Testimony of Julian Sanchez
Senior Fellow
Cato Institute

Before the United State House of Representatives Committee on Homeland Security, Subcommittee on Intelligence and Counterterrorism

Hearing on
Artificial Intelligence and Counterterrorism: Possibilities and Limitations

Cannon House Office Building Room 310 Washington, D.C.
June 25, 2019

My thanks to the chair, ranking member, and all members of this subcommittee for the opportunity to speak to you today.

As a firm believer in the principle of comparative advantage, I don't intend to delve too deeply into the technical details of automated content filtering, which my co-panelists are far better suited than I to address. Instead I want to focus on legal and policy considerations, and above all to urge Congress to resist the temptation to intervene in the highly complex—and admittedly highly imperfect—processes by which private online platforms seek to moderate both content related to terrorism and "hateful" or otherwise objectionable speech more broadly. (My colleague at the Cato Institute, John Samples, recently published a policy paper dealing still more broadly with issues surrounding regulation of content moderation policies, which I can enthusiastically recommend to the committee's attention.[1])

The major social media platforms all engage, to varying degrees, in extensive

[1] John Samples, "Why the Government Should Not Regulate Content Moderation of Social Media" (Cato Institute) https://www.cato.org/publications/policy-analysis/why-government-should-not-regulate-content-moderation-social-media#full

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, typically including nudity, individual harassment, and—more germane to our subject today—the promotion of extremist violence and, more broadly, 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. [2]

Elected officials and users of these platforms are often dissatisfied with these efforts—both with the speed and efficacy of content removal and the scope of individual platforms' policies. Yet it is clear that *all* the major platforms' policies go far further in restricting speech than would be permissible under our Constitution via state action.

The First Amendment protects hate speech. The Supreme Court has ruled in favor of the constitutional right of American neo-Nazis to march in public brandishing swastikas[3], and of a hate group to picket outside the funerals of veterans displaying incredibly vile homophobic and anti-military slogans.[4]

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,"[5] 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."[6]

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 with a much freer hand to restrict offensive content that our Constitution forbids the law from reaching. The Supreme Court reaffirmed that principle just this month, in a

---

[2] 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

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

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

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

[6] 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

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."[7]

Perversely, then, legislative efforts to compel more aggressive removal of hateful or extremist content risk 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 mandates had rendered First Amendment standards applicable to online platforms, the ultimate result would almost certainly be *more* hateful and extremist 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 hateful or extremist content is not as great as is commonly supposed. 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.[8] 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."[9] 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 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. Algorithms may be able to determine that a post contains images of extremist

---

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

[8] 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

[9] 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

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[10], and erased evidence of war crimes in Syria.[11] Just this month, 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.[12]

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."[13] 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[14] and fiercely opposed by a wide array of civil society groups.[15]

A recent high-profile case illustrates the challenges platforms face: The efforts by

---

[10] 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/

[11] 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/

[12] 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

[13] 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

[14] 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/

[15] "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

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*[16], while reporters Elizabeth Dwoskin and Craig Timberg wrote about the parallel struggles of YouTube's team for *The Washington Post*[17]—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.[18] 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 broad legislative mandates are 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.

I thank the committee again for the opportunity to testify, and look forward to your questions.

---

[16] 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

[17] 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

[18] 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