Response to Written Questions Submitted by Hon.
Jerry Moran
Written Questions for the Record to
Carlos Monje

Question 1. How is the battle against terrorist content different from misinformation and abuse?

Response. Twitter is committed to combating terrorist content, abusive activity (including malicious automated activity) and the spread of misinformation on our platform. Because bad actors often rely on the same methods to propagate such content, to some extent, we deploy similar tools to detect and stop all malicious activity on Twitter, including the proliferation of terrorist content. At the same time, we recognize that each of these areas presents unique challenges to deploying our technology at scale. For example, with terrorist content, we can more readily identify the signals of bad actors intending to disseminate terrorist propaganda and can efficiently find and suspend many of these accounts using machine learning. In contrast, for abuse and misinformation, the context of conversations and the content itself are often needed to determine whether something crosses a policy line.

However, Twitter's approach to addressing the spread of malicious automation and inauthentic accounts on our platform is to focus wherever possible on identifying problematic behavior, and not on the content itself. Those who are seeking to influence a wide audience often find ways to try to artificially amplify their messages across Twitter. As with spam, these behaviors frequently provide more precise signals than focusing on content alone.

Accordingly, we monitor various behavioral signals related to the frequency and timing of Tweets, Retweets, likes, and other such activity, as well as to similarity in behavioral patterns across accounts, in order to identify accounts that are likely to be maliciously automated or acting in an automated and coordinated fashion in ways that are unwelcome to our users. We monitor and review unsolicited targeting of accounts, including accounts that mention or follow other accounts with which they have had no prior engagement. For example, if an account follows 1,000 users within the period of one hour, or mentions 1,000 accounts within a short period of time, our systems are capable of detecting that activity as aberrant and as potentially originating from suspicious accounts.

Twitter is continuing its effort to detect and prevent malicious automation by leveraging our technological capabilities and investing in initiatives aimed at understanding and addressing behavioral patterns associated with such accounts. For example, in early 2017, we launched the Information Quality initiative, an effort aimed at enhancing the strategies we use to detect and stop bad automation, improve machine learning to spot spam, and increase the precision of our tools designed to prevent such content from contaminating our platform.

In 2017, we have made significant improvements to reduce external attempts to manipulate content visibility. These improvements were driven by investments in methods to detect malicious automation through abuse of our API, limit the ability of malicious actors to create new accounts in bulk, detect coordinated malicious activity across clusters of accounts, and better enforce policies against abusive third-party applications.

In addition, we have developed new techniques for identifying patterns of activity inconsistent with legitimate use of our platform (such as near-instantaneous replies to Tweets, nonrandom Tweet timing, and coordinated engagement), and we are currently implementing these detections across our platform. We have improved our phone verification process and introduced new challenges, including reCAPTCHA (utilizing an advanced risk-analysis engine developed by Google), to give us additional tools to validate that a human is in control of an account. We have enhanced our capabilities to link together accounts that were formed by the same person or that are working in concert. And we are improving how we detect when accounts may have been hacked or compromised.

With our improved capabilities, we are now detecting and blocking approximately 523,000 suspicious logins each day that we believe to be generated through automation. In December 2017, our systems identified and challenged more than 6.4 million suspicious accounts globally per week—a 60% increase in our detection rate from October 2017. Over three million of those accounts were challenged upon signup, before their content or engagements could impact other users. Since June 2017, we also suspended more than 220,000 malicious applications for API abuse. These applications were collectively responsible for more than 2.2 billion Tweets in 2017. We plan to continue building upon our 2017 improvements, including through collaboration with our peers and investments in machine-learning capabilities that help us detect and mitigate the effect on users of fake, coordinated, and malicious automated account activity.

We have also observed the expansion of malicious activity on our platform from automated accounts to human-coordinated activity, which poses additional challenges to making our platform safe. We are determined to meet those challenges and have been successful in addressing such abusive behavior in other contexts. We are committed to leveraging our technological capabilities in order to do so again by carefully refining and building tools that respond to signals in the account behavior.

Those tools have also been successful at detecting and removing terrorist content on our platform. For example, as of September 2017, 95% of account suspensions for promotion of terrorist activity were accomplished using our existing proprietary detection tools—up from 74% in 2016. These tools focus on indicia of violating activity beyond the content of the Tweet. Although they have proved successful, our efforts to address terrorist content on our platform do not end with investments in our proprietary detection tools. We recognize that the spread of terrorist and extremist content online is not unique to Twitter, and we are committed to collaborating with our industry peers to address this shared thread. Accordingly, 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 GIFTC announced a multimillion dollar commitment to support research on terrorist abuse of the Internet, and how governments, tech companies, and civil society can respond effectively. We are looking to establish a network of experts that can develop these platform-agnostic research questions and analysis that considers a range of geopolitical contexts. The GIFCT opened a call for proposals in December 2017, and we look forward to sharing further details of the initial projects this year.

The GIFCT has created a shared industry database of "hashes"—unique digital "fingerprints"—for violent terrorist imagery or terrorist recruitment videos or images that have been removed from our individual services. 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 even block extremist content before it is posted in the first place. The database now contains more than 40,000 hashes. Instagram, Justpaste.it, LinkedIn, Oath, and Snap have also joined this initiative, and we are working to add several additional companies in 2018. Twitter also participates in the Technology Coalition, which shares images to counter child abuse.

As part of our work with the GIFCT, we have hosted more than 50 small companies at workshops through the Tech Against Terrorism initiative, our partners under the UN CounterTerrorism Executive Directorate. Twitter believes that this partnership will provide a unique opportunity for us to share our knowledge and technical expertise with smaller and emerging companies in the industry and for all industry actors to harness the expertise that has been built up in recent years.

We have also focused on NGO outreach and, since 2013, and have participated in more than 100 Countering Violent Extremism training and events around the world, including in Beirut, Bosnia, Belfast and Brussels, and summits at the White House, at the United Nations, London, and Sydney. Twitter has partnered with groups like the Institute of Strategic Dialogue, the Anti-Defamation League and Imams Online to bolster counterspeech that offers alternatives to radicalization. As a result of that work, NGOs and activists around the world are able to harness the power of our platform in order to offer positive alternative narratives to those at risk and their wider communities.

Finally, in addressing abuse directed at users on the platform, context matters. A turn of phrase can be playful or offensive, depending on the circumstance, topic, and author. This means we need more nuanced and creative approaches to our machine learning models in order to address abusive activity at scale. One example where we have made progress is in our improving ability to action reports of abuse by witnesses (instead of by victim directly). By looking at various signals, including the relationship and activity between the reported abuser and reported victim, we can better identify, escalate, and take action against instances of abuse.

Question 2. Can you walk through your track record of removing terrorist content?

Response. As noted above, we have made considerable inroads against the proliferation of terrorist content on our platform. For example, in February 2016 when we first started sharing metrics for our enforcement efforts, we announced that, since the middle of the preceding year, we had suspended more than 125,000 accounts for threatening or promoting terrorist acts. See https://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html. By August 2016, we announced that we had suspended an additional 235,000 accounts for violating Twitter policies related to the promotion of terrorism.
https://blog.twitter.com/official/en_us/a/2016/anupdate-on-our-efforts-to-combat-violent-

extremism.html. We also announced at that time that our daily suspension records increased by more than 80% compared to the previous year, and that our response time for suspending reported accounts decreased dramatically.

We made additional improvements the following year. As we noted in our September 2017 Transparency Report, for the reporting period between January 1 and June 30, 2017, we suspended nearly 300,000 accounts for violations of Twitter policies prohibiting the promotion of terrorism. Of those suspensions, 95% were accomplished using our proprietary tools—up from 74% in 2016. Approximately75% of those accounts were suspended before posting their first Tweet. In total, between August 1, 2015 and June 30, 2017, we suspended nearly 1 million accounts for violating Twitter rules and policies prohibiting the promotion of violence or terrorist content.

Question 3. What is the next challenge on the Common Vulnerabilities and Exposures (CVE) front? How do we empower smaller platforms that the terrorists are moving to?

Response. As noted above, we plan to continue building upon our 2017 improvements, including through collaboration with our peers and investments in machine-learning capabilities that help us detect and mitigate the effect on users of fake, coordinated, and malicious automated account activity.

We have observed the expansion of malicious activity on our platform from automated accounts to human-coordinated activity, which poses additional challenges to making our platform safe. We are determined to meet those challenges and have been successful in addressing such abusive behavior in other contexts. We are committed to leveraging our technological capabilities in order to do so again by carefully refining and building tools that respond to signals in the account behavior.

Those tools have also been successful at detecting and removing terrorist content on our platform. For example, as of September 2017, 95% of account suspensions for promotion of terrorist activity were accomplished using our existing proprietary detection tools—up from 74% in 2016. These tools focus on indicia of violating activity beyond the content of the Tweet.

Although they have proved successful, our efforts to address terrorist content on our platform do not end with investments in our proprietary detection tools. We recognize that the spread of terrorist and extremist content online is not unique to Twitter, and we are committed to collaborating with our industry peers to address this shared thread. Accordingly, 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 GIFTC announced a multimillion dollar commitment to support research on terrorist abuse of the Internet and how governments, tech companies, and civil society can respond effectively. We are looking to establish a network of experts that can develop these platform-agnostic research questions and analysis that considers a range of geopolitical contexts. The GIFCT opened a call for proposals in December 2017, and we look forward to sharing further details of the initial projects this year.

The GIFCT has created a shared industry database of “hashes”—unique digital “fingerprints”—for violent terrorist imagery or terrorist recruitment videos or images that have been removed from our individual services. 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 even block extremist content before it is posted in the first place. The database now contains more than 40,000 hashes. Instagram, Justpaste.it, LinkedIn, Oath, and Snap have also joined this initiative, and we are working to add several additional companies in 2018.

As part of our work with the GIFCT, we have hosted more than 50 small companies at workshops through the Tech Against Terrorism initiative, our partners under the UN CounterTerrorism Executive Directorate. Twitter believes that this partnership will provide a unique opportunity for us to share our knowledge and technical expertise with smaller and emerging companies in the industry and for all industry actors to harness the expertise that has been built up in recent years.

Response to Written Questions Submitted by Hon.
Ron Johnson
Written Questions for the Record to
Carlos Monje

Question 1. Social media companies are increasingly able to remove terrorist recruitment, incitement, and training materials before it posts to their platforms by relying on improved automated systems. Other than content removal, what else can be done to limit the audience or distribution of these dangerous materials?

Response. Twitter has been at the forefront of developing a comprehensive response to the evolving challenge of preventing terrorist exploitation of the Internet. We initially focused on scaling up our own, in-house proprietary spam technology to detect and remove accounts that promote terrorism. In early 2016, the technological tools we had at our disposal detected approximately one-third of terrorism-related accounts that we removed at that time. In 2017, 95% of account suspensions for promotion of terrorist activity were accomplished using our existing proprietary detection tools—up from 74% in 2016. Approximately 75% of those accounts were suspended prior to sending their first Tweet. In total, since 2015, we have suspended nearly a million accounts that we determined violated our terms of service. In December 2016, for example, we took steps toward a hash-sharing agreement with Facebook, Microsoft, and YouTube, intended to further curb the spread of terrorist content online. Pursuant to this agreement, the four companies created an industry database of "hashes"— unique digital "fingerprints"—for violent terrorist imagery or terrorist recruitment videos or images that we have removed from our services. By sharing this information with each other, we may use the shared hashes to help identify potential terrorist content on our respective hosted consumer platforms.

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 GIFTC announced a multimillion dollar commitment to support research on terrorist abuse of the Internet and how governments, tech companies, and civil society can respond effectively. We are looking to establish a network of experts that can develop these platform-agnostic research questions and analysis that considers a range of geopolitical contexts. The GIFCT opened a call for proposals last month, and we look forward to sharing further details of the initial projects early in 2018.

The GIFCT has created a shared industry database of "hashes"—unique digital "fingerprints"—for violent terrorist imagery or terrorist recruitment videos or images that have been removed from our individual services. 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 even block extremist content before it is posted in the first place. The database now contains more than 40,000 hashes. Instagram, Justpaste.it, LinkedIn, Oath, and Snap have also joined this initiative, and we are working to add several additional companies in 2018. Twitter also participates in the Technology Coalition, which shares images to counter child abuse.

As part of our work with the GIFCT, we have hosted more than 50 small companies at workshops through the Tech Against Terrorism initiative, our partners under the UN CounterTerrorism Executive Directorate. Twitter believes that this partnership will provide a unique opportunity for us to share our knowledge and technical expertise with smaller and emerging companies in the industry and for all industry actors to harness the expertise that has been built up in recent years.

We have also focused on NGO outreach and, since 2013, and have participated in more than 100 Countering Violent Extremism training and events around the world, including in Beirut, Bosnia, Belfast and Brussels and summits at the White House, at the United Nations, London, and Sydney. Twitter has partnered with groups like the Institute of Strategic Dialogue, the Anti-Defamation League and Imams Online to bolster counterspeech that offers alternatives to radicalization. As a result of that work, NGOs and activists around the world are able to harness the power of our platform in order to offer positive alternative narratives to those at risk and their wider communities.

Question 2. Terrorist how-to guides are protected by the First Amendment in the United States, but violate the content policies of many social media companies as well as the laws of some international partner nations. What countries have laws that go beyond your company's content policies and can you give examples of how you have worked with those countries to de-conflict those differences?

Response. The Twitter Rules prohibit violent threats and the promotion or incitement of violence, including terrorism. Twitter is committed to removing such content swiftly from the platform. In addition, our Hateful Conduct policy is designed to protect users from harassment on the basis of protected categories, such as race, ethnicity, national origin, gender identity, age and religion. Examples of hateful conduct that we do not tolerate include targeting users with: (1) harassment; (2) wishes for the physical harm, death, or disease of individuals or groups; (3) references to mass murder, violent events, or specific means of violence in which or with which such groups have been the primary victims; (4) behavior that incites fear about a protected group; and (5) repeated and/or or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories. We have also updated our policies to clearly prohibit users who affiliate with organizations that—whether by their own statements or activity both on and off the platform—use or promote violence against civilians to further their causes.

We have established channels for law enforcement agencies to request removal of content that may be illegal under the requesting jurisdiction's laws. Specifically, Twitter publishes global guidelines for law enforcement personnel that explain our policies and the process for submitting requests for content removal. See https://help.twitter.com/en/rules-andpolicies/twitter-law-enforcement-support. We accept requests from law enforcement agencies in countries in which

Twitter operates, and we evaluate each request and, if appropriate, we will take action against the content at issue within the jurisdiction from which the removal request originated. As part of our commitment to transparency, since 2012, Twitter has published biannual Transparency Reports, reflecting the number of requests that we have received for user information and content removal on a per-country basis. See https://transparency.twitter.com.

Those reports indicate the number of requests that we have received and the number of requests with which we have complied.

Response to Written Questions Submitted by Hon.
Senator Wicker
Written Questions for the Record to
Carlos Monje

Question 1. Has Twitter placed any restrictions on the U.S. Government's use of publicly available information on your platform? If yes, please describe what those restrictions are, why they have been imposed and on which U.S. Government agencies?

Response. Twitter is a public platform. When users choose to share information by posting it to their public profile, the information is available to anyone who visits those users' profiles. With respect to Twitter's application programming interface ("API") (public and commercial), through which we provide developers and other third parties access to subsets of public Twitter content, all users of our developer products must comply with Twitter's developer terms and policies. Those policies include long-standing provisions that prohibit, among other things, the use of Twitter data for surveillance purposes or for purposes in contravention of the Universal Declaration of Human Rights. While Twitter works closely with its developer community to address questions and investigate instances of potential abuse, each developer is responsible for compliance with Twitter's applicable policies.

Twitter maintains strong working relationships with law enforcement. We publish guidelines for law enforcement personnel that explain our policies and the process for submitting requests for information. We regularly respond to law enforcement requests, have a dedicated 24/7 response team for that purpose, and have developed a user-friendly online submission form to streamline responses to law enforcement agencies through properly scoped valid legal process. There are also a number of news alert products that are available and used by law enforcement, including the Federal Bureau of Investigation.

Question 2. Are U.S. Government agencies or intelligence organizations permitted to search for or monitor—either directly with Twitter or through third-party aggregators— counterterrorism information or specific Twitter accounts that are likely affiliated with terrorist organizations within the publicly available content found on Twitter's platform? If not, why? Please explain.

Response. The answer to Question 2 has been provided in response to Question 1.

Question 3. Are companies (such as casinos) allowed to monitor—either directly with Twitter or through third-party aggregators—specific Twitter accounts that have made public threats against their venues or staff?

Response. The answer to Question 2 has been provided in response to Question 1.

Question 4. Does Twitter have any policies that prohibit the use of its data, by any public or private third-party, for counterterrorism analyses focused on terrorist organizations? If so, can you please explain the purpose of that policy and the parameters of it?

Response. The answer to Question 2 has been provided in response to Question 1.

Question 5. During the hearing, Mr. Monje testified that Twitter works with law enforcement through the "proper legal process". Please describe the legal process to which Mr. Monje was referring and how it applies to law enforcement's use of Twitter's aggregate user data.

Response. As we noted above in response to Question 1, Twitter maintains strong working relationships with law enforcement. We publish guidelines for law enforcement personnel that explain our policies and the process for submitting requests for information. See https://help.twitter.com/en/rules-and-policies/twitter-law-enforcement-support. We regularly respond to law enforcement requests, have a dedicated 24/7 response team for that purpose, and have developed a user-friendly online submission form to streamline responses to law enforcement agencies through valid legal process. Before launching this system to all U.S. law enforcement agencies, we conducted a pilot with the Federal Bureau of Investigation. We have begun rolling out this tool for global use.

In addition, we have offered and conducted training sessions to law enforcement officials to familiarize them with our policies and procedures. In 2017, we have attended and provided training at a national conference for investigators of crimes against children, training events for FBI legal attachés posted to U.S. embassies abroad, and other conferences with the participation of federal, state and local law enforcement. We continue to build upon and invest in our law enforcement outreach and training. And we welcome feedback from law enforcement experts and professionals about how we can improve our systems.

We regularly and directly engage with law enforcement officials on a wide range of issues, including extremist content online. We receive and respond to "Internet Referral Unit" reports of extremist content. Our recently published Transparency Report for the first half of 2017 details the statistics of those responses. See https://blog.twitter.com/official/en_us/topics/company/2017/New-Data-Insights-Twitters-Latest-Transparency-Report.html. In addition, we receive briefings from government experts on terrorist use of online platforms, which help inform our proactive efforts.

Law enforcement requests to Twitter must comply with applicable laws in the jurisdiction where they are issued. For federal law enforcement, this includes the Electronic Communications Privacy Act 18 U.S.C. 2510 et seq. These requirements only apply to data sought from Twitter. If law enforcement is able to access publicly available information from the Twitter service they may do so subject to any other legal or policy restrictions that may apply to their conduct (e.g., Department of Justice guidance). If law enforcement seeks access to Twitter data via our API directly or through a third party developer, they must do so in a manner that complies with the applicable Twitter terms and policies for our API.

Question 6. Does a U.S. Government agency have to obtain a warrant (or go through a similar legal process as discussed in Question #5) to search publicly available information found on Twitter? If yes, why? If no, does Twitter allow U.S. Government agencies to gain access to publicly available information on its platform through third-parties that have purchased aggregate user data from Twitter?

Response. The answer to Question 6 has been provided in response to Question 1.

Question 7. Twitter’s platform allows users to “follow” other users. In your view, what is the difference between “following” someone and “surveilling” someone?

Response. A user may follow another account holder via the Twitter service. An account holder may view their list of followers at any time. The account holder may take a range of actions in response to receiving a “follow” from another user. They may decide to follow that user in return. They may also choose to block that follower, preventing the follower from viewing in their timeline content posted that account holder or receiving notifications of posts by the account holder Twitter users may also choose to make their account private so as to restrict new followers to those that they have expressly allowed as followers. These choices, including the ability to restrict followers and the transparency inherent in our platform, are important aspects of the Twitter service.