United States House of Representatives Committee on Homeland Security Hearing Titled "Examining Social Media Companies' Efforts to Counter Online Terror Content and Misinformation"

June 26, 2019

Written Testimony of Nick Pickles
Senior Public Policy Strategist
Twitter, Inc.

Chairman Thompson, Ranking Member Rogers, and Members of the Committee:

Twitter's purpose is to serve the public conversation. Twitter is a place where people from around the world come together in an open and free exchange of ideas. My statement today will provide information and deeper context on (I) Twitter's work to protect the health of the public conversation, including combating terrorism, violent extremist groups, hateful conduct, and platform manipulation, and (II) our partnerships and societal engagement.

I. TWITTER'S WORK TO PROTECT THE HEALTH OF THE PUBLIC CONVERSATION

All individuals accessing or using Twitter's services must adhere to the policies set forth in the Twitter Rules. Accounts under investigation or which have been detected as sharing content in violation with the Twitter Rules may be required to remove content, or in serious cases, will see their account permanently suspended. Our policies and enforcement options evolve continuously to address emerging behaviors online.

A. *Policy on Terrorism*

Individuals are prohibited from making specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people. This includes, but is not limited to, threatening or promoting terrorism.

We have now suspended more than 1.5 million accounts for violations related to the promotion of terrorism between August 1, 2015, and December 31, 2018. In 2018, a total of 371,669 accounts were suspended for violations related to promotion of terrorism. We continue to see more than 90 percent of these accounts suspended through proactive measures.

The trend we are observing year-over-year is a steady decrease in terrorist organizations attempting to use our service. This is due to zero-tolerance policy enforcement that has allowed us to take swift action on ban evaders and other identified forms of behavior used by terrorist entities and their affiliates. In the majority of cases, we take action at the account creation stage — before the account even Tweets.

Government reports constituted less than 0.1 percent of all suspensions in the last reporting period. Continuing the trend we have seen for some time, the number of reports we received from governments of terrorist content from the second half of last year decreased by 77 percent compared to the previous reporting period (January-June 2018).

We are reassured by the progress we have made, including recognition by independent experts. For example, Dublin City University Professor Maura Conway found in a detailed study that "ISIS's previously strong and vibrant Twitter community is now...virtually non-existent."

In tandem with removing content, our wider efforts on countering violent extremism going back to 2015 have focused on bolstering the voices of non-governmental organizations and credible outside groups to use our uniquely open service to spread positive and affirmative campaigns that seek to offer an alternative to narratives of hate.

We have partnered with organizations delivering counter and alternative narrative initiatives across the globe and we encourage the Committee to consider the role of government in supporting the work of credible messengers in this space at home and abroad.

B. *Policy on Violent Extremist Groups*

In December 2017, we broadened our rules to encompass accounts affiliated with violent extremist groups. Our prohibition on the use of Twitter's services by violent extremist groups — i.e., identified groups subscribing to the use of violence as a means to advance their cause — applies irrespective of the cause of the group.

Our policy states:

> *Violent extremist groups are those that meet all of the below criteria:*
>
> - *identify through their stated purpose, publications, or actions as an extremist group;*
> - *have engaged in, or currently engage in, violence and/or the promotion of violence as a means to further their cause; and*
> - *target civilians in their acts and/or promotion of violence.*

An individual on Twitter may not affiliate with such an organization — whether by their own statements or activity both on and off the service — and we will permanently suspend those who do so.

We know that the challenges we face are not static, nor are bad actors homogenous from one country to the next in how they behave. Our approach combines flexibility with a clear, consistent policy philosophy, enabling us to move quickly while establishing clear norms of unacceptable behavior.

Since the introduction of our policy on violent extremist groups, we have taken action on 184 groups under this policy and permanently suspended 2,182 unique accounts. Ninety-three of these groups advocate violence against civilians alongside some form of extremist white supremacist ideology.

C. *Policy on Hateful Conduct*

People on Twitter are not permitted to promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm toward others on the basis of these categories.

We do not allow individuals to use hateful images or symbols in their profile image or profile header. Individuals on the platform are not allowed to use the username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate toward a person, group, or protected category.

Under this policy, we take action against behavior that targets individuals or an entire protected category with hateful conduct. Targeting can happen in a number of ways, for example, mentions, including a photo of an individual, or referring to someone by their full name.

When determining the penalty for violating this policy, we consider a number of factors including, but not limited to the severity of the violation and an individual's previous record of rule violations. For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent account suspension. If an account is engaging primarily in abusive behavior, or is deemed to have shared a violent threat, we will permanently suspend the account upon initial review.

D. *Manipulation of the Public Conversation*

Our policies regarding terrorism, violent extremist groups, and hateful conduct are strictly enforced, as are all our policies. We take additional steps to safeguard the public conversation from manipulation.

As a uniquely open, public service, the clarification of falsehoods could happen in seconds on Twitter. We proactively enforce policies and use technology to halt the spread of content propagated through manipulative tactics, such as automation or attempting to deliberately game trending topics.

Our Site Integrity team is dedicated to identifying and investigating suspected platform manipulation on Twitter, including activity associated with coordinated malicious activity that we are able to reliably associate with state-affiliated actors. In partnership with teams across the company, we employ a range of open-source and proprietary signals and tools to identify when attempted coordinated manipulation may be taking place, as well as the actors responsible for it. We also partner closely with governments, law enforcement, academics, researchers, and our peer companies to improve our understanding of the actors involved in information operations and develop a holistic strategy for addressing them.

For example, we typically challenge 8 to 10 million accounts per week for these behaviors, requesting additional details, like email addresses and phone numbers in order to authenticate the account. We also recently acquired a new business to augment our efforts in this regard. This strategic investment will be a key driver as we work to protect the public conversation and help all individuals on our service see relevant information.

Attempts to execute misinformation campaigns rely on tactics like coordinated account manipulation or malicious automation — all of which are against Twitter's Rules. We are continuing to explore ways at how we may take action — through both policy and product — on these types of issues in the future. We continue to critically examine additional safeguards we can implement to protect the conversation occurring on Twitter.

In October 2018, we published the first comprehensive archive of Tweets and media associated with known state-backed information operations on Twitter and since then we have provided two further updates covering a range of actors. Thousands of researchers from across the globe have now made use of these datasets, which contain more than 30 million Tweets and more than one terabyte of media, using our archive to conduct their own investigations and to share their insights and independent analysis with the world.

By making this data open and accessible, we seek to empower researchers, journalists, governments, and members of the public to deepen their understanding of critical issues impacting the integrity of public conversation online, particularly around elections. This transparency is core to our mission.

E. *Investing in Tech: Behavior vs. Content*

Twitter's philosophy is to take a behavior-led approach, utilizing a combination of machine learning and human review to prioritize reports and improve the health of the public conversation. That is to say, we increasingly look at how accounts behave before we look at the content they are posting. This is how we seek to scale our efforts globally and leverage technology even where the language used is highly context specific. Twitter employs extensive content detection technology to identify potentially abusive content on the service, along with allowing users to report content to us either as an individual or a bystander.

We have made the health of Twitter our top priority, and our efforts will be measured by how we help encourage more healthy debate, conversations, and critical thinking on the platform. Conversely, abuse, automation, hateful conduct, terrorism, and manipulation will detract from the health of our platform.

For abuse, this two-pronged strategy has allowed us to take three times the amount of enforcement of action on abuse within 24 hours than this time last year. We now proactively surface nearly 40 percent of abusive content we remove compared to 20 percent a year ago to reduce the burden on the individual. Since we started using machine learning three years ago to reduce the visibility on abusive content:

- 80 percent of all replies that are removed were already less visible;
- Abuse reports have been reduced by 7.6 percent;
- The most visible replies receive 45 percent less abuse reports;
- 100,000 accounts were suspended for creating new accounts after a suspension during January through March 2019 — a 45 percent increase from the same time last year;
- 60 percent faster response to appeals requests with our new in-app appeal process;
- 3 times more abusive accounts suspended within 24 hours after a report compared to the same time last year; and
- 2.5 times more private information removed with a new, easier reporting process.

II. PARTNERSHIPS AND SOCIETAL ENGAGEMENT

We work closely with the Federal Bureau of Investigation, along with law enforcement and numerous public safety around the world. As our partnerships deepen, we are able to better respond to the changing threats we all face, sharing valuable information and promptly responding to valid legal requests for information.

A. *Industry Collaboration*

Collaboration with our industry peers and civil society is also critically important to addressing common threats from terrorism globally. In June 2017, we launched the Global Internet Forum to Counter Terrorism (the “GIFCT”), a partnership among Twitter, YouTube, Facebook, and Microsoft.

The GIFCT facilitates, among other things, information sharing; technical cooperation; and research collaboration, including with academic institutions. In September 2017, the members of the GIFCT announced a significant financial commitment to support research on terrorist abuse of the Internet and how governments, tech companies, and civil society can respond effectively. Our goal is to establish a network of experts that can develop platform-agnostic research questions and analysis that consider a range of geopolitical contexts.

Technological collaboration is a key part of GIFCT’s work. In the first two years of GIFCT, two projects have provided technical resources to support the work of members and smaller companies to remove terrorist content.

First, the shared industry database of "hashes" — unique digital "fingerprints" — for violent terrorist propaganda now spans more than 100,000 hashes. The database allows a company that discovers terrorist content on one of its sites to create a digital fingerprint and share it with the other companies in the forum, who can then use those hashes to identify such content on their services or platforms, review against their respective policies and individual rules, and remove matching content as appropriate or block extremist content before it is posted.

Second, a year ago, Twitter began working with a small group of companies to test a new collaborative system. Because Twitter does not allow files other than photos or short videos to be uploaded, one of the behaviors we saw from those seeking to promote terrorism was to post links to other services where people could access files, longer videos, PDFs, and other materials. Our pilot system allows us to alert other companies when we removed an account or Tweet that linked to material that promoted terrorism hosted on their service. This information sharing ensures the hosting companies can monitor and track similar behavior, taking enforcement action pursuant with their individual policies. This is not a high-tech approach, but it is simple and effective, recognizing the resource constraints of smaller companies.

Based on positive feedback, the partnership has now expanded to 12 companies and we have shared more than 14,000 unique URLs with these services. Every time a piece of content is removed at source, it means any link to that source — wherever it is posted — will no longer be operational.

We are eager to partner with additional companies to expand this project, and we look forward to building on our existing partnerships in the future.

B. *The Christchurch Attack*

The Christchurch attack was unprecedented both in the way it exploited the online environment but also the disparate range of online communities that were involved in sharing the Christchurch video and the hateful manifesto of the attacker.

We saw a wide range of individuals on the service continue to upload excerpts of the attacker's video even after it had been removed. This included those who sought to condemn the attack, including those who combined video of their condemnation and prayers with video of the attack, and others who saw baseless conspiracies and wanted to provide evidence to refute such claims. There were those who believed to remove the content was censorship and those who wanted to amplify the hatred the video embodied. Our analysis found 70 percent of the views of footage of the attack in Christchurch on Twitter were from content posted by verified accounts, including media outlets and those seeking to condemn the violence. In all of these circumstances we removed the relevant content.

As a uniquely open service, we see regular examples around the world of our users, communities, and groups challenging hate and division, particularly following violent acts. As the world began to comprehend the horror of what took place in Christchurch, some may have sought to promote hate, but there was another conversation taking place, one that reached many more people. The hashtag #HelloBrother saw people around the world recognizing the brave act of one victim and rejecting the terrorist's narrative, while hundreds of thousands of Tweets expressed similar sentiments in their own way. This is the potential of open public conversation and what it can empower — a global platform for the best of society to challenge violence and hatred.

C. *The Christchurch Call to Action*

In the months since the attack, New Zealand Prime Minister Jacinda Ardern has led the international policy debate, and that work has culminated in the Christchurch Call. Twitter's Chief Executive Officer Jack Dorsey attended the launch of the Christchurch Call in Paris, meeting with the Prime Minister to express our support and partnership with the New Zealand Government.

Because terrorism cannot be solved by the tech industry alone, the Christchurch Call is a landmark moment and an opportunity to convene governments, industry, and civil society to unite behind our mutual commitment to a safe, secure open, global Internet. It is also a moment to recognize that however or wherever evil manifests itself, it affects us all.

In fulfilling our commitments in the Call, we will take a wide range of actions. We continue to invest in technology to prioritize signals, including user reports, to ensure we can respond as quickly as possible to a potential incident, building on the work we have done to harness proprietary technology to detect and disrupt bad actors proactively.

As part of our commitment to educate users about our rules and to further prohibit the promotion of terrorism or violent extremist groups, we have updated our rules and associated materials to be clearer on where these policies apply. This is accompanied by further data being provided in our transparency report, allowing public consideration of the actions we are taking under our rules, as well as how much content is detected by our proactive efforts.

Twitter will take concrete steps to reduce the risk of livestreaming being abused by terrorists, while recognizing that during a crisis these tools are also used by news organizations, citizens and governments. We are investing in technology and tools to ensure we can act even faster to remove video content and stop it spreading.

Finally we are committed to continuing our partnership with industry peers, expanding on our URL sharing efforts along with wider mentoring efforts, strengthening our new crisis protocol arrangements, and supporting the expansion of GIFCT membership.

D. *A Whole of Society Response*

The challenges we face as a society are complex, varied, and constantly evolving. These challenges are reflected and often magnified by technology. The push and pull factors influencing individuals vary widely and there is no one solution to prevent an individual turning to violence. This is a long-term problem requiring a long-term response, not just the removal of content.

While we strictly enforce our policies, removing all discussion of particular viewpoints, no matter how uncomfortable our customers may find them, does not eliminate the ideology underpinning them. Quite often, it moves these views into darker corners of the Internet where they cannot be challenged and held to account. As our peer companies improve in their efforts, this content continues to migrate to less-governed platforms and services. We are committed to learning and improving, but every part of the online ecosystem has a part to play.

We have a critical role. Tech companies and content removal online cannot alone, however, solve these issues. They are systemic and societal and so they require an whole-of-society approach. We welcome the opportunity to continue to work with our industry peers, government, academics, and civil society to find the right solutions.

* * *

Our goal is to protect the health of the public conversation and to take immediate action on those who seek to spread messages of terror and violent extremism. However, no solution is perfect, and no technology is capable of detecting every potential threat.

Twitter's efforts around the globe to support civil society voices and promote positive messages have seen Twitter employees train groups on five continents and we have provided pro-bono advertising to groups to enable their messages to reach millions of people. When we at Twitter talk about the health of the public conversation, we see the principles of civility, empathy, and mutual respect as foundational to our work. We will not solve problems by removing content alone. We should not underestimate the power of open conversation to change minds, perspectives, and behaviors.

We stand ready to assist the Committee in its important work regarding the issue of the tools that Internet companies can employ to stop the spread of terrorist content and misinformation on our services.