019)

[7] Alex Schultz and Guy Rosen "Understanding the Facebook Community Standards Enforcement Report"
https://fbnewsroomus.files.wordpress.com/2018/05/understanding_the_community_standards_enforcement_report.pdf

[8] Sheera Frenkel, "Facebook Says It Deleted 865 Million Posts, Mostly Spam" New York Times (May 15, 2018). Facebook Says It Deleted 865 Million Posts, Mostly Spam
https://www.nytimes.com/2018/05/15/technology/facebook-removal-posts-fake-accounts.html

the false-positive ratio: How much content was misidentified as terror-related, or how often such content appeared in the context of posts either reporting on or condemning terrorist activities.

There is ample reason to believe that such false positives impose genuine social cost, even in what one would expect to be this easiest type of case. Algorithms may be able to determine that a post contains images of extremist content, but they are far less adept at reading contextual cues to determine whether the purpose of the post is to glorify violence, condemn it, or merely document it—something that may in certain cases even be ambiguous to a human observer. Journalists and human rights activists, for example, have complained that tech company crackdowns on violent extremist videos have inadvertently frustrated efforts to document human rights violations[9], and erased evidence of war crimes in Syria.[10] Last year, a YouTube crackdown on white supremacist content resulted in the removal of a large number of historical videos posted by educational institutions, and by anti-racist activist groups dedicated to documenting and condemning hate speech.[11]

Of course, such errors are often reversed by human reviewers—at least when the groups affected have enough know-how and public prestige to compel a reconsideration. Government mandates, however, alter the calculus. As three United Nations special rapporteurs wrote, objecting to a proposal in the European Union to require automated filtering, the threat of legal penalties were "likely to incentivize platforms to err on the side of caution and remove content that is legitimate or lawful."[12] If the failure to filter to the government's satisfaction risks stiff fines, any cost-benefit analysis for platforms will favor significant overfiltering: Better to pull down ten benign posts than risk leaving up one that might expose them to penalties. For precisely this reason, the EU proposal has been roundly condemned by human rights activists[13] and fiercely opposed by a wide array of

---

[9] Dia Kayyali and Raja Althaibani, "Vital Human Rights Evidence in Syria is Disappearing from YouTube" https://blog.witness.org/2017/08/vital-human-rights-evidence-syria-disappearing-youtube/

[10] Bernhard Warner, "Tech Companies Are Deleting Evidence of War Crimes" The Atlantic (May 8, 2019). https://www.theatlantic.com/ideas/archive/2019/05/facebook-algorithms-are-making-it-harder/588931/

[11] Elizabeth Dwoskin, "How YouTube erased history in its battle against white supremacy" Washington Post (June 13, 2019). https://www.washingtonpost.com/technology/2019/06/13/how-youtube-erased-history-its-battle-against-white-supremacy/?utm_term=.e5391be45aa2

[12] David Kaye, Joseph Cannataci, and Fionnuala Ní Aoláin "Mandates of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; the Special Rapporteur on the right to privacy and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism" https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24234

[13] Faiza Patel, "EU 'Terrorist Content' Proposal Sets Dire Example for Free Speech Online" (Just Security) https://www.justsecurity.org/62857/eu-terrorist-content-proposal-sets-dire-free-speech-online/

civil society groups.[14]

A recent high-profile case illustrates the challenges platforms face: The efforts by platforms to restrict circulation of video depicting the brutal mass shooting of worshippers at a mosque in Christchurch, New Zealand. Legal scholar Kate Klonick documented the efforts of Facebook's content moderation team for *The New Yorker*[15], while reporters Elizabeth Dwoskin and Craig Timberg wrote about the parallel struggles of YouTube's team for *The Washington Post*[16]—both accounts are illuminating and well worth reading.

Though both companies were subject to vigorous condemnation by elected officials for failing to limit the video quickly or comprehensively enough, the published accounts make clear this was not for want of trying. Teams of engineers and moderators at both platforms worked around the clock to stop the spread of the video, by increasingly aggressive means. Automated detection tools, however, were often frustrated by countermeasures employed by uploaders, who continuously modified the video until it could pass through the filters. This serves as a reminder that even if automated detection proves relatively effective at any given time, they are in a perennial arms race with determined humans probing for algorithmic blind spots.[17] There was also the problem of users who had—perhaps misguidedly—uploaded parts of the video in order to condemn the savagery of the attack and evoke sympathy for the victims. Here, the platforms made a difficult real-time value judgment that, in this case, the balance of equities favored an aggressive posture: Categorical prohibition of the content regardless of context or intent, coupled with tight restrictions on searching and sharing of recently uploaded video.

Both the decisions the firms made and the speed and adequacy with which they implemented them in a difficult circumstance will be—and should be—subject to debate and criticism. But it would be a grave error to imagine that political intervention is likely to produce better results than such context-sensitive judgments, or that smart software will somehow obviate the need for a difficult and delicate balancing of competing values.

---

[14] "Letter to Ministers of Justice and Home Affairs on the Proposed Regulation on Terrorist Content Online" https://cdt.org/files/2018/12/4-Dec-2018-CDT-Joint-Letter-Terrorist-Content-Regulation.pdf

[15] Kate Klonick, "Inside the Team at Facebook That Dealt With the Christchurch Shooting" *The New Yorker* (April 25, 2019) https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting

[16] Elizabeth Dwoskin and Craig Timberg "Inside YouTube's struggles to shut down video of the New Zealand shooting — and the humans who outsmarted its systems" Washington Post (March 18, 2019) https://www.washingtonpost.com/technology/2019/03/18/inside-youtubes-struggles-shut-down-video-new-zealand-shooting-humans-who-outsmarted-its-systems/?utm_term=.6a5916ba26c1

[17] See, e.g., Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran "Deceiving Google's Perspective API Built for Detecting Toxic Comments" Arvix (February 2017) https://arxiv.org/abs/1702.08138

Finally, even restricting our consideration to the uncoerced private efforts of social media platforms, there are a number of strong reasons to counsel restraint—especially when dealing not with overt endorsement of violence, but the ever more varied strains of disinformation and conspiracy theorizing spreading online.

One potential pitfall is the so-called "Streisand effect," named for the ill-fated effort of singer and actress Barbra Streisand to suppress distribution of an aerial photograph of her home via litigation.[18] To Streisand's surprise in 2003, but to the surprise of nobody in 2020, the litigation itself drew vastly more attention—and resulted in far wider dissemination of the photograph—than the initial publication. This risk may be compounded in the case of misinformation linked to conspiracy theories, which often posit some cabal of elites determined to conceal the truth from the general population. To those inclined to such thinking, corporate attempts to suppress false content may perversely be seen as a kind of validation. For such users, mockery and debunking by trusted peers may prove a more effective antidote than a fact-check from a formal journalistic entity.

There is also the risk of driving false and extreme content into the ideological equivalent of the black market. The popular Web forum Reddit recently banned some 7,000 groups or "subreddits" that trafficked in hateful content, including many associated with the QAnon movement.[19] Many of these, unsurprisingly, migrated to completely unrestricted online fora, where content far more extreme and violet than Reddit's moderators would have allowed is welcome. This is not to say Reddit acted in error—not all users will migrate to fringe fora, and Reddit's vastly larger user base is at less risk of chancing upon such content and slipping down the rabbit hole. But it is a tradeoff that must be borne in mind.

Until the general population develops stronger antibodies against ideological pathogens, the best course may be to focus on light-touch user interface interventions that hamper the spread of disinformation without removing the content entirely. Fact checks are one strategy already implemented by many platforms. Another is the disabling of single-click sharing of content flagged as false or misleading, and the adjustment of content-recommending algorithms to ensure that such content is not foisted upon users who do not willingly seek it out. The memetic equivalent of a face mask, this does not prevent transmission by a determined user, but it does at least reduce the rate of casual or unthinking transmission.

Another possibility is to visually distinguish content articles from reputable news outlets. In the pre-Internet era, the difference between a *New York Times* or *Wall Street Journal* report and a mimeographed screed could be seen at a glance, independently of the differences in style and content. On a Facebook feed, everything looks more or less

---

[18] "Words We're Watching: 'Streisand Effect'" https://www.merriam-webster.com/words-at-play/words-were-watching-streisand-effect-barbra

[19] Ray, Siladitya, "Reddit Says It Has Removed 7,000 'Hateful' Subreddits Since Changing Hate Speech Policy In June" *Forbes* (Aug 20, 2020) https://www.forbes.com/sites/siladityaray/2020/08/20/reddit-says-it-has-removed-7000-hateful-subreddits-since-changing-hate-speech-policy-in-june/#dac6459233ee

identical. This strategy is only viable, of course, to the extent platforms can resist both political and user pressure to give their imprimatur to unreliable information sources with large constituencies.

This brings us, finally, to the primary obstacle to such efforts: What one person dubs memetic hygiene, another is apt to see as ideological bias and discrimination. As noted earlier, moderation compelled by state action is apt to invite legal challenges that would perversely yield far *less* moderation. But conversely, platforms need assurances that their attempts to restrict the rapid spread of disinformation and extremist rhetoric will not provoke a political reaction, such as the revocation of legal protections against liability for user-posted content. Any efforts platforms make in this domain are bound to involve a substantial element of subjective judgment and be prone to error. For the conspiracy theorist, anyone who calls it a "conspiracy theory" is biased. If we expect inherently risk-averse businesses to be proactive about curating content to stanch the spread of extremist rhetoric and disinformation, they must be confident they are free to muddle through the process of developing adaptive responses—and, inevitably, to make many mistakes along the way—without incurring ruinous legal consequences as a result.

Sincerely,

/s/

Julian Sanchez
Senior Fellow
Cato Institute