Statement of Nick Pickles
Senior Strategist, Public Policy, Twitter, Inc.

Before the Committee on the Judiciary
United States House of Representatives

July 17, 2018

Chairman Goodlatte, Ranking Member Nadler, and Members of the Committee:

Thank you for the opportunity to appear here today. We appreciate the Committee's inquiries about Twitter's content moderation policies, and we are pleased to be here to share our story.

We are delighted that in the United States, all 100 senators, 50 governors and almost every member of the House of Representatives have official Twitter accounts, which are used to engage in local, national, and global conversations on a wide range of issues of civic importance. We also partner with news organizations on a regular basis to live-stream congressional hearings and political events, giving users a front-row seat to history from their smartphones or computer screens.

Our purpose is to serve the public conversation.

As part of that, there are certain responsibilities we consider core to our company. We must ensure that all voices can be heard. We must make it so that everyone feels safe participating in the conversation − whether they are speaking or simply listening. And we must ensure people can trust in the credibility of the conversation and those taking part.

Our commitment to this work is at the very heart of why people come to Twitter.

We are also committed to improving the health of the public conversation and the health of the conversation on Twitter. Consistent with that effort, we have made more than 30 policy and product changes since the beginning of last year aimed at improving health on our platform. Our data show we are making progress in advancing this goal.

We are also seeking to collaborate with outside experts to better define and measure the health of the public conversation. We have requested proposals and will be announcing the outcome of that process soon. And we are committed to sharing the results of our collaboration so that other organizations can benefit from that work.

Threats of violence, abusive conduct, and harassment can have the effect of bullying voices into silence, thereby robbing other Twitter users of valuable perspectives and threatening the freedom of expression that Twitter seeks to foster. We want to ensure that Twitter continues to be a safe space for our users to share their viewpoints with the broader Twitter community.

To do that, we must keep Twitter safe for all viewpoints and opinions, even those viewpoints and opinions that some of our users may find objectionable or with which they vehemently disagree − so long as the content is not in violation of the Twitter Rules. We do not believe that censorship will solve political or societal challenges or that removing certain content could resolve disagreements or address prejudices. We are committed to protecting speech and promoting the health of the public conversation on our platform.

Accordingly, the Twitter Rules prohibit certain types of behavior on our platform. Because abusive activity − and the challenge of detecting and curtailing it − is not static and has evolved over time, those rules continue to be developed and refined over the years as well.

Our rules are laid out in detailed and plain language in our Help Center, which can be found at help.twitter.com. They are not based on ideology or a particular set of beliefs. Instead, the Twitter Rules are based on behavioral contexts. For example, our rules prohibit making specific threats of violence or wishing for the serious physical harm, death, or disease of an individual or group of people.

Our rules also govern other abusive activity on the platform. For example, we prohibit malicious automation, spam, impersonation accounts, and the use of multiple accounts for overlapping purposes. Again, all of these rules are based on problematic behaviors, not the content of the Tweets or any ideology.

We also have stringent rules and policies that govern advertising on the platform. Advertising on Twitter generally takes the form of promoted Tweets, which advertisers can use to reach new users. Because promoted Tweets are presented to users from accounts they have not yet chosen to follow, Twitter applies a robust set of policies that prohibit, among other things, ads for illegal goods and services, ads making misleading or deceptive claims, ads for drugs or drug paraphernalia, ads containing hate content, sensitive topics, and violence, and ads containing offensive or inflammatory content. These policies are laid out at twitter.com/adspolicy.

We see a range of groups across the political spectrum use our advertising products to promote content about a variety of issues ranging from immigration to tax reform. Like any account that uses our advertising products, those groups are all bound by the same Twitter ads policies and Twitter Rules.

Both organic and promoted content can be reported by our users. We address such reports with a combination of technology and human review approaches. Machine learning improvements are enabling us to be more proactive in finding those who are being disruptive, but user reports are still a highly valuable part of our work. When evaluating these reports, we take into account a variety of factors and context, including whether the behavior is directed at an individual, a group, or a protected category of people. We also take into account whether the user has a history of violating our policies as well as the severity of the violation.

Accounts that violate our policies and the Twitter Rules can be subject to a range of enforcement actions, including temporary and, in some cases, permanent suspension. We recognize that a lack of transparency in enforcement actions can lead to a lack of public understanding about what an individual may have done to warrant action, and we have taken meaningful steps to address this where possible. For example, where appropriate, users are now notified of the specific Tweet that we determined to be in violation of our rules; we also alert those users to the specific rule they violated.

Our Safety Center houses information about our rules, tools, philosophy, and partnerships to further explain our work in this area. We explain our approach to enforcement in greater detail here: help.twitter.com/en/rules-and-policies/enforcement-philosophy.

Because our enforcement process typically relies on both automated and manual human review, we often have to make tough calls, and we do not always get things right − especially given the scope and scale of a platform such as Twitter, where users collectively post hundreds of millions of Tweets each day. When we make a mistake, we acknowledge it and strive to learn from it. We are committed to being direct and engaged with our users and the public − including elected officials − when we get things wrong.

Where we identify suspicious account activity (*e.g.*, exceptionally high-volume Tweeting with the same hashtag, or mentioning the same @handle without a reply from the account being addressed), we automatically send the account owner a test to confirm he or she is still in control of the account. These automated tests vary depending on the type of suspicious activity we detect, and may involve the account owner completing a simple reCAPTCHA challenge or a password reset request. We do not immediately remove content as part of these automated tests, but limit its visibility until the test is passed.

This approach has proven effective in helping us address malicious automation and spam on our platform. In May 2018, for example, our systems identified and challenged more than 9.9 million potentially spammy or automated accounts per week. That is an increase from 6.4 million in December 2017, and 3.2 million in September.

Due to technology and process improvements during the past year, we are now removing 214 percent more accounts for violating our spam policies on a year-on-year basis.

At the same time, the average number of spam reports we received through our reporting flow continued to drop − from an average of approximately 25,000 per day in March, to approximately 17,000 per day in May. We've also seen a 10 percent drop in spam reports from search as a result of our recent changes.

These metrics demonstrate our progress, but our work will never be complete. Bad actors change their behavior and we are constantly evaluating new threats and behavior. Among other things, we rely on our detection tools to identify people who have been suspended from the platform and who have created a new Twitter account or those who use multiple accounts for the same purpose.

We have also taken additional proactive steps recently to make follower counts more meaningful and accurate by removing locked accounts from follower counts globally. This step is a reflection of our ongoing commitment to the health of Twitter and a desire to ensure indicators that users rely on to make judgements about an account are as accurate as possible. Our process applies to all accounts active on the platform, regardless of the content they post.

Another critical part of our commitment to health is changing how we think about the areas on Twitter where our systems curate how information is presented. In places like search and conversations where we try to present content we believe you are most likely to find interesting, we are increasingly relying on behavior to help us make those determinations.

To help us do that, we recently took steps to more effectively address behaviors and activity on the platform that do not necessarily violate our policies, but that distort and detract from the public conversation. Most significantly, this approach enables us to improve the overall health of the conversation without needing to remove content from Twitter. Ultimately, everyone's comments and perspectives are available, but those who are simply looking to disrupt the conversation will not be rewarded by having their Tweets placed at the top of the conversation or search results.

Early results demonstrated that this approach has a positive impact, resulting in a four percent decrease in abuse reports from search and eight percent fewer abuse reports from conversations.

Some critics have described the sum of all of this work as a banning of conservative voices. Let me make clear to the Committee today that these claims are unfounded and false. In fact, we have deliberately taken this approach as a robust defense against bias, as it requires us to define and act upon bad conduct, not a specific type of speech. Our purpose is to serve the conversation, not to make value judgments on personal beliefs.

Our success as a company depends on making Twitter a safe place for free expression and a place that serves healthy public conversation. We know that Twitter plays an increasingly vital role in the world, and we know there is much work for us to do to make it even better. And we are committed to continue to improve transparency and visibility to the people using our service. Thank you again, and I look forward to your questions.