January 19, 2018

Twitter, Inc.

Senate Committee on the Judiciary,
Subcommittee on Crime and Terrorism Hearing on Extremist Content and Russian Disinformation Online: Working to Find Solutions, October 31, 2017

Sean Edgett's Answers to Questions for the Record

QUESTIONS FOR THE RECORD—CHAIRMAN GRASSLEY

1. **To follow up on a request made during the hearing, please provide a detailed written update on what internal investigations have found regarding all accounts, advertisements, and posts with connections to Russia that relate to the lead-up and aftermath of the 2016 presidential campaign.**

As we explained in connection with the Committee hearing, we conducted a retrospective review that included two components. First, we reviewed organic activity on the core Twitter product during the period between September 1, 2016 and November 15, 2016. Our goal has been to identify election-related content that appears to have originated from automated accounts or from human-coordinated activity associated with Russia. We have attached as Appendix 1 a detailed written update on that component of our review.

In the second part of our review, we focused on determining whether or how malicious Russian actors may have sought to abuse our platform using advertising. To evaluate the scope and impact of election-related advertisements by Russian actors, we used a custom-built machine-learning model and conducted a manual review of nearly 6,500 accounts to determine if those accounts had any Russian-specific characteristics or non-Russian international characteristics.

We identified nine accounts that both exhibited at least one characteristic of a Russian-linked account and promoted election-related Tweets that violated our ads policies, including Twitter policies prohibiting inflammatory or low-quality content.

Of those nine accounts, @RT_COM and @RT_America represented the vast majority of the promoted Tweets, spend and impressions for the suspect group, with a combined advertising spend of $516,900 in 2016. Of that amount, $234,600 was spent on ads that were served to users in the U.S. The two accounts promoted 1,912 Tweets and generated approximately 192 million impressions across all ad campaigns, with approximately 53.5 million of those impressions generated by U.S.-based users.

The remaining seven accounts that our review identified represented small, apparently unconnected actors who, in 2016, spent a combined total of $2,282 on advertising, ran 404 promoted Tweets, and generated a total of 2.29 million impressions across all ad campaigns (222,000 of which were by U.S.-based users). We have since off-boarded these advertisers.

2. Globally, leaders and law enforcement—including in the United States and in Europe—have been highly critical of Facebook and other tech companies for not doing more to combat extremist content online. In June 2017, Google's General Counsel said, "Terrorism is an attack on open societies, and addressing the threat posed by violence and hate is a critical challenge for us all. Google and YouTube are committed to being part of the solution. We are working with government, law enforcement and civil society groups to tackle the problem of violent extremism online. There should be no place for terrorist content on our services."

a. How <u>exactly</u> is tech working to help (1) government, (2) law enforcement, and (3) civil society groups tackle the problem of extremist content online?

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 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 their 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 Counter-Terrorism 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 also focused on NGO outreach and, since 2013, 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.

b. Please list all law enforcement agencies, domestic and international, with which tech is partnering or assisting and describe the assistance.

i. If no such partnership or assistance exists, please explain why not?

1. Regardless of current status, what are future plans in this area?

ii. What <u>specific</u> proposals does tech have to assist, or further assist, law enforcement in this area?

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 response 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 are now opening this tool for global use.

We have offered and conducted training sessions to law enforcement officials to familiarize them with our policies and procedures. This year alone, 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.

3. What more should tech be doing to incorporate new and existing technologies to independently, as a matter of corporate responsibility, detect, remove, and report – both to their users and law enforcement – extremist content?

As detailed in our latest Government Terms of Service Report, *available at* https://transparency.twitter.com/en/gov-tos-reports.html, we strive to improve our internal efforts to remove the burden from users to report content promoting terrorism by identifying and removing terrorist content through the use of proprietary technology. We also work with our industry peers to identify and invest in technologies that further strengthen the ability of all platforms to share and leverage a variety of signals for surfacing content promoting terrorism, and we are committed to continuing this collaboration to identify and share best practices and effective strategies for countering terrorist content online.

We respond to law enforcement requests related to emergencies involving imminent harm to persons on a 24/7 basis. In the rare event that we come across information that we believe directly involves or threatens imminent harm to a person, we work to bring such threats to the attention of appropriate law enforcement agencies as soon as possible.

Twitter is continually improving its notices to users and the public about Terms of Service and Twitter Rules violations. It has been our long-standing practice to note publicly when accounts are suspended for violations of our rules. And we have recently taken steps to further explain what other enforcement actions we may pursue against content and accounts that violate our rules, including through expanding our responses to user complaints and to users against whom we have taken action.

4. In March 2017, advertisers left Google's video-sharing platform YouTube after finding that their ads were appearing next to extremist content. Google had allowed some videos to remain on YouTube with a "warning" label — this has not been done with other content such as pornography or copyright violations that defy YouTube's Terms of Service. In June 2017, Google and YouTube introduced new measures to curb extremist video online.

a. How is tech differentiating between extremist content to be removed and content that needs a warning label?

The Twitter Rules prohibit violent threats and the promotion or incitement of violence including terrorism. Twitter removes such content swiftly from the platform. As described further below in response to question 4(b), we have taken a number of steps to detect and remove terrorist content.

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 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 recognize that context matters, and we take into account the intention of the user in sharing media as we assess graphic content. Although we may permit some forms of graphic violence or adult content in Tweets when they are marked as sensitive media, users may not include this type of content in live video, or in profile or header images. Additionally, we may sometimes require users to remove media containing excessively graphic violence out of respect for the deceased and their families if we receive a request from their families or an authorized representative.

In addition to this policy, we enable content control from the user side. For example, users can choose to activate "safe search" mode, which excludes potentially sensitive content, along with accounts the user chose to mute or has blocked, from the user's search results. Users also have the option to activate our quality filter, which filters lower-quality content from their notifications. Such content includes, for example, duplicate Tweets or content that appears to be maliciously automated. Both the safe search and the quality filter are default settings, which are utilized by more than 97% of our users.

Twitter has additional restrictions on ads content. Our ads policies apply not only to the content we allow advertisers to promote on the platform, but also to the types of content on

which ads are served. We use a combination of machine-learning models, manual human review, and user reports to determine whether ads are served in safe environments across Twitter. We categorize content using both visual and text-based machine learning, and we supplement those processes with a manual human review of the content. We also give our users the ability to self-identify their content as potentially sensitive or to flag others' content as not adhering to our policies. User-based signals are incorporated into and calibrate our machine-learning models to further improve them.

b. What is tech doing to remove extremist content that violates terms of service?

We have taken a number of steps to combat violent extremism. Our efforts continue to drive meaningful results, including a significant shift in this type of activity off of Twitter.

For example, we increased the size of Twitter teams that review reports, thereby significantly reducing our response time to such content. We also look into other accounts similar to those reported and leverage proprietary spam-fighting tools to surface other potentially violating accounts for review. Those efforts have proved successful: we have seen an increase in account suspensions as well as a shifting off of the platform of this type of activity.

In part through this work, between August 2015 and June 2017, Twitter has suspended more than 935,000 accounts for the promotion of terrorism. And over 95% of those accounts were detected using our in-house, enhanced detection tools. We observed an 80% reduction in accounts reported by governments compared to the previous reporting period of July 1, 2016 through December 31, 2016. Notably, government requests accounted for less than 1% of account suspensions for the promotion of terrorism during the first half of 2017.

We continue to work with law enforcement agencies that seek assistance with investigations to prevent suspected terrorist attacks. Specifically, as we outline in our Law Enforcement Guidelines, Twitter responds to valid legal process issued in compliance with applicable law, and we report on these government requests (in aggregate) in our biannual Transparency Report.

Our global Public Policy team has expanded its partnerships with organizations working to counter violent extremism online. Those partnerships are intended to empower credible nongovernmental voices against violent extremism and include respected organizations such as Parle-moi d'Islam (France), Imams Online (UK), Wahid Foundation (Indonesia), The Sawab Center (United Arab Emirates), and True Islam (U.S.).

As many industry actors and experts in the field have long recognized, there is no magic algorithm for identifying terrorist-related content online. Accordingly, global Internet platforms must make challenging calls based on very limited information and guidance. Despite those challenges, Twitter is committed to enforcing our rules aggressively in this area and to engaging with law enforcement agencies, governmental authorities, and other relevant organizations to find solutions to this critical issue and promote powerful counterspeech narratives.

c. What is tech doing to define, catalogue, and categorize content it has removed and share it:

i. i. within any family of apps,

ii. ii. other Internet and social media platforms,

iii. iii. and also with law enforcement?

In June 2017, we announced the formation of the Global Internet Forum to Counter Terrorism, a collaboration with Facebook, YouTube, and Microsoft. That initiative helps us continue to make our hosted consumer services hostile to terrorists and violent extremists. Twitter believes that, by working together and sharing the best technological and operational elements of our individual efforts, we can have a greater impact on the threat of terrorist content online.

The forum builds on other joint initiatives, including the EU Internet Forum and the Shared Industry Hash Database, discussions with the UK government, and the conclusions of the recent G7 and European Council meetings. It formalizes and structures existing and future areas of collaboration between our companies and fosters cooperation with smaller tech companies, civil society groups, and academic institutions, governmental organizations, and supranational bodies such as the EU and the UN. Our work as part of the initiative is as follows:

Technological solutions: Our companies work together to (1) refine and improve existing joint technical work, such as the Shared Industry Hash Database; (2) exchange best practices as we develop and implement new content detection and classification techniques using machine learning; and (3) define standard transparency reporting methods for terrorist content removals.

Research: We have commissioned research to inform our counter-speech efforts and guide future technical and policy decisions around the removal of terrorist content.

Knowledge-sharing: We work with counter-terrorism experts, including governments, civil society groups, academic institutions, and other companies to engage in shared learning about terrorism. Through a joint partnership with the UN Security Council Counter-Terrorism Executive Directorate and the ICT4Peace Initiative, we are establishing a broad knowledge-sharing network both to engage with smaller companies to help them develop the technology and processes necessary to tackle terrorist and extremist content online, and to develop best practices through existing partnerships with organizations such as the Center for Strategic and International Studies, Anti-Defamation League, and Global Network Initiative. Through such collaborations, we strive to identify how best to counter extremism and online hate, while respecting freedom of expression and privacy, foster education by allowing us to learn from and contribute to one another's counterspeech efforts, and discuss how to further empower and train civil society organizations and individuals who may be engaged in similar work and support ongoing efforts such as the Civil Society Empowerment Project.

Our partnership with law enforcement is detailed in question 2(b) above.

5. Last year Google, Facebook, Twitter, and Microsoft announced a "hashing coalition" designed to share signatures of known extremist content and to remove this content.

a. How large is the signature database?

b. How much is the signature database growing on a weekly and on a monthly basis?

c. How much content is actually being found and removed?

d. Is the content deployed on all platforms (e.g., for Google, on YouTube, Google Drive, Google Photos)?

e. What is being done to share this hashing coalition with law enforcement, both domestically and globally?

On June 26, 2017, Facebook, Microsoft, Twitter and YouTube announced the formation of the Global Internet Forum to Counter Terrorism. We believe that by working together and sharing the best technological and operational elements of our individual efforts, we can have a greater impact on stifling the spread of terrorist content online. The new forum builds on initiatives including the European Internet Forum and the Shared Industry Hash Database. It seeks to formalize and structure existing and future areas of collaboration between our companies and foster cooperation with smaller tech companies, civil society groups, as well as academics, governments and supranational bodies, such as the EU and the UN.

The scope of the work of the GIFCT will evolve over time—we intend to be responsive to the ever-evolving terrorist and extremist tactics. One workstream is devoted to technological solutions. Each company has worked together to refine and improve existing joint technical work, most importantly the Shared Industry Hash Database.

On December 5, 2016, Facebook, Microsoft, Twitter, and YouTube announced a commitment to create 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 their 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 and Snap have also joined this joint initiative, and we are working to add several additional companies over the coming months.

6. Tech has promised solutions based on artificial intelligence to identifying the problem of extremist content online.

a. How large are the current teams working on this technology?

b. What is the timeline of deployment?

c. What is the current accuracy of detection and false alarms?

d. Will these technologies be shared across the tech industry?

e. Will these technologies be shared with law enforcement?

The answers to question 6(a)-(e) have been provided in response to questions 2 through 4.

QUESTIONS FOR THE RECORD—SENATOR FEINSTEIN

Information on Russian-Connected Accounts

1. Accounts and ads created by the Internet Research Agency (IRA) or other Russia-linked entities have been identified to varying degrees by your company.

a. How do you know whether all accounts tied to the IRA or other suspected Russian-connected entities that are using your platforms have been identified?

To identify IRA-linked and related accounts, Twitter conducted a comprehensive platform-wide review based on account behavior and account identity features of known IRA accounts to identify linked accounts. In the fall, we reported that we had found 2,752 IRA-linked accounts. We noted that this was an active area of inquiry and that we planned to update the Committee as we continued the analysis. Through that continued review, we have identified 1,062 more IRA-linked accounts during the relevant period, for a total of 3,814 such accounts. Those accounts posted 175,993 Tweets, approximately 8.4% of which were election-related. Many of their Tweets—just over 53%—were automated. All 3,814 IRA-linked accounts were suspended for Terms of Service violations, and all but a few compromised accounts that have subsequently been restored to their legitimate account owners remain suspended.

b. Have you found other troll farms or other organizations like the IRA? (If so, please describe those organizations and their use of your social media platform.)

The expansion of coordinated activity from automated to human-driven poses additional challenges to making our platform safe. Our work in this arena is ongoing and we continue to monitor, review, and analyze suspicious activity that exhibits similar patterns and characteristics to the IRA accounts. At this time, we are not aware of any other troll farms originating from Russia or government-sponsored organizations similar to the IRA that have been active on our platform. But we are deploying all available Twitter tools and resources to identifying such accounts, and we are continually exploring third-party leads and information from the public domain in order to conduct in-depth reviews of accounts that may fall into this category.

c. What criteria do you use to identify inauthentic accounts?

Although Twitter permits users to Tweet under any name they choose, we take action against accounts that are inauthentic because they are maliciously automated or part of a human-coordinated effort to engage in malicious or abusive conduct. Twitter's approach to addressing the spread of malicious automation and inauthentic accounts on our platform is to focus on problematic behavior and abuse, not primarily on the content that such accounts attempt to disseminate. We are committed to addressing the spread of misinformation on our platform—and to prevent future attempts of interfering with U.S. elections—but we recognize that spam and malicious automation are not limited to political content and can undermine the positive user experience we seek to offer irrespective of the content.

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 which are likely to be maliciously automated or

acting in an automated and coordinated fashion. 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.

d. What do you do with inauthentic accounts once you've identified them?

Our systems are built to detect malicious automated and spam accounts across their lifecycles, including detection at the account creation and login phase and detection based on unusual activity (e.g., patterns of Tweets, likes, and follows). Our ability to detect such activity on our platform is bolstered by internal, manual reviews conducted by Twitter employees. Those efforts are further supplemented by user reports, which we rely on not only to address the content at issue but also to calibrate our detection tools to identify similar content as spam.

Once our systems detect an account as generating spam, we can take action against that account at either the account level or the Tweet level. Depending on the mode of detection, we have varying levels of confidence about our determination that an account is violating our rules. We have a range of options for enforcement; generally, the higher our confidence that an account is violating our rules, the stricter our enforcement action will be, with immediate suspension as the harshest penalty. If we are not sufficiently confident to suspend an account on the basis of a given detection technique, we may challenge the account to verify a phone number or to otherwise prove human operation, or we may flag the account for review by Twitter personnel. Until the user completes the challenge, or until the review by our teams has been completed, the account is temporarily suspended; the user cannot produce new content (or perform actions like Retweets or likes), and the account's contents are hidden from other Twitter users.

We also have the capability to detect suspicious activity at the Tweet level and, if certain criteria are met, to internally tag that Tweet as spam or otherwise suspicious. Tweets that have been assigned those designations are hidden from searches, do not count toward generating trends, and generally will not appear in feeds unless a user follows that account. Typically, users whose Tweets are designated as spam are also put through the challenges described above and are suspended if they cannot pass.

For safety-related Terms of Service ("TOS") violations, we have a number of enforcement options. For example, we can stop the spread of malicious content by categorizing a Tweet as "restricted pending deletion," which requires a user to delete the Tweet before the user is permitted to continue using the account and engaging with the platform. So long as the Tweet is restricted—and until the user deletes the Tweet—the Tweet remains inaccessible to and hidden from all Twitter users. The user is blocked from Tweeting further unless and until he or she deletes the restricted Tweet. This mechanism is a common enforcement approach to addressing less severe content violations of our TOS outside the spam context; it also promotes education among our users. More serious violations, such as posting child sexual exploitation or promoting terrorism, result in immediate suspension and may prompt interaction with law enforcement.

2. For the accounts your companies have identified as linked to the Internet Research Agency (IRA):

a. How many people followed these accounts?

In total, IRA-linked accounts had approximately 2.7 million followers. Of those, @TEN_GOP had the highest number of followers—152,099 in total prior to suspension.

b. What did they see when they went to the IRA webpages?

Each IRA-linked account had a different profile page. By following those accounts, users gained access to public profile information, including username, location (if specified), the accounts that were following the IRA-linked account and those accounts that the IRA-linked account was following, profile images, and any publicly available content those accounts generated.

c. How did the IRA messages spread on your platforms?

In the fall, we reported that we had found 2,752 IRA-linked accounts. We noted that this was an active area of inquiry and that we planned to update the Committee as we continued the analysis. Through that continued review, we have identified 1,062 additional IRA-linked accounts during the relevant period, for a total of 3,814 such accounts. Those accounts posted 175,993 Tweets, approximately 8.4% of which were election-related. Many of their Tweets—just over 53%—were automated. All 3,814 IRA-linked accounts were suspended for Terms of Service violations, and all but a few compromised accounts that have subsequently been restored to their legitimate account owners, remain suspended.

Between September 1 and November 15, 2016, Tweets from IRA-linked accounts received a total of 351,632,679 impressions within the first seven days after posting the Tweet and were Retweeted 4,509,781 times.

d. You said that you removed all posts by IRA, but did you also take down versions of those posts shared by other users?

Once the IRA-linked accounts were suspended, all Retweets (native shares) of content from those accounts were automatically removed as well.

e. Have you confirmed that the IRA has not been able to create new inauthentic (or "fake") accounts once existing ones are found and taken down? What are you doing to make sure that these copycats are also taken down?

We have taken steps to block future registrations related to IRA-linked accounts that we have already detected and suspended, and we continue to monitor activity in order to ensure that such accounts and similar accounts do not gain access to Twitter.

AD Purchasing

3. Did your company have any restrictions before the 2016 election on who could buy ads?

Yes. Twitter believes that freedom of expression is paramount and we want to ensure that our users feel safe when they engage and interact with others on our site. We also want to ensure advertisers bring value to our users and enhance rather than detract from their experience. Because advertisers purchase Twitter ads and can present them to an audience beyond their followers, there are greater limitations on the type of content that can be promoted with Twitter ads compared to organic content that our users generate.

Some of the criteria we take into account in determining whether accounts may run ads on our platform include, but are not limited to, whether the account (1) has a valid billing address; (2) runs the proposed ad in a Twitter-supported language (Twitter only supports a select number of languages); (3) has not been deactivated or suspended organically or as an advertiser account in the past; (4) is not from an embargoed country; (5) and the content and business model are compliant with our Advertising Policies (*available at* https://support.twitter.com/articles/20169693#).

4. Is there any way for your company to tell if an ad buyer is a mere intermediary or proxy for someone else? For example, can your company detect when an ad buyer is serving as a proxy for the Russian government or a Russian troll farm that actually paid for the ad campaign?

As part of our ads transparency and electioneering ads efforts announced in October 2017, Twitter will require advertisers to go through an onboarding process, which will obligate them to provide information about how they are funding their media buys.

5. What are you doing to make sure that you know when foreign state actors buy ads? What are you doing to disclose that fact to other users?

As part of the electioneering advertising onboarding process, we will clearly label ads to make it possible for our users to quickly identify the nature of the ad. In addition, we will include links to more information about the ad, including the identity of the advertiser and the source of the reported source of funding for the ad. For advertisers who do not self-identify but who run electioneering ads, we will use a combination of machine-learning models and human manual review to detect and halt these advertisers until they have correctly onboarded with us as an electioneering advertiser.

To make it clear when a user is viewing or engaging with content considered to be an electioneering ad, our policy will require that advertisers that meet the definition of electioneering identify their campaigns as such. We will also change the interface of such ads and include a visual political ad indicator (*see, e.g.*, Fig. 1 below).

Fig. 1: Template for New Electioneering Ad

6. What is your company doing to identify businesses and organizations that run election ads?

The requested information has been provided in response to question 5.

7. Do you believe that other platform users should be notified regarding the identity of individuals or entities purchasing election ads on your platform?

Twitter believes that providing more information about who is funding electioneering ads on our platform will increase transparency and user literacy about the ads they are viewing on our platform. Our new Ads Transparency Center is designed to achieve that goal.

8. What specific documentation are you using to verify that ad purchasers are who they say they are?

We use a combination of billing entity details as well as information from third parties and financial/sanctions databases to verify that the organization and/or individual is legitimate and should be allowed to run ads on our platform according to our policies.

Shutting Down Suspect Accounts

9. What criteria does your company use to determine if an account should be shut down?

When a user violates the Twitter Rules, we can take a variety of actions depending on the nature of the violation and the user's violation history. For example, if the user Tweets content that violates our policies, creates an account that impersonates another user, or otherwise violates our rules, we can temporarily suspend the account while we ask the user to edit their profile or

delete the prohibited content. We keep the account suspended until the user complies with our request.

We also have the ability to place an account in read-only mode, which limits the user's ability to Tweet, Retweet, or like content from 12 hours to 14 days—or until the user passes a challenge—depending on the severity of the violation. For accounts with suspicious logins or aberrant Tweeting activity, we can temporarily suspend the account and require that the account go through a number of challenges—*i.e.*, reCAPTCHA, phone verification—in order to restore it.

We can also permanently suspend the account. A permanent suspension is our most severe enforcement action because doing so removes the account from global view; the user is not allowed back on the platform—either through the suspended account or by creating new accounts.

10. What steps are being taken to prevent your platforms from being used to incite violence or lawlessness?

Twitter Rules prohibit making specific threats of violence or wishing for serious physical harm, death, or disease of an individual or group of people. This includes, but is not limited to, threatening or promoting terrorism. The rules further prohibit users from engaging in the targeted harassment of another individual, or inciting other people to do so.

At this time, Twitter removes content that includes a violent threat or wish of serious physical harm. And we recently clarified that prohibited violent content includes content that glorifies or condones acts of violence that result in death or serious physical harm. Furthermore, as of December 18, 2017, we began suspending accounts for organizations that use violence to advance their cause and we will not allow users to display names that are abusive.

In addition, we now prohibit the use of hateful imagery in avatars or profile headers. Accounts containing such content are hidden and not viewable by other users.

11. Are you considering changes to your terms of service to address this content?

Twitter has introduced a number of new rules governing this activity, which took effect on December 18, 2017. Those include, among other things, additional updates on the rules governing violence, harassment, and the use of hateful content and imagery, as well as changes designed to target accounts that are affiliated with violent groups or those that display hateful imagery and hate symbols.

Delay in Removing Fake GOP Account

12. Twitter took 11 months to close a Russian troll account that impersonated the Tennessee Republican Party. This account was particularly active, pushing out fake news and inflammatory material. The actual Tennessee Republican Party notified Twitter three times between September 2016 and August 2017 that the account was fake before Twitter finally closed the account. ("Twitter was warned repeatedly about this fake account run by a Russian Troll Farm and refused to take it down," Buzzfeed News, October 18, 2017.)

a. Why did it take Twitter nearly a year to take down an account that it was notified was fake?

We have rightfully received criticism for this oversight. We recognize that we should have acted sooner than we did and we are committed to taking swift action in the future in response to similar reports and notifications.

b. What steps are being taken to speed up the process in the future?

Once we receive a report of potential user impersonation, we investigate the reported accounts to determine if the accounts are in violation of the Twitter Rules, which prohibit such profiles. Accounts determined to be in violation of our impersonation policy, or those not in compliance with our parody, commentary, and fan account policy, are either suspended or asked to update their profile so they no longer violate our policies.

c. What steps does Twitter take to identify and remove impersonation accounts?

Twitter Rules prohibit impersonation accounts. In response to reports—from either the user who is being impersonated or their authorized representatives—Twitter takes action against accounts that deceptively impersonate another user or account. Users and non-users alike can report impersonation accounts through a dedicated form in our Help Center or directly from the impersonated account's profile on the platform.

Accounts with similar usernames or that are similar in appearance (*e.g.*, the same avatar image) are not automatically in violation of the impersonation policy. In order to be deemed as an impersonation account, the account must also portray another person in a misleading or deceptive manner. In addition, so long as they meet certain requirements—*e.g.*, indicating that the user is not affiliated with the account subject and utilizing a different username than the account subject—Twitter users are allowed to create parody, commentary, or fan accounts.

d. Have you had any other instances of fake accounts claiming to be legitimate political organizations?

Impersonation can happen to individuals and entities, including public officials or political organizations. In most cases, upon receiving a report of impersonation and verification of the identity of a reporter, Twitter is able to take action against impersonation accounts. In some cases, the accounts being reported for impersonation may be engaged in parody or satire.

Such accounts must comply with Twitter's parody policy in order to remain active on the platform. Twitter's government team has worked with many political figures and organizations to address impersonation issues and provide a higher level of assurance to the public that those accounts are authentic.

e. Can you provide an estimate of the reach of the fake Tennessee Republican Party account?

i. How many followers did it have?

ii. How many tweets?

While active, @TEN_GOP account gained 152,099 followers and posted a total 10,985 total Tweets and Retweets, of which 9,852 were original Tweets (2,092 posted during the election time period). Original Tweets from this account received more than 67 million impressions within the first seven days after posting.

iii. How did this compare to the actual Tennessee Republican Party's reach? If Russia's reach was more extensive, can you explain why that was the case?

The account associated with the Tennessee Republican Party (@TNGOP), a verified account, currently has 13,800 followers. As of November 18, 2017, it Tweeted or Retweeted 8,768 times. Of those, 200 were original Tweets that received nearly 240,000 impressions within the first seven days after posting.

Russian use of bots during the 2016 election

13. What efforts is Twitter currently taking to prevent the spread of "bots" – programs that automatically make repetitive internet posts – which can amplify disinformation?

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.

Since the 2016 election, 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. 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. We are confident that the combination of our dedicated teams, our detection tools, and other technological advancements at our disposal will prove essential in addressing malicious human-coordinated activity as well.

14. How many Russian bots were active in the run-up to the 2016 election?

In the course of our retrospective review, we identified a total of 50,258 automated accounts that were both Russian-linked and Tweeting election-related content. This represents 0.016% (approximately two one-hundredths of a percent) of the total accounts on Twitter at the time.

15. Can you give us a general sense of what these bot networks were pushing specifically during the 2016 campaign? Did you see a preference on Presidential candidates in these efforts?

Our review suggested that IRA-linked accounts Tweeted on a wide variety of topics, but primarily focused on divisive social and political issues.

16. How much circulation did they get? How widespread are the retweets, shares, or likes?

As noted in Appendix 1, as part of our retrospective review and through our supplemental analysis we have identified an additional 13,512 accounts, for a total of 50,258 automated accounts that we identified as Russian-linked and Tweeting election-related content. This represents approximately two one-hundredths of a percent (0.016%) of the total accounts on Twitter at the time. The 2.12 million election-related Tweets that we identified through our retrospective review as generated by Russian-linked, automated accounts constituted approximately one percent (1.00%) of the overall election-related Tweets on Twitter at the time. Those 2.12 million Tweets received only one-half of a percent (0.49%) of impressions on election-related Tweets, based on impressions generated within the first seven days of posting by users logged into the system. In the aggregate, automated, Russian-linked, election-related Tweets generated significantly fewer impressions relative to their volume on the platform.

Voter Suppression

17. Twitter produced images from tweets that contained false voting information (e.g., telling voters they could vote by sending a text message), all targeting likely Clinton voters. Twitter has said there were at least 918 of these voter suppression tweets. Just before the election, Twitter initially responded to complaints saying Twitter had "determined that it was not in violation of [our] Rules." At least some posts were removed only after Twitter's CEO was directly notified about these voter suppression efforts by a Twitter user. (*Id.*) ("No, you can't text your vote. But these fake ads tell Clinton supporters to do just that," Washington Post, 11/3/16.) Twitter has said that there was no obvious connection between these efforts and Russia.

a. Why was this content allowed to remain in place?

Twitter did not permit the 918 voter suppression Tweets to remain visible on the platform. Rather, Twitter labeled those Tweets as "restricted pending deletion." Assigning that label to a Tweet requires an account user to delete the Tweet before the user is permitted to continue using the account and engage with the platform. So long as the Tweet has that label—and until the user deletes the Tweet—the Tweet remains inaccessible to and hidden from all Twitter users. The user is blocked from Tweeting unless and until he or she deletes the restricted Tweet.

In addition, Twitter permanently suspended 106 accounts that were collectively responsible for 734 "vote-by-text" Tweets.

Twitter identified, but did not take action against, an additional 286 Tweets of the relevant content from 239 Twitter accounts. With respect to those Tweets, Twitter determined that they propagated the content in order to refute the message and alert other users that the information is false and misleading. And partly as a result of our enforcement decisions, those refuting Tweets generated significantly greater engagement across the platform compared to the Tweets spreading the misinformation—eight times as many impressions, engagement by ten times as many users, and twice as many replies.

b. How does Twitter identify tweets that try to suppress the vote or otherwise interfere with voters' rights? How does Twitter know it has identified all tweets that tried to suppress voter turnout? What does Twitter do once it identifies this kind of tweet?

Twitter's automated spam and abuse systems are designed to capture malicious automated activity on our platform. We focus on a range of aberrant behavioral patterns common across accounts that attempt to propagate malicious content. Designing a system that detects non-automated malicious content is a significantly more challenging task, but it is one we are continually examining and committed to address.

Tweets aimed at suppressing voter turnout generally surface through user reports. This content is reviewed, then promptly removed as illegal interference with voting rights: the content is either restricted as inaccessible pending deletion by the user (*i.e.*, other users are unable to see the content) or the responsible accounts are permanently suspended. In addition, in order to proactively surface additional Tweets with a given text-to-vote meme, Twitter utilizes technology for identifying instances where the same image appears across multiple Tweets. Content identified through this process is then subject to manual review.

c. The examples Twitter gave to the Committee are all aimed at suppressing the Clinton vote. Are there others aimed at Trump voters?

We are not aware of any similar incidents of voter suppression efforts that appeared to specifically target Trump voters, such as through the use of pro-Trump hashtags or imagery featuring Trump.

d. Did the people who saw these tweets have things in common like social or political interests or location?

Because the voter suppression Tweets were organic Tweets, not Promoted Tweets, they could not be targeted to any particular user.

e. Provide all tweets that were posted from September 1, 2016, through November 8, 2017, that contained false voting information or otherwise attempted to suppress or interfere with the exercise of voting rights. Please also provide the following information for these tweets:

i. i. account that posted the tweet; date and time of tweet;

ii. ii. the number of views, comments, retweets, and likes;

iii. iii. any action taken by Twitter in response to the tweet; and

iv. iv. the date and time of any action taken by Twitter in response to the tweet.

Twitter will provide to the Committee information concerning the accounts against which Twitter took action in connection with this content, including the text of the Tweets at issue. If the Committee should need additional information concerning these accounts, we respectfully ask that the Committee follow up with us with the additional information necessary to your investigation.

f. Have you identified the actors behind the various tweets? Are the accounts that sent these tweets still publicly accessible? Did you take down tweets that were shared by other users?

Twitter identified and suspended 106 accounts responsible for 734 text-to-vote Tweets, and we rendered inaccessible pending deletion an additional 918 Tweets from 529 users. As described above, we left available Tweets that refuted the information and thus contributed to correcting any misinformation the malicious Tweets created on the platform. Twitter also Tweeted information to users indicating these text-to-vote Tweets were false. That corrective Tweet received 11.5 million impressions.

g. Once tweets that contain false voting information or otherwise attempt to suppress or interfere with the exercise of voting rights like these are identified, will Twitter shut down or freeze the accounts of the users who post voter suppression tweets?

Depending on the number of violations for any given account disseminating voter suppression Tweets, Twitter will either restrict access to the Tweet or suspend the account. During the period leading up to the 2016 election, for example, Twitter labeled and restricted access to the vote-to-text Tweets pursuant to the Twitter User Agreement, which contains the Twitter Terms of Service, Twitter Privacy Policy, and Twitter Rules. According to the unlawful use provision of the Twitter Rules, users are prohibited from using Twitter's "service for any unlawful purpose or in furtherance of illegal activities" and "[i]nternational users agree to comply with all local laws regarding online conduct and acceptable content." Twitter User Agreement—Twitter Rules, *available at* https://twitter.com/en/tos. Because the Tweets in question appeared to mislead users into believing that they could vote online or vote by text, Twitter viewed the Tweets as an unlawful interference with the voting process.

Twitter labeled as "restricted pending deletion" a total of 918 such Tweets from 529 Twitter accounts, which rendered the Tweets inaccessible and disabled the accounts' ability to use the platform until those Tweets were deleted. In connection with this activity, Twitter also suspended 106 of those accounts, a majority of which were found to be in violation of the Twitter Rules prohibiting spam, including posting duplicate content over multiple accounts or multiple duplicate updates on one account. In a few instances, however, Twitter suspended accounts of users who shared the voting-related content and had previous, but otherwise unrelated, violations of the Twitter Rules against abusive behavior.

Data Deletion During the 2016 Campaign

18. It has been reported that during the presidential campaign, Twitter updated its policies to require permanent deletion of certain data. Specifically, Twitter updated its privacy policy and user agreements reminding firms that when anyone deleted or revised tweets, or closed accounts, the commercial firms accessing Twitter data would be required to destroy that deleted data as well. Firms failing to follow Twitter's directive risked being cut off from Twitter, and that happened in at least one case. (Politico, "Twitter urged firms to delete data during 2016 campaign," October 27, 2017.) Researchers have complained that Twitter's deletion policy was undermining efforts to analyze the extent to which Russians were spreading misinformation using the platform. (*Id.*)

a. Why did Twitter make these updates in the middle of the presidential campaign?

Reports that we updated our policies specifically to require deletion of certain campaign-related data reflect a misunderstanding of our API policies. Through our API, we give developers and other third parties access to subsets of public Twitter content. Access to this publicly available data through our API is conditioned on acceptance of our policies, including the requirement that developers not use the API to undertake activities with respect to content that users have removed from the platform. Examples of situations this policy is designed to address include a parent deciding to remove pictures of their children if they have safety concerns or a college student removing Tweets as they prepare to apply for jobs.

This is a long-standing Twitter policy and was not a new policy implemented during the election. Twitter's recent API policy updates were entirely unrelated to the presidential election, investigations of election interference, or any specific content.

We do update our policy from time to time to clarify the language or otherwise improve how we communicate with our developers.

b. Does Twitter have any record of the data that was deleted, particularly as to posts from Russian trolls during the presidential election?

Twitter has been able to identify information about malicious Russian activity on the platform during the U.S. election for its analysis described in Appendix 1 based on a range of internal signals and data. The retrospective review included signals from malicious users who were suspended from the platform and could not delete their data, as well as signals from malicious users who did attempt deletion while active on the platform. As noted in our testimony, for malicious users utilizing VPNs to access our systems, we would be unable to determine whether those users accessed Twitter's platform from Russia based on IP address, but that is the case regardless of whether or not an account was active on the platform.

c. Instead of fully deleting the data, why didn't Twitter simply hide the content from users but keep it available for bona fide researchers to analyze?

With respect to the policy updates referenced in Question 18(a), researchers and other third parties must adhere to Twitter's policies to access to data via our API. Through our API, we give developers and other third parties access to subsets of public Twitter content. Our policies are designed to apply equally and consistently across developers. These policies limit our ability to provide third party researchers all of the internal signals and data available to us.

d. If Twitter has deleted data from its system, how can we be sure that Twitter has identified all Russia-connected malicious users?

As noted above, Twitter has been able to identify information about malicious Russian activity on the platform during the U.S. election for its analysis described in Appendix 1 based on a range of internal signals and data. The retrospective review included signals from malicious users who were suspended from the platform and could not delete their data, as well as signals from malicious users who did attempt deletion while active on the platform. As noted in our testimony, for malicious users utilizing VPNs to access our systems, we would be unable to determine whether those users accessed Twitter's platform from Russia based on IP address, but that is the case regardless of whether or not an account was active on the platform.

e. How do we know that copycat trolls like the Internet Research Agency have not simply covered their tracks by deleting their original malicious posts?

As described above in response to Question 18(d), we have been able to conduct an extensive analysis of activity on the Twitter platform. In addition, through analysis of our internal signals and information provided by third-parties, we have been able to identify accounts associated with the IRA that were active during the relevant time period and to analyze the content they generated and propagated through the platform.

Russian State-Sponsored Media

19. What steps did your company take to evaluate how its platform is being exploited by Russian organizations before and after the Intelligence Community Assessment was released in January 2017?

In the period preceding the 2016 election we observed new ways in which accounts were abusing automation to propagate misinformation on our platform. Among other things, we noticed accounts that Tweeted false information about voting in the 2016 election, automated accounts that Tweeted about trending hashtags, and users who abused their access to the platform we provide developers. At the time, we understood these to be isolated incidents, rather than manifestations of a larger, coordinated effort at misinformation on our platform. And, as a result, we targeted such activity on an incident-by-incident level rather than as examples of a platform-wide problem.

Once we understood the systemic nature of the problem in the aftermath of the election, we launched a dedicated initiative to research and combat that new threat. Immediately

following the election, we adopted a forward-looking approach and prioritized bolstering existing systems, improving detection capabilities, and implementing new safeguards in advance of additional elections. This has resulted in significant enhancements to our detection tools. Platform-wide, 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.

We also conducted an intensive retrospective review of how Russian actors utilized our platform to interfere with the 2016 election. The methodology and latest results of that review are described in Appendix 1.

We are committed to continue to bolster our detection tools and mechanisms through information we obtain from third parties, our peers, and from our own retrospective review and analysis.

20. How does your [company] identify state-sponsored propaganda? What steps does your company take once such propaganda is identified?

Twitter faces the same technological challenges in identifying state-sponsored propaganda as it does in determining the source of any other content. Users can mask geographic origin and affiliation by using data centers, Virtual Private Networks, and proxy servers, which makes it very difficult to determine whether an account's content is part of a coordinated foreign propaganda effort. To the extent that we are able to identify any foreign links or signals associated with an account and determine that the account has been functioning as state-sponsored propaganda, Twitter reserves the right to take action on the account. We also look to governments to identify organizations who may appear independent, but are in fact state-sponsored and seeking to exploit our platform. For example, Twitter recently ended its advertising relationship with Russia Today and Sputnik on the basis of their efforts to disrupt the 2016 Presidential election as reported by the Intelligence Community and due to violations of our advertising policies.

21. How does your company treat content from state-sponsored propaganda accounts or suspect accounts in its news feed and elsewhere?

There are numerous state-sponsored accounts on Twitter that openly engage in advocacy and will not be suspended or actioned by Twitter. However, in circumstances where the accounts' content or engagement is damaging to the users and the integrity of the platform or violates our policies—as was the case with RT and Sputnik—Twitter will take actions that it deems appropriate to minimize that damage going forward. For example, advertising will no longer be made available to RT and Sputnik, because many of their ads violated policies adopted following the election and relied on such promoted content in their efforts to influence the U.S. election without disclosing state-sponsorship.

22. Identify how much money your company has made, directly or via third-party intermediaries, through its relationships with RT, Sputnik, and any other Russian state-run media entities, whether by (i) selling these entities' ads, (ii) placing ads on these entities' websites or webpages, or (iii) in any other way. Please provide this

information broken down by year, Russian entity, and company product.

We examined advertisement spend between 2011 and 2017 by a number of accounts linked to RT and Sputnik—regardless of content— as well as seven additional accounts that our review identified as (1) having at least one of the criteria we used to identify potential Russia-linked accounts and (2) having promoted election-related content that violated our ads policies, including our policies related to inflammatory content or ad quality.

We deemed an account to be Russian-linked if any of the following criteria were present: (1) the account had a Russian email address, mobile number, credit card, or login IP; (2) Russia was the declared country on the account; or (3) Russian language or Cyrillic characters appeared in the account information or name. And we treated as election-related any promoted Tweets that referred to any candidates (directly or indirectly), political parties, notable debate topics, the 2016 election generally, events associated with the election, or any political figures in the U.S.

RT accounts—@RT_com, @RT_America, @ActualidadRT, @RT_Deutsch, @RT_russian, @RTarabic, @RTenfrancais, @RTUKnews, and @RT_1917—spent a combined total of $34,264 on ad campaigns in 2011, $165,316 in 2012, $60,009 in 2014, $388,469 in 2015, $1,083,065 in 2016, and $134,888 in 2017. RT did not run any ad campaigns in 2013. Campaigns served to U.S. users represented only $17,764 of the combined RT spend in 2014, $459 in 2015, $274,100 in 2016, and $7,130 in 2017. No portion of the RT accounts' global advertising spend in 2011, 2012 or 2013 was for advertisements served in the U.S. Overall, RT's global ad spend on Twitter between 2011 and 2017 was $1,866,012, $299,461 of which represented spend on U.S.-based campaigns.

In 2016 and 2017, certain of the RT accounts also purchased advertisements for the Twitter Audience Platform ("TAP"), a service that enables Twitter advertisers to run ads outside the platform and on mobile applications. RT's TAP spend was $71,259 in 2016 and $20 in 2017. Only $453 of the RT's TAP spend (all from 2016) was for ads targeting U.S. users.

Sputnik accounts—@SputnikInt, @sputnikbelarus, @de_sputnik, @sputnik_fr, and @sputnik_ir—spent a combined $160 on advertisements in 2015 and $6 on advertisements in 2016. Sputnik accounts did not run any ad campaigns from 2011 through 2014, or in 2017. Of those amounts, only $35 (spent in 2015) was for ad campaigns served in the U.S. Unlike RT, Sputnik did not make use of TAP.

We also identified much smaller ad spending—starting in 2015—from seven other accounts that met at least one of our criteria for Russian-linked:

- @replyua spent $47 and $239 on advertising in 2015 and 2016, respectively, none of which represented ads served to users in the U.S. This account also spent $61 on TAP in 2016, but the TAP ads were not served to U.S.-based users.
- @RussiaConnects spent $765 on advertising in 2016, $86 of which was spent on ads served to users in the U.S.
- @RadHumanity spent $60 on advertising in 2016, all of which was spent on ads served to users in the U.S.

- @VijayIngam spent $592 on advertising in 2016, $488 of which was spent on ads served to users in the U.S.
- @TrumpForBoston spent $100 on advertising in 2016, all of which was spent on ads served to users in the U.S.
- @Publius_Philo spent $450 on advertising in 2016 and $84 in 2017, all of which was for ads served to users in the U.S.
- @ericlichmann spent $12 in 2016, $1 of which was spent on ads served to users in the U.S. This account also spent $3 on TAP in 2016, but the TAP ads were not served to U.S.-based users.

Targeting Voters

23. It has been reported that social media companies offered to embed their personnel with the presidential campaigns so that they could make more effective use of your ad buying tools. ("How Facebook, Google and Twitter 'embeds' helped Trump in 2016" Politico, 10/26/17.) For example, your personnel could assist the campaigns in refining their voter targeting to maximize the effectiveness of their ads.

a. What tools did your company offer the campaigns to target voters?

Twitter did not "embed" staff with any political campaign in the 2016 election. We did have staff visit the campaigns to educate staff on how to use the platform and get the most use out of their advertising spend. This is a service we offer to all of our large advertisers; it is not unique to any campaign or to election campaigns.

b. Did your company's employees provide voter profiles to the campaigns to allow for "microtargeting" of prospective voters?

Twitter did not provide that information to any campaign. Twitter personnel that visited campaign locations provided guidance and education on how to use our platform and products, not data and information about our users.

c. Did your company's employees provide input on the content of ads to make them more effective?

Twitter sales personnel assigned to work with the campaigns offered guidance on how to maximize reach and engagement on our platform. This guidance included advising the campaign on Twitter advertising best practices.

24. We know that Russian operatives used Facebook, Twitter, and Google platforms to build deceptive online presences. We also know that Russia-linked ads targeted U.S. users in various ways, including interests and location.

a. Did the Russia-linked advertisers target people in similar ways – by similar interests, locations, etc. – as the Trump campaign?

In connection with our retrospective review of Russian activity on our platform in 2016, we identified nine accounts as being potentially linked to Russia that promoted election-related, English-language content. Of the nine accounts that we identified as being potentially linked to Russia and promoting election-related, English-language content, the most significant use of advertising was by @RT_com and @RT_America. Those two accounts collectively ran 44 different ad campaigns, accounting for nearly all of the relevant advertising we reviewed.

Of all of RT's ad campaigns, 11 were targeted exclusively at English-language speakers, and several others—including all of @RT_America's seven campaigns—used geographic targeting to focus on U.S. users. Though many of the campaigns did not include specialized, non-geographic targeting, a subset of the @RT_com campaigns targeted followers of other RT accounts or followers of other leading news organizations based in the U.S. and other countries. A small number of short "quick promote" campaigns used keyword targeting to attempt to reach audiences searching for particular words or phrases.

The remaining seven accounts, which collectively ran approximately 50 ad campaigns, used a broad range of targeting strategies. We did not identify a trend across the targeting criteria used by those accounts. The accounts sporadically used English-language targeting and location targeting at the country level (including the U.S., Canada, the UK, France, and Ukraine). A handful of campaigns also sought to reach followers of certain accounts.

White Nationalism and Online Cyberhate

25. During the last Presidential election (from August 2015-July 2016), the Anti-Defamation League found 2.6 million tweets that had anti-Semitic language, with nearly 20,000 tweets directed at 50,000 U. S. journalists. One Jewish reporter received threats over twitter, including a photoshopped picture of her face on a corpse in a concentration camp. (USA Today, "Massive Rise in Hate Speech on Twitter during Presidential Election," 10/21/16.) The photo included a message saying, "Don't mess with our boy Trump, or you will be first in line for the camp." This type of cyberhate has targeted other minority communities as well, including Muslim and immigrant communities.

a. What is your company doing to take down these types of messages and advertisements?

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.

In December 2017, we began enforcing new 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. In addition, we now prohibit the use of hateful imagery in avatars or profile headers. Accounts containing such content are hidden and not viewable by other users.

Twitter has worked with various organizations, including the Anti-Defamation League, Dangerous Speech Project, Muslim Advocates, and academics whose scholarship focuses on those and related issues to better understand and address the problem of online hate speech while simultaneously observing the principles of free expression. That collaboration focuses on developing approaches to reduce and counter the proliferation of hate speech online, finding technical solutions to this problem, enabling the exchange of ideas, and enhancing our approach to identifying and combating online hate.

In light of our deeper appreciation for and understanding of those issues as a result of our long-standing partnership with those organizations and other members of our Trust & Safety Council, we implemented three measures. First, in order to provide our users with increased control over their Twitter experience, we expanded our "mute" feature to enable users to mute not only accounts, but also keywords, phrases, as well as entire conversations. Second, we expanded our hateful conduct policy, which now prohibits specific conduct that targets people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. And we do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories. Third, to ensure proper enforcement and implementation of our hateful conduct and related policies, we provided additional training to all support teams, which included sessions on cultural and historical contextualization of hateful conduct. We also implemented an ongoing refresher program to ensure continued familiarity with the policies and any updates to those policies. Finally, we improved our internal tools and systems to more effectively address reports of hateful conduct on our platform.

QUESTIONS FOR THE RECORD—SENATOR WHITEHOUSE

1. a. Please identify the specific ways in which Twitter has improved since this time last year with respect to identifying, preventing, and addressing the use of its platform for purposes of foreign interference in our elections (including by individuals or entities spreading disinformation.)

We are committed to addressing the spread of misinformation on our platform—and to prevent future attempts to interfere with U.S. elections—but we recognize that spam and malicious automation are not limited to political content and can undermine the positive user experience we seek to offer regardless of content. Twitter's approach to addressing the spread of malicious automation and inauthentic accounts on our platform is to focus on identifying problematic behavior and abuse, not primarily on the content that such accounts attempt to disseminate. This is not to say that the content is not important—or that content has no place in our analysis—but we recognize that those who are seeking to influence a wide audience must find ways to amplify their messages across Twitter. As with spam and terrorist content, 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.

Since the 2016 election, 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. 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. We are confident that the combination of our dedicated teams, our detection tools, and other technological advancements at our disposal will prove essential in addressing malicious human-coordinated activity as well.

b. How does Twitter define success with respect to combating use of its platform for purposes of foreign interference in our democracy, and what goal posts will the company use to make progress toward success?

We believe that our significantly increased detection rates of malicious automation and malicious third-party applications, as described above, are meaningful improvements in combatting these risks. We are conscious, however, that this work is ongoing and that there is room for improvement. Preventing future efforts to interfere in our democratic process is a dynamic challenge, as the tactics and strategies used by our adversaries will shift as we adapt to and anticipate their actions. Although it is difficult to set affirmative goal posts to measure success, we are committed to our goal of remaining a free and open platform for the exchange of ideas while preventing action taken by bad actors to influence the democratic process, and we will continue to invest resources accordingly.

c. Do shell corporations impede your company's progress in achieving any of the goals enumerated in (b)? If so, how? Would incorporation transparency laws (e.g., laws requiring the disclosure of beneficial ownership information at the time of incorporation) enhance your ability to overcome those impediments?

Twitter recognizes that, even with our newly announced Ads Transparency Center, which will considerably improve users' visibility into those who run and fund the ads visible to them, there are limits to the due diligence that Twitter can conduct on users and accounts that purchase ads. We are working on improved approaches to getting to know our clients and gaining information about who is behind the entities that are purchasing ads on our platform. And we are committed to continually reexamine ways in which we can further improve transparency and provide additional information to the Twitter community.

2. What is your understanding about how creators of misinformation use Twitter hashtags and handles to promote their messages?

Hashtags can be used by both legitimate and malicious actors to create "trends" that draw attention to a topic. Trends are words, phrases, or hashtags that may relate to an event or other topic (*e.g.*, #CommitteeQFRs). Twitter detects trends through an advanced algorithm that picks up on topics about which activity is growing quickly and thus showing a new or heightened interest among our users. Trends thus do not measure the aggregate popularity of a topic, but rather the velocity of Tweets with related content. The trends that a user sees may depend on a number of factors, including their location and their interests. If a user clicks on a trend, the user can see Tweets that contain that hashtag.

In connection with our retrospective review, and in an effort to measure Russian use of this tactic, we examined several election-related hashtags that trended during the period leading up to the 2016 election. For example, we analyzed data concerning Tweets promoting the #PodestaEmails hashtag, which originated with Wikileaks' publication of thousands of emails from the Clinton campaign chairman John Podesta's Gmail account. We found that slightly under 5% of Tweets containing #PodestaEmails came from accounts with potential links to Russia, and that those Tweets accounted for less than 20% of impressions generated within the first seven days of posting. The core of the hashtag was propagated by Wikileaks, whose account sent out a series of 118 original Tweets containing variants on the hashtag #PodestaEmails referencing the daily installments of the emails released on the Wikileaks website. In the two months preceding the election, around 64,000 users posted approximately 484,000 unique Tweets containing variations of the #PodestaEmails hashtag. Our automated spam detection systems identified in real time approximately 25% of those Tweets, hiding them from searches. Based on information we had available at the time we submitted our hearing testimony, we know that approximately 75% of impressions on the trending topic within the first seven days were views by U.S.-based users. A significant portion of these impressions, however, are attributable to a handful of high-profile accounts, primarily @Wikileaks. At least one heavily-Retweeted Tweet came from another potentially Russia-linked account that showed signs of automation.

We also analyzed #DNCLeak, which concerned the disclosure of leaked emails from the Democratic National Committee, approximately 26,500 users posted around 154,800 unique Tweets with that hashtag in the relevant period. Of those Tweets, roughly 3% were from potentially Russian-linked accounts. Our automated systems at the time detected, labeled, and hid just under half (47%) of all the original Tweets with #DNCLeak. Of the total Tweets with the hashtag, 0.95% were hidden and also originated from accounts that met at least one of the criteria for a Russian-linked account. Those Tweets received 0.35% of overall Tweet impressions within the first seven days after posting. We learned that a small number of Tweets from several large accounts were principally responsible for the propagation of this trend. In fact, several of the most-viewed Tweets with #DNCLeak were posted by @Wikileaks, an account with millions of followers.

In addition to hashtags, accounts may also engage with other accounts by mentioning those accounts in their Tweets. A mention is a Tweet that contains another person's @username anywhere in the body of the Tweet. Mentions increase a user's potential of reaching Twitter users who may not be following the Tweeting account. While this practice was not common and generally was limited to accounts associated with the Internet Research Agency, we saw the use of mentions by several Russia-linked accounts to attempt to communicate with journalists, activists, and other users on Twitter.

Mentions can also appear in replies to existing Tweets. For example, on December 2, 2016, we learned of @PatrioticPepe, an account that automatically replied to all Tweets from @realDonaldTrump with spam content (and mentioned @realDonaldTrump in its replies). On the same day that we identified @PatrioticPepe, we suspended the API credentials associated with that user for violation of our Twitter Rules and our Automation Rules.

3. Many Americans heard about the fake TEN_GOP Twitter account, which called itself the unofficial account of the Tennessee Republican party and enjoyed 100,000 followers before being shut down. For months, it sent out a stream of fake news such as a tweet falsely stating that there was voter fraud in Florida. These false stories got plenty of amplification, including in the form of retweets by Kellyanne Conway and Donald Trump Jr. Donald Trump himself thanked the account for its support during the election.

a. How did this happen, and how do we know similar accounts aren't spreading fake news and being amplified right now?

We have rightfully received criticism for not promptly suspending the @TEN_GOP when we initially received the report. We recognize that we should have acted sooner than we did and we are committed to taking swift action in the future response to similar reports and notifications.

Twitter Rules prohibit impersonation accounts. In response to reports—from either the user who is being impersonated or their authorized representatives—Twitter takes action against accounts that deceptively impersonate another user or account. Users and non-users alike can report impersonation accounts through a dedicated form in our Help Center or directly from the impersonated account's profile on the platform.

We are determined to expedite the suspension process for accounts deemed to be impersonating other users. Once we receive a report of potential user impersonation, we investigate the reported accounts to determine if the accounts are in violation of the Twitter Rules, which prohibit such profiles. Accounts determined to be in violation of our impersonation policy, or those not in compliance with our parody, commentary, and fan account policy, are either suspended or asked to update their profile so they no longer violate our policies.

With respect to the proliferation of fake news and general misinformation, we share the Committee's concern, and we are taking steps to prevent this activity in the future. Although we believe that Twitter's open and real-time environment is a powerful antidote to the spread of deceptive information, we recognize that we cannot rely on user activity alone when faced with organized misinformation campaigns.

For that reason, we have launched new tools to combat malicious disinformation moving forward, and we have reached out to engage with journalistic organizations with regard to the issue of fake news. We are creating a dedicated media literacy program to demonstrate how Twitter can be an effective tool in media literacy education.

b. Does Twitter differentiate between fake news by a user who does not claim to be anyone other than himself or herself and fake news propagated by an account that fraudulently purports to speak for an entity? If so, how?

Impersonation is a violation of the Twitter Rules. While Twitter allows parody, commentary, or fan accounts that clearly identify themselves as such, any account that portrays another in a confusing or deceptive manner may be permanently suspended pursuant to the Twitter impersonation policy.

To the extent that content posted on our platform violates the Twitter Terms of Service or the Twitter Rules, we will take action against that account regardless whether the content originates from a user who does not claim to be someone else.

4. According to a study conducted by the Universities of Southern California and Indiana and published in May, approximately 15 percent of Twitter users are bots.

a. Is this figure accurate, to the best of your knowledge? If not, what is your best estimate as to the accurate figure?

Based on a review of a representative sample of accounts, we estimate that false or spam accounts represent less than 5% of our MAUs. Our estimates are lower than those reported by outside researchers because those researchers do not have access to critical internal information necessary to make an accurate determination of the scale of spam, fake accounts or automated bots on Twitter. As a result, reports from third-party researchers often overestimate the true volume of such accounts on our platform—sometimes by large orders of magnitude.

While our detection tools for false or spam accounts rely on a number of inputs and variables and do not operate with 100% precision, they are informed by information not available outside of Twitter. Our internal researchers have access to and can analyze a number of different signals including, among other things, email addresses, phone numbers, login

history, and other non-public account and activity characteristics that enable us to conduct a more thorough review and reach more accurate conclusions as to whether the account in question is fake or spam. We keep such information confidential and do not make it available to researchers in order to protect the privacy of our users.

Because third-party researchers do not have access to internal signals that Twitter can access, their bot and spam detection methodologies must be based on public information and often rely on human judgment, rather than on internal signals available to us. One common model for determining whether an account is fake or automated is the "Botometer model," which compares publicly available account features, such as Tweet count, follower count, and use of language, to the characteristics exhibited by purportedly "known" bots. The initial evaluation of whether an account is or is not a bot, however, relies on an individual assessment and is, therefore, inherently imprecise.

There are also studies that use the limited public Tweet data that we offer researchers through an application programming interface ("API"). The studies that rely on information from the Twitter API to identify automated accounts similarly overestimate both the number and impact of these accounts because our internal detection tools and filtering techniques are not available to third parties. Those tools enable us to remove from the platform malicious automated accounts (and content generated by such accounts), but the accounts may nevertheless appear in the data stream that researchers access through our API, thus inaccurately reflecting the traffic on Twitter.

In the study conducted by the University of Southern California and Indiana University, the 15% estimate was based on a prediction of whether a user may or may not be an automated account and was derived from human judgments about an account's attributes. The authors of the study acknowledge that detecting automated accounts "is a hard task. Many criteria are used in determining whether an account is controlled by a human or a bot, and even a trained eye gets it wrong sometimes." See https://botometer.iuni.iu.edu/#!/faq#bot-threshold.

We are committed to continuing to work on refining our spam detection tools and to update the Twitter community and the public periodically about our estimates and analysis of these things on our platform. We regularly receive and welcome input from researchers and Twitter users about ways in which we can optimize our detection and enforcement methods. In addition, as we have announced, we are also committed to donating the $1.9 million we projected to have earned from RT advertising to support external research into the use of Twitter in civic engagement and elections, including the use of malicious automation and misinformation.

b. What steps, if any, is the company taking to remove these bots from its network?

We regularly take action to challenge and remove malicious automated accounts on Twitter. Once our systems detect an account as generating spam, we can take action against that account at either the account level or the Tweet level. Depending on the mode of detection, we have varying levels of confidence about our determination that an account is violating our rules. We have a range of options for enforcement; generally, the higher our confidence that an account is violating our rules, the stricter our enforcement action will be, with immediate suspension as

the harshest penalty. If we are not sufficiently confident to suspend an account on the basis of a given detection technique, we may challenge the account to verify a phone number or to otherwise prove human operation, or we may flag the account for review by Twitter personnel. Until the user completes the challenge, or until the review by our teams has been completed, the account is temporarily suspended; the user cannot produce new content (or perform actions like Retweets or likes), and the account's contents are hidden from other Twitter users.

We also have the capability to detect suspicious activity at the Tweet level and, if certain criteria are met, to internally tag that Tweet as spam or otherwise suspicious. Tweets that have been assigned those designations are hidden from searches, do not count toward generating trends, and generally will not appear in feeds unless a user follows that account. Typically, users whose Tweets are designated as spam are also put through the challenges described above and are suspended if they cannot pass.

For safety-related Terms of Service ("TOS") violations, we have a number of enforcement options. For example, we can stop the spread of malicious content by categorizing a Tweet as "restricted pending deletion," which requires a user to delete the Tweet before the user is permitted to continue using the account and engaging with the platform. So long as the Tweet is restricted—and until the user deletes the Tweet—the Tweet remains inaccessible to and hidden from all Twitter users. The user is blocked from Tweeting further unless and until he or she deletes the restricted Tweet. This mechanism is a common enforcement approach to addressing less severe content violations of our TOS outside the spam context; it also promotes education among our users. More serious violations, such as posting child sexual exploitation or promoting terrorism, result in immediate suspension and may prompt interaction with law enforcement.

c. Assuming that bots count toward your company's "daily active user" (DAU) number and that a high DAU is beneficial to the company in terms of advertising revenue, is there a conflict when it comes to Twitter's incentives to decrease or remove bots?

We estimate the number of false or spam accounts as a percentage of our Monthly Active Users ("MAU"). Such accounts do not count toward MAU measures reported in our SEC filings. In addition, as we discussed in our testimony, malicious automation results in a bad user experience and undermines the confidence of our users and advertisers in the Twitter platform, which could potentially hinder our ability to grow MAU and advertising revenue. We are incentivized to identify and remove accounts responsible for such behavior as well as those engaged in other forms of abuse.

5. Do Twitter's terms of service require that profiles be linked to real names or other unique identifiers to ensure that they are associated with human beings? If not, why not?

We do not require users to include their real names or other unique identifiers when opening an account. We believe that allowing users to create accounts under any name or identity they choose is essential to promoting and facilitating an open and safe platform that

supports free expression. Anonymous and pseudonymous accounts provide critical protections for political dissidents, embedded journalists, and human rights activists.

Anonymity allows a college student who disagrees with the administration, and who might otherwise be reluctant to say anything, to speak against the university's policies; it enables Christians in China to speak about their persecution and oppression; and it empowers individuals to stand up to powerful people on our platform, including government officials, and encourages the expression of minority views.

Twitter recognizes that anonymity can be abused. And we strive to strike the right balance to prevent abuse or harassment. Users on our platform are not completely anonymous. For example, Twitter captures some essential user data, such as email addresses, phone numbers and/or IP addresses. That information is often critical to law enforcement efforts, and Twitter maintains a dedicated 24/7 team to respond to law enforcement requests for such information.

6. In materials Twitter produced to Committee staff prior to the hearing, we saw examples of illegal voter suppression that you say Twitter identified and took down.

a. Can you tell us how many times these illegal messages were retweeted and how many users may have seen them?

Twitter identified a total of 918 vote-by-text Tweets that originated from 529 accounts. Our review indicates that those Tweets were viewed 222,111 times (an average of 242 views per Tweet (also known as "impressions")), were Retweeted by 801 users, and received 318 replies.

b. You testified that retweets highlighting the falsity and illegality of the suppressive messages far outnumbered the number of retweets intended to amplify those messages. Do you have any data to support this claim? How does Twitter identify a user's intent in retweeting a message? Were tweets that shared the illegal message while warning that the message was false also deleted from the platform?

We identified 286 Tweets that shared the illegal messages with commentary refuting the misleading messages. The Tweets correcting the misinformation were viewed 1,634,063 times and Retweeted by 11,620 users.

To determine the user's intent, we conducted a human review of the content of these users' Tweets, in part to assess whether some Retweets were alerting other users that the information was false. We aimed to take action on only Tweets that violated the rules by interfering with voting rights, not Tweets that aided in minimizing the harmful impact of misinformation on the platform.

c. Did Twitter take any proactive steps to clarify to users that any messages urging people to vote by text (or any other illegal voter-suppression messages) should be discounted as false?

On November 6, 2016, Twitter's Government and Elections account, @TwitterGov, proactively Tweeted: "#ElectionDay is almost here! Remember: you cannot vote via text or

Tweet. Direct Message @Gov to find your polling place and ballot info." In addition to Tweeting correct information, when we received reports of "text-to-vote" Tweets, we used proprietary tools to search for linked accounts that also violated our rules. As noted above, after careful review, we took action on over one hundred accounts and hundreds of Tweets.

QUESTIONS FOR THE RECORD—SENATOR COONS

1. **On January 6, 2017, the U.S. Intelligence Community released a public report that concluded, "Vladimir Putin ordered an influence campaign in 2016 aimed at the US presidential election." The documents provided to the Committee confirm the intelligence community's conclusion and highlight the need for online platforms to work with the government to prevent threats going forward.**

    a. **Who is your contact person at the Department of Justice and/or the FBI as you work to counter these ongoing threats?**

    b. **Have the DOJ or FBI made any recommendations to you for preventing interference going forward?**

We maintain an ongoing working relationship with a number of departments and sections of the Department of Justice and the FBI. We frequently meet to discuss law enforcement needs and concerns and to continually improve our response to lawful process from all law enforcement agencies. During the course of our relationship, we have discussed a number of threats to U.S. interests, as well as to Twitter services. While we have not received any specific advice on avoiding election interference, we have seen value in our ongoing exchanges with those agencies, and we remain committed to a continued dialogue.

2. **Do you believe that computer algorithms and machine learning are sufficient to catch foreign political ads, fake accounts, and false information?**

    a. **What new technologies or capabilities will you introduce to prevent these abuses?**

    b. **At the hearing, you testified that Twitter has hundreds of people in its trust and safety team and user services team. What do these employees do to improve safety and security?**

    c. **How many employees will be dedicated to preventing foreign political ads from being published on the platform?**

Twitter dedicates significant resources to addressing malicious automation, bots, and other coordinated activities. We believe we have the right resources and strategies in place. We dedicated nearly the entire engineering, product, and design teams to look at these issues at the beginning of 2017, and we regularly reexamine staffing and resources and adjust as needed.

For example, 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. We are confident that the combination of our dedicated teams, our detection tools, and other technological advancements at our disposal will prove essential in addressing malicious human-coordinated activity as well.

We also recognize that, at this time, computer algorithms alone may not be sufficient to address the problem. Accordingly, those tools are complemented by manual review teams, collaboration and information sharing with industry peers and participants, reliance on data and intelligence from third-party security vendors, and partnerships with other companies and civil society.

Twitter understands that, to succeed, we must combine information, knowledge, and effort with industry partners, civil society, academic institutions, and government. We do not compete against other companies on our ability to detect and label malicious content on our platform; instead, we recognize that we will all be stronger if we view this as a shared threat. We are committed to a continued collaborative approach and believe it will prove successful going forward.

3. Although Twitter catches some bot accounts, many others go undetected and can quickly disseminate false news. What improvements are you making to counter increasingly sophisticated bots?

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.

Since the 2016 election, 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. 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. We are confident that the combination of our dedicated teams, our detection tools, and other technological advancements at our disposal will prove essential in addressing malicious human-coordinated activity as well.

4. Russian operatives were able to increase their influence by hacking or purchasing online accounts that were originally authentic but no longer maintained by their owners. In fact, buying unmaintained accounts has become a cottage industry. What steps are you taking to prevent unmaintained accounts from falling into the hands of inauthentic users?

Twitter is continually investing in new techniques for identifying potentially compromised or hacked accounts as quickly as possible. We use signals like suspicious login activity and location to detect when an account may have been accessed by a malicious actor, and we regularly use information about large-scale breaches of personal information across the web (such as password leaks from other websites) to help us proactively secure our users' accounts.

5. As we saw with the Comet Pizza incident, where a man brought a gun into a D.C. pizza restaurant based on false reports that criminal activity was occurring there, fake news can stoke hatred and violence. What is Twitter doing to prevent the proliferation of fake news across the site?

Our detection systems combat malicious automation and spam content, which can significantly aid the proliferation of fake news. Additionally, we are actively engaging with journalistic organizations on the issue of misinformation. Enhancing media literacy is critical to ensuring that voters can discern which sources of information have integrity and which may be suspect. We anticipate that our commitment to improving media literacy will make Twitter a more effective tool in media literacy education. Moreover, we engage in collaborations and trainings with NGOs, such as the Committee to Protect Journalists and Reporters without Borders, in order to ensure that journalists and journalistic organizations are familiar with how to utilize Twitter effectively to convey timely and accurate information.

6. Does Twitter support the Honest Ads Act? If you do not support this bill or are unable to commit to a position, please explain why.

Twitter supports the goals of the Honest Ads Act. Through our own initiative, we have announced voluntary, industry-leading steps to improve transparency and accountability in our ads platform that strongly aligns with the goals and standards in the Act. In fact, in some cases, we expect our new transparency requirements will go further than the draft legislation—for example, by requiring transparency measures for a broad range of advertisers.

We do have suggestions for potential improvements of the bill. First, we want to be sure that the proposed requirements, including in-ad disclosure language, are sufficiently flexible to account for character-constrained platforms like Twitter. Second, we hope that legislation on this topic would clarify that, while the duty to collect and display disclosure information lies with the platforms, the duty to provide accurate information lies with the advertisers.

7. Can you assure us that political ads will include permanently displayed disclosure notifications like in print or television ads? If you cannot, please explain why.

Yes. Twitter is updating its policies to ensure that political ads will include permanently displayed disclosure notifications. We will require that electioneering advertisers identify their campaigns as such, and will add a visual political ad indicator to electioneering ads.

8. What reforms will Twitter enact to address issue ads that do not mention political candidates by name?

Twitter's recent updates focus on electioneering ads, which refer to a clearly identified candidate or party associated with that candidate for any elected office. Although there is no clear industry definition for issue-based ads, we are committed to work with our peer companies, other industry leaders, policy makers, and ad partners to define them so that we can integrate them into the new approach. Those advertisements will also still be subject to increased disclosures in Twitter's upcoming Transparency Center. Users are also able to report any advertisement, which will prompt us to review and remove inappropriate advertisements.

9. Foreign entities will continue to try to use social media to interfere with U.S. elections. Has Twitter identified attempts by foreign entities to interfere with post-2016 elections? Please describe such attempts.

While our improved systems for detecting automation have allowed us more effectively to mitigate the impact of malicious, automated election-related content, we have identified the use of both automated accounts and coordinated campaigns as part of apparent efforts to influence post-2016 elections (*e.g.*, an ultimately unsuccessful attempt to influence hashtag trends during the French election).

QUESTIONS FOR THE RECORD—SENATOR DURBIN

1. On January 6, the U.S. Intelligence Community issued a report on Russian election interference and described what happened last year as the "new normal in Russian influence efforts." The IC Report said "we assess Moscow will apply lessons learned from its campaign aimed at the U.S. presidential election to future influence efforts in the United States and worldwide."

We are less than a year away from Election Day in 2018. The campaign season will be upon us before we know it. We do not have much time to safeguard our nation's social media platforms against Russian disinformation efforts and election propaganda.

a. Will your company be ready before Election Day 2018 to reassure Americans that your platform is not tainted by foreign disinformation or influence efforts?

We have made great strides already in improving the quality of information on Twitter and preventing disinformation campaigns. We are committed to ensuring that the 2018 election is the safest ever on Twitter.

b. Will you be ready before then to ensure that consumers can quickly identify who is truly responsible for election ads or election-related content that they see on your platform?

As outlined in our updated advertisement transparency policies, we are committed to ensuring that our users can identify who is responsible for election ads on Twitter's platform. First, Twitter will ensure that political ads will include permanently displayed disclosure notifications, so that the public can quickly identify that it is an election ad in the first place. Second, our upcoming Transparency Center will offer the public increased visibility into all advertising on the platform, and to provide users with tools to share feedback with us. With respect to electioneering ads and the Transparency Center, we intend to better enable users and outside parties to conduct their own research or evaluation regarding particular ads. Electioneering ads information accessible through the Transparency Center will include, among other things, the identity of the organization funding the campaign, all ads that are currently running or have run on Twitter, campaign spend, and targeting demographics for specific ads or campaigns. We plan to launch the Transparency Center as soon as feasible after rolling out our electioneering policy in the first quarter of 2018, and we are continuing to refine the tools we will make available in conjunction launching the Transparency Center to ensure the best experience for our users.

c. If you cannot provide reassurance that you will be ready before Election Day 2018, what else needs to happen in the next year to provide that reassurance?

Twitter is committed to devoting all necessary resources to keep up with the evolving threats posed by our malicious adversaries. We are working every day to enhance our defenses, and we fully intend to continue that progress up to and through the 2018 election.

2. We've heard a lot about the Russian "troll farm" model best exemplified by the Internet Research Agency in St. Petersburg. It is astonishing that we are seeing these types of businesses sprout up for the purpose of spreading disinformation and sowing division online. Your company has taken steps to remove some accounts and ads created by these troll farms, but I fear that a reactive strategy is not going to be good enough.

a. What additional legislative or administrative actions do you think Congress or federal agencies should pursue against these troll farms to prevent them from spreading lies and discord across the internet?

Twitter is committed to addressing malicious activities that originate from coordinated human activity such as troll farms. We work to identify these organizations and take action where they violate our rules. However, as with most technology-based threats, cooperation and information-sharing is essential to ensure that the challenges are properly met.

Although prompt information about such organizations would be most helpful to Twitter, the federal government has other powers at its disposal to address this issue. We are not well positioned to opine on regulatory or enforcement options for addressing the issue of foreign troll farms, but we would encourage legislative solutions that would increase transparency surrounding this issue for more informed enforcement on our platform.

b. Should there be a special designation or "watch list" set up by the government for troll farms which would carry certain penalties or obligations for companies that fit this designation?

We do not believe that we are in the best position to opine on all policy considerations arising from potential watch lists. But more broadly, Twitter would benefit from information sharing concerning organizations that are state-sponsored propaganda organizations engaged in activity harmful to the United States so that Twitter can take appropriate steps to identify abuse of its platform by associated users.

3.

a. Is it your view that the federal Departments of Justice and Homeland Security are taking the problem of Russian disinformation over social media seriously?

Twitter has no reason to believe that the Department of Justice and the Department of Homeland Security are not taking the problem seriously.

b. Is your company getting support, guidance and collaboration from those two agencies?

c. Who are the point people in those agencies dedicated to working with your company on this challenge?

We maintain an ongoing, working relationship with a number of departments and sections of the Department of Justice and the FBI. We frequently meet to discuss law enforcement needs and concerns, and we strive to continually improve our response to lawful

process from all law enforcement agencies. During the course of our relationship, we have discussed a number of threats to U.S. interests, as well as to Twitter services. While we have not received any specific advice on avoiding election interference, we have seen value in our ongoing exchanges with those agencies, and we remain committed to a continued dialogue.

4. Much of the discussion about combatting extremist content on social media has centered around the global terrorism threat. However, we are also facing a rising threat posed by white supremacist and other domestic extremist groups, who are all too often motivated by bigotry and hate.

An unclassified May 2017 FBI-DHS joint intelligence bulletin found that "white supremacist extremism poses [a] persistent threat of lethal violence," and that white supremacists "were responsible for 49 homicides in 26 attacks from 2000 to 2016 … more than any other domestic extremist movement." And *Politico* reported recently that "suspects accused of extreme right-wing violence have accounted for far more attacks in the U.S. than those linked to foreign Islamic groups like al Qaeda and ISIS, according to multiple independent studies."

What steps is your company taking to address extremist content from white supremacists and other domestic terrorist threats?

Twitter is continually working to make the platform a safe place for our users. For example, we are introducing changes to our Twitter Rules, including how we correspond with those who violate them, and to our rules' enforcement process. We recently unveiled clarifications and updates to rules regarding hateful display names, hateful imagery, violent groups, and content that glorifies violence, which we began enforcing in December 2017. We made the various updates available prior to enforcement in order to provide our users and the general Twitter community with sufficient time to review and understand them.

QUESTIONS FOR THE RECORD—SENATOR LEAHY

1. Columbia Law Professor Tim Wu recently wrote about how Russian and Chinese government "troll armies" on Twitter have perfected "reverse censorship" – essentially drowning disfavored speech in a flood of propaganda or simply irrelevant distraction.[1]

What more can Twitter do to prevent authoritarian regimes from deliberately trying to quash already suppressed voices of dissent on your platform?

Our products are designed to show users the best and most relevant content first, while filtering or removing malicious, abusive, or low-quality Tweets. We continue to invest in improving our spam and malicious automation detection systems in order to identify and prevent the amplification of malicious automated content and attempts to drown out legitimate speech using malicious automation. And we continue to enhance and improve our recommendations and relevance systems.

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.

Since the 2016 election, 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

[1] https://www.nytimes.com/2017/10/27/opinion/twitter-first-amendment.html.

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. 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. We are confident that the combination of our dedicated teams, our detection tools, and other technological advancements at our disposal will prove essential in addressing malicious human-coordinated activity as well.

2. In September 2017, Twitter released a statement about what it is doing to combat automated "bots."[2] Bots were routinely used by Russia during the 2016 election to spread false, divisive, and malicious content. Roughly 37,000 Russian "bots" spread election-related tweets that were viewed 288 million times between September and November last year.

a. Twitter was only able to identify approximately half of the 37,000 Russian bots in real time. Yet already this year Twitter suspended almost 300,000 accounts for promoting terrorism – 95 percent of which were detected automatically.[3] Given this discrepancy, can Twitter be doing more to combat Russian "bots"?

As described in detail in Appendix 1, through our analysis of relevant activity on our platform, we have identified additional 13,512 accounts, for a total of 50,258 automated accounts that we identified as Russian-linked and Tweeting election-related content during the period between September 1, 2016 and November 15, 2016. Those accounts represent a fraction of the average four million accounts globally that we take action on for malicious automated activity each week.

Although we believe that we have vastly improved our detection capabilities since the 2016 election, we are committed to pursuing and implementing additional improvements going forward. In the coming year, we plan to build upon our 2017 improvements by making additional investments in machine-learning capabilities that help us detect and mitigate the effect on users of fake, coordinated, and automated account activity. Our engineers and product specialists continue this work every day, further refining our systems so that we capture and address as much malicious content as possible. We are committed to continuing to invest all necessary resources into making sure that our platform remains safe for our users.

b. Beyond the Twitter's Information Quality Initiative, will Twitter give this Committee <u>quantifiable and verifiable metrics</u> on how it will be cracking down on the deceptive political use of "bots"?

Twitter has developed new techniques for identifying patterns of activity that are consistent with automated use of our platform such as near-instantaneous replies to Tweets, non-random Tweet timing, and coordinated engagement with content. We are currently implementing those detections tools across our platform. 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.

[2] https://blog.twitter.com/official/en_us/topics/company/2017/Update-Russian-Interference-in-2016--Election-Bots-and-Misinformation.html

[3] https://transparency.twitter.com/en/gov-tos-reports.html

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. Those 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.

3. As I stated in my questions at the hearing on October 31, 2017, Russia's "Internet Research Agency" purportedly set up a fake Twitter account posing as the Tennessee Republican party, and sent out a stream of fake claims including allegations of voter fraud.

What more can Twitter do to prevent a user from fraudulently pretending to be someone they're not?

Twitter Rules prohibit impersonation accounts. In response to reports—from either the user who is being impersonated or their authorized representatives—Twitter takes action against accounts that deceptively impersonate another user or account. Users and non-users alike can report impersonation accounts through a dedicated form in our Help Center or directly from the impersonated account's profile on the platform.

Once we receive a report of potential user impersonation, we investigate the reported accounts to determine if the accounts are in violation of the Twitter Rules, which prohibit such profiles. Accounts determined to be in violation of our impersonation policy, or those not in compliance with our parody, commentary, and fan account policy, are either suspended or asked to update their profile so they no longer violate our policies.

QUESTIONS FOR THE RECORD—SENATOR BLUMENTHAL

A. Data Production to the Judiciary Committee

1. Has Twitter identified Russian advertisements or accounts that did not originate with the Internet Research Agency? What is the status of your efforts to identify such advertisements?

Using a custom-built machine-learning model, we identified 6,493 advertisement accounts with election-related content that promoted English-language Tweets in 2016. We considered those accounts to be Russian-linked if they exhibited one of several criteria indicating a Russian account, including (1) Russian contact information, billing information or login IP; (2) Russia as the declared country on the account; and (3) Russian language or Cyrillic characters in the account information or username.

Nine accounts that had at least one of the criteria for a Russian-linked account also promoted election-related content Tweets that, based on our manual review, violated existing or recently implemented ads policies, such as those prohibiting inflammatory or low-quality content. Two of those accounts—@RT_COM and @RT_America—together spent $516,900 in advertising in 2016, $234,600 of which was spent on ads served to users in the U.S. The two accounts promoted 1,912 Tweets and generated approximately 192 million impressions across all ad campaigns, with approximately 53.5 million representing impressions generated by U.S.-based users. Twitter recently ended its advertising relationship with RT on the basis of their efforts to disrupt the 2016 Presidential election.

The remaining seven accounts were small and apparently unconnected actors. In 2016, those seven accounts (1) spent a combined total of $2,282 on advertising, including $1,184 spent on ads that were served to users in the U.S.; (2) ran 404 promoted Tweets; and (3) generated 2.29 million impressions across all ad campaigns (approximately 222,000 of which were impressions generated by U.S. users). We have since off-boarded those advertisers.

2. When can you provide these advertisements to this Committee?

Twitter will provide to the Committee the advertisements from the nine accounts described in response to question 1.

B. Role of Social Media Consultants/Social Media Management Companies

3. Have you done any analysis to determine the degree to which Russia relied on social media consultants/management companies to purchase ads designed to influence the election?

Although we analyzed the advertisements purchased by Russian government-affiliated entities such as RT from the period leading up to the 2016 election, Twitter does not have any insight into those accounts' use of social media consultants in their advertising content creation or targeting.

4. To what degree will your new transparency policies help the public identify ads purchased by foreign governments if these ads are purchased through social media consultants/management companies?

Twitter is launching an industry-leading Transparency Center that will offer the public visibility into who is advertising on Twitter and how those ads are targeted. Users will have tools to share feedback with us. While it is possible that foreign governments may attempt to purchase ads through consultants/management companies, we intend that Twitter's upcoming Transparency Center will provide identifying information and transparency into other advertising activities on Twitter. That information will better enable users and outside parties to conduct their own research or evaluation regarding particular ads.

5. Are you working with social media consultants/management companies to ensure that they cannot be used to shield political ads from transparency efforts?

Twitter is not working with any social media consultants on this effort, as we do not believe that it is necessary. Twitter's new policies regarding political advertising require disclosure of the person or entity paying for the advertisement. A political advertiser would be unable to rely on a consulting company as a shell "purchaser" of the advertisement, but instead must designate who actually has purchased the ad. This is analogous to political ads on television, which are often "purchased" by media consulting firms, but must nonetheless disclose the person or entity paying for the advertisement

C. Responding to Search Engine Optimization/Search Engine Manipulation

6. It is my understanding that you have systems for detecting attempts to manipulate search results. Are you using these detection systems to identify manipulation originating in Russia? If so, what have you been able to identify?

We have identified instances where users attempted to influence trends through using certain hashtags. Trends are words, phrases, or hashtags that may relate to an event or other topic (*e.g.*, #CommitteeQFRs). Twitter detects trends through an advanced algorithm that picks up on topics about which activity is growing quickly and thus showing a new or heightened interest among our users. Thus, trends do not measure the aggregate popularity of a topic, but rather the velocity of Tweets with related content. The trends that a user sees may depend on a number of factors, including their location and their interests. If a user clicks on a trend, the user can see Tweets that contain that hashtag.

Two examples of our systems in action dealt with election-related hashtags—#PodestaEmails and #DNCLeak—for which our automated systems detected, labeled, and hid a portion of Tweets at the time they were created. We found that slightly under 5% of Tweets containing #PodestaEmails came from accounts with potential links to Russia, and that those Tweets accounted for less than 20% of impressions generated within the first seven days of posting. With respect to #DNCLeak, which concerned the disclosure of leaked emails from the Democratic National Committee, approximately 26,500 users posted around 154,800 unique Tweets with that hashtag in the relevant period. Of those Tweets, roughly 3% were from potentially Russian-linked accounts. Our automated systems at the time detected, labeled, and

hid just under half (47%) of all the original Tweets with #DNCLeak. Of the total Tweets with the hashtag, 0.95% were hidden and also originated from accounts that met at least one of the criteria for a Russian-linked account. Those Tweets received 0.35% of overall Tweet impressions within the first seven days after posting. We learned that a small number of Tweets from several large accounts were principally responsible for the propagation of this trend. In fact, several of the most-viewed Tweets with #DNCLeak were posted by @Wikileaks, an account with millions of followers.

7. How are you addressing the challenge of search engine optimization or search engine manipulation in this context? Are you prioritizing this issue?

Twitter's search capabilities function only within Twitter. In other words, searches only yield Twitter content, or content contained within or linked to from Tweets. And Twitter provides users with several controls over the content made available to them through searches. Those include the "safe search" mode, which excludes potentially sensitive content, along with accounts the user chose to mute or has blocked, from the user's search results. Users also have the option to activate our quality filter, which filters lower-quality content from their notifications. Such content includes, for example, duplicate Tweets or content that appears to be automated. In addition, content that Twitter detects as malicious and automated is excluded from search results.

8. Is there a "paper trail" for this sort of manipulation? What other challenges do you face in identifying search engine manipulation?

Although automated content has the potential to alter search results, Twitter has made significant advances in detecting and labeling malicious automated content before accounts that generate such content are able to proliferate it on our platform. And Tweets that we detect and label as malicious and automated—regardless of whether they were generated by automated or human accounts—are internally tagged as spam and are automatically hidden from searches.

D. Transparency for Issue-Based Advertisements

9. How do you intend to bring greater transparency to issue-based advertisements that potentially originate in Russia?

Twitter's upcoming Transparency Center is intended to provide broad transparency into advertising on the Twitter platform, covering disclosures such as (but not limited to) campaign spend and the identity and advertising history of the organization funding the campaign. We are also committed to work with our peer companies, other industry leaders, policy makers, and ad partners to define issue-based ads so that they can be subjected to the same requirements as electioneering ads. Although we are confident that our new measures will improve transparency, we recognize that this is an ongoing effort that requires a long-term commitment in order to succeed.

10. Do you believe your recent transparency policies go far enough, or do you intend to build on them?

We recently announced stricter, industry-leading transparency policies. We recognize, however, that developing and enforcing our transparency policies is an ongoing challenge. We are committed to working with peer companies, policy makers, and industry leaders to build on our current policies on an ongoing basis.

E. Speed of Transparency Efforts

11. I want to impress upon you the importance of dealing with this issue swiftly—otherwise we may be facing the same situation in November 2018 as we did in November 2016. What assurances can you provide this Committee that your company is working as fast as possible to implement new transparency features?

Twitter understands the urgency of these issues in light of the upcoming elections. As we announced several weeks ago, we have already rolled out new policies focused on advertising transparency, and we are working as quickly as possible to begin implementing and enforcing them. We plan to launch the Transparency Center as soon as feasible after rolling out our electioneering policy in the first quarter of 2018, and we are continuing to refine the tools we will make available in conjunction launching the Transparency Center to ensure the best experience for our users.

F. Impact of Disinformation Campaigns

12. Do you have information regarding how many voters were impacted by posts that were part of foreign disinformation campaigns? Do you have data on how these posts may have impacted election results?

Between September 1 and November 15, 2016, Russian-linked accounts generated approximately 2.12 million automated, election-related Tweets, which collectively received approximately 454.7 million impressions generated within the first seven days of posting. Those 2.12 million Tweets received only one-half of a percent (0.49%) of impressions of all election-related Tweets within the first seven days after posting.

At the time, we detected and labeled as automated over half of the Tweets (approximately 1.34 million) from approximately half of the accounts (23,601), representing 0.63% of overall election-related Tweets and 0.20% of election-related Tweet impressions generated within the first seven days of posting.

Unfortunately, there is no way for Twitter to determine how these impressions impacted actual voting decisions or election results. In the aggregate, however, the automated, Russian-linked, election-related Tweets consistently underperformed in terms of impressions relative to their volume.

13. When can you provide this information to this Committee?

The requested information has been provided in response to question 12.