January 15, 2018

Questions for the Record
Senate Select Committee on Intelligence
Hearing on Social Media Influence in the 2016 U.S. Elections
November 29, 2017

Questions for the Record for Mr. Sean Edgett, Acting General Counsel, Twitter.

[From Chairman Burr]

1. What procedures must the Russian government follow to compel the production of customer-created content or personally identifiable information from your company?

Twitter publishes global guidelines for law enforcement personnel that explain our policies and the process for submitting requests for information. *See* https://help.twitter.com/en/rules-and-policies/twitter-law-enforcement-support. Since 2012, Twitter has published bi-annual Transparency Reports, reflecting the number of requests that we have received for user information and content removal on a per-country basis, including requests from Russia. *See* https://transparency.twitter.com/.

Although we have received requests for user information from Russian government entities, the number of requests per reporting period has been relatively small compared to requests from other jurisdictions. As the Transparency Reports indicate, Twitter has not complied with any of those requests. *See, e.g., Fig. 1 below* (showing Twitter's Transparency Report table reflecting 0% compliance for information requests from Russia).

Fig. 1: Summary of Information Requests from Russia

Russia

Russia information requests

Report	Account information requests	Percentage where some information produced	Accounts specified
January - June 2017	2	0%	2
July - December 2016	-	-	-
January - June 2016	2	0%	2
July - December 2015	82	0%	87
January - June 2015	43	0%	46
July - December 2014	108	0%	108

2. Has the Russian government compelled the production of customer-created content or personally identifiable information from your company?

3. If so, has your company complied with such efforts by the Russian government to compel the production of customer-created content or personally identifiable information?

4. Has your company ever refused to comply with efforts by the Russian government to compel the production of customer-created content or personally identifiable information? If so, have any of these efforts been successful?

5. Has your company provided any content created by a U.S. person or personally identifiable information about a U.S. person to the Russian government?

6. More specifically, has your company provided to the Russian government the content of any direct messages sent to or from a U.S. person?

7. Has your company provided to the Russian government any information that could be used to determine the location of a U.S. person?

The answers to questions 2-7 are provided in response to question 1.

8. The persona, GUCCIFER2.0 (@GUCCIFER_2), emerged in June 2016 and almost immediately started to post material purportedly hacked from the Democratic National Committee (DNC), linking to a seemingly related WordPress site.

- **When did Twitter become aware that the @GUCCIFER_2 account was posting links to a site containing material hacked from the DNC?**
- **What investigation, if any, did Twitter perform into the @GUCCIFER_2 account?**
- **If an investigation was conducted, to whom were the results provided?**
- **Did Twitter provide any information about @GUCCIFER_2 to U.S. law enforcement or the U.S. Intelligence Community?**
- **Why is the @GUCCIFER_2 permitted to remain active on Twitter?**

Our ability to respond to these questions is limited because we are unable to comment on whether or not we received requests related to any specific law enforcement investigations.

Twitter maintains a dedicated 24/7 team to respond to law enforcement requests. We work closely with U.S. law enforcement agencies around the country and promptly respond to properly scoped and valid legal process. To the extent we are permitted to do so by law, we make available to the Twitter community and the general public bi-annual, high-level reports summarizing the types of requests we receive and Twitter's responses, if any. *See* https://transparency.twitter.com.

Our Rules prohibit users from posting content that violates our private information policy, unless those individuals provide express authorization and permission for Twitter users to do so. An example of private information covered by the policy includes, but is not limited to, a non-public personal email address. Sharing private information on the platform could pose serious safety and security risks for the person whose information is shared and is a violation of the Twitter Rules, which we take very seriously.

When we receive a complaint that private information was shared on our platform, Twitter will take appropriate action under its policies, which can include temporarily suspending the account in question pending removal of the private information posted in violation of the Twitter Rules. Subsequent violations of the Twitter Rules can result in a permanent suspension. Twitter applies these rules to all accounts, including the account referenced in the question, and has taken action when that account or others do not comply with the Twitter Rules.

9. Relatedly, another account, currently also named Guccifer2.0 (@Guccifer2) was one of the first accounts to welcome @GUCCIFER_2 to Twitter, responding to the latter's first Tweet. The @Guccifer2 account has been identified, through Twitter's application programming interface (API), to have been created on June 9, 2016. The moniker "Guccifer2.0" did not publicly exist until @GUCCIFER_2 emerged, which did not occur until June 16, 2016.

- **Has Twitter conducted an investigation as to how the account @Guccifer2 and the @GUCCIFER_2 account interacted?**
- **Has the @Guccifer2 account used any other names or handles? If so, when did the change to Guccifer2.0/@Guccifer2 occur?**
- **What investigation, if any, did Twitter perform into the @Guccifer2 account?**
- **If an investigation was conducted, to whom were the results provided?**
- **Did Twitter provide any information about @GUCCIFER 2 to U.S. law enforcement or the U.S. Intelligence Community?**

Our ability to respond to these questions is limited because we are unable to comment on whether or not we received requests related to any specific law enforcement investigations.

As noted above, Twitter works closely with U.S. law enforcement to promptly respond to properly scoped and valid legal process. To the extent we are permitted to do so by law, we make available to the Twitter community and the general public bi-annual, high-level reports summarizing the types of requests we receive and Twitter's responses, if any. *See* https://transparency.twitter.com. Any law enforcement requests that Twitter received in reference to this account would be addressed consistent with Twitter's Law Enforcement Guidelines.

Twitter users are permitted to change their usernames (also referred to as their handles) and display names over time while maintaining an account on our platform. Once a Twitter user changes a username, the prior username and date of the change is no longer visible on Twitter or

via our application programming interface ("API"). More information about how Twitter account holders can update this information is available through Twitter's Help Center. *See* https://help.twitter.com/en/managing-your-account/change-twitter-handle.

[From Vice Chairman Warner]

10. Outside researchers suggest that as much as 15% of the accounts — or 48 million accounts — on Twitter are automated bots, while Twitter has consistently contended the proportion of automated accounts is much lower.

- **If the researchers are inaccurate with respect to the number of fake accounts on your platform, as you suggest, what steps will you commit to take to assist these researchers with providing the real information?**

We regularly receive and welcome input from researchers and Twitter users about ways in which we can optimize our detection and enforcement methods with respect to false or spam accounts. We are committed to continuing to work on refining those tools and to update the public and the Twitter community periodically about our estimates and analysis. And Twitter is unique in the transparency it provides about public Tweet data via our application programming interface ("API"). Through our API, we provide developers, researchers, and other third parties access to public Twitter content. This service is a hallmark of our commitment to transparency, collaboration, and innovation.

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 although we can make a significant amount of information public, researchers do not have access to critical internal information necessary to make an accurate determination of the scale of spam, false 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 data 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.

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.

A study conducted by the University of Southern California and Indiana University estimated that as much as 15% of Twitter accounts are automated, spam accounts. That estimate, however, 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.

In addition to our ongoing steps to promote transparency about our platform through the API, 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.

11. What are your policies regarding the distribution of hacked and stolen emails from your platform?

The use of Twitter to distribute hacked or stolen emails or documents may implicate a number of the Twitter Rules, including our prohibitions on posting private information, infringement of intellectual property rights, and/or unlawful use. For example, users are prohibited from posting other people's private information, unless those individuals provide express authorization and permission for Twitter users to do so. This policy protects information such as non-public personal email addresses and mobile phone numbers.

In addition, under 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." *See* Twitter User Agreement—Twitter Rules, https://twitter.com/en/tos. Twitter has processes in place for law enforcement and private parties to report Twitter Rules violations to us and to submit legal requests and court orders for removal of illegal content. When Twitter detects content posted in violation of our rules and policies, we may take a range of enforcement actions, including requiring the deletion of specific content, withholding an account or specific Tweets where they are unlawful, or suspending the account.

12. We've seen that bad actors are working across the various platforms to spread their disinformation. It has proven true in the past several months that identifying fake accounts on one platform can facilitate identification on other platforms.

- **Do you share data with other firms?**

- **What concrete steps are you taking to share information to improve detection?**

Twitter has partnered with other platforms to make progress against common threats. In June 2017, for example, 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.

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.

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.

13. Press investigations have found that for as little as $100 it is possible to buy a "bot army" via unscrupulous underground online actors, enabling the coordinated propagation of tweets, including of false information, to create trending topics. The *Daily Beast* publication investigated and was able to buy 1000 Twitter accounts for just $45.[1]

- **Is it a violation of your Terms of Service to buy and sell bots?**

- **What safeguards do you have in place to prevent this type of activity?**

Twitter strictly prohibits the purchasing and selling of account interactions on our platform. We advise our users that, by purchasing followers, Retweets, and likes, they are often purchasing bots, fake, or hacked accounts. Accounts found to have purchased, sold, or promoted the selling of followers, Retweets, and likes are in violation of the Twitter Rules and may be subject to suspension. *See* https://help.twitter.com/en/rules-and-policies/twitter-rules.

The Twitter Rules prohibit using automation tools for the purpose of generating spam—unwanted content consisting of multiple postings either from the same account or from multiple coordinated accounts. While "spam" is frequently viewed as having a commercial element, since it is a typical vector for spreading advertising, Twitter's Rules take an expansive view of spam because it negatively impacts the user experience. Examples of spam violations on Twitter include automatically Retweeting content to reach as many users as possible, automatically

[1] https://www.thedailybeast.com/i-bought-a-russian-bot-army-for-under-dollarl100

Tweeting about topics on Twitter in an attempt to manipulate trends, generating multiple Tweets with hashtags unrelated to the topics of those hashtags, repeatedly following and unfollowing accounts to tempt other users to follow reciprocally, Tweeting duplicate replies and mentions, and generating large volumes of unsolicited mentions.

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.

14. If Twitter feels it is important to allow for bots on its platform, then why not require a disclosure that a bot account is non-human? Why not ensure that your real users know when they are interacting with an automated bot?

Because automated content can originate with any type of account, such disclosure would necessarily be both over-inclusive and under-inclusive. For example, certain automation tools that do not violate our rules enable users to schedule their Tweets to post automatically at a particular time or in response to specific activity. Since they may be utilized by individual users, applying a categorical designation to accounts that employ those tools would not accurately characterize such accounts.

It is important to note, moreover, that not all automation is malicious. Automation is essential for certain informational content, particularly when time is of the essence, including for law enforcement or public safety notifications. Examples include Amber Alerts, earthquake and other storm warnings, and notices to "shelter in place" during active emergency situations. Automation is also used to provide customer service for a range of companies. For example, as of April 2017, users are able to Tweet @TwitterSupport to request assistance from Twitter. If a user reports a forgotten password or has a question about our rules, the initial triage of those messages is performed by our automated system—a Twitter-developed program to assist users in troubleshooting account issues.

To maintain the integrity of our platform and to ensure a positive user experience, Twitter's approach to addressing the spread of malicious automation is to focus on problematic behavior and abuse, and on accounts that engage in such behavior. To that end, 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 automated and acting in a malicious automated and coordinated fashion. When we identify such accounts or abusive activity, we take action to prevent them from interfering with the integrity of the platform and detracting from the positive experience of our users.

[From Senator Collins]

15. What provision in your Terms of Service ensures that political advertisements targeted toward the United States are purchased by an American citizen?

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. 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.

While it is possible that foreign governments may attempt to purchase ads through consultants or management companies, Twitter's upcoming Transparency Center is intended to provide identifying information about any such companies and their other advertising activities on Twitter. That information will better enable users and outside parties to conduct their own research or evaluation regarding particular ads.

If Twitter concluded that any advertiser was running an unlawful ad on Twitter, Twitter would restrict the user from such further action, remove the ad, or both as appropriate.

16. Do your Terms of Service prohibit users from influencing elections in other countries?

Twitter is a global platform with more than 330 million monthly active users around the world. Facilitating organic, robust debate and commentary about important events around the world is a core part of Twitter's mission. So long as our users do not violate the Twitter Rules or the Twitter Terms of Service, we do not place restrictions on the organic content that they choose to Tweet about and share with their followers, including when that content contains commentary about elections in other countries.

By contrast, unlawful interference with elections is prohibited. As noted in 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. For example, during the period leading up to the 2016 election, Twitter took action on Tweets that suggested to users that they can vote by text message. Because the Tweets in question appeared to mislead users into believing that they could vote online or vote by text, Twitter viewed them as an unlawful interference with the voting process and took action against those accounts and Tweets.

17. If a foreign national working on behalf of a foreign intelligence service was an authentic user in real name on your platform, could he post divisive, but non-violent content related to a U.S. election without violating your Terms of Service? Would he be able to purchase political advertising?

Twitter is committed to providing a service that fosters and facilitates free and open democratic debate and that promotes positive change in the world. Twitter has a history of

facilitating civic engagement and political freedom, and we intend for Twitter to remain a vital avenue for free expression here and abroad. As a global platform, we believe that our users benefit from the exchange of ideas with other users around the world. Accordingly, unless an account or a user violates the Twitter Rules or the Twitter Terms of Service—including by engaging in abusive behavior, promoting violence, harassing individuals, or using the platform for unlawful purposes—we do not restrict the type of content they choose to share with their followers based on their nationality or citizenship.

Advertisements generally receive a different type of review than organic Tweets. This is because organic Tweets are generally shown to people who choose to follow the user that sends the Tweet, while ads—Promoted Tweets—can be served to a broader audience, including users who have not chosen to follow the user or account that generated the ad. And Twitter places greater limitations on the type of content that can be promoted with Twitter ads compared to organic content that our users generate. In addition, foreign nationals may be subject to legal restrictions governing campaign contributions and electioneering. If Twitter concluded that any advertiser, including a foreign national, was running an unlawful ad on Twitter, Twitter would restrict the user from such further action, remove the ad, or both as appropriate.

[From Senator Feinstein]

18. Twitter has conceded that the number of people exposed to content from foreign groups online is far more pronounced through organic traffic and fake accounts than it is through paid advertising. Troublingly, it does not appear there is a proven method for combatting the spread of fake accounts created to sow division in society. Although the Committee has heard testimony indicating that Twitter has a number of ways to detect "bot-like" activity, as recently as August 2017, divisive foreign unpaid content designed to polarize and anger the American people could be found on Twitter.

- **What specific actions is Twitter taking to combat this type of divisive unpaid activity on an on-going basis?**

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 450,000 suspicious logins each day that we believe to be generated through automation. In September 2017, our systems identified and challenged an average of four million suspicious accounts per week, which represents more than double our rate of detection at this time last year. Over three million of those accounts were challenged upon signup, before their content or engagements could impact other users. Between June and September 2017, we also suspended more than 117,000 malicious applications for API abuse. Those applications were collectively responsible for more than 1.5 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.

19. One of the more troubling findings from this investigation is the number of targeted voter disengagement efforts promoted through social media.

- **Can you say with certainty that foreign actors did not use U.S. voter registration data to target individuals through <u>both</u> paid and unpaid activity?**

As we explained in connection with the November 1, 2017, hearing before the Committee, as a result of our retrospective review, we have identified ways in which Russian actors engaged with our platform in the period leading up to the 2016 election. Although we have no visibility into what information those actors had at their disposal, users posting organic (unpaid) content cannot target other users based on any criteria, and voter registration is not an

available targeting criterion for ads and promoted Tweets on our platform (*i.e.*, paid Twitter services).

[From Senator Cotton]

20. Do Twitter's Terms of Service prohibit collaboration with Russian intelligence services to influence an election?

Unlawful interference with elections is strictly prohibited and users who engage in malicious activity designed to exert artificial influence on an election are likely to violate any number of Twitter's rules and policies. These may include our rules against spam, malicious automation, abuse or harassment, posting private information, or advertising policies. In addition, any illegal collaboration with Russian intelligence services to influence an election would implicate Twitter's prohibition against use of the "service for any unlawful purpose or in furtherance of illegal activities." Twitter's Rules specify that "international 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.

21. Provided an individual or entity does not violate Twitter's Terms of Service, will they be allowed to use your platform to work with hostile, foreign intelligence services to potentially influence the 2018 and 2020 U.S. elections?

The answer to question 21 has been provided in response to question 20.

22. What is Twitter's justification for allowing entities and individuals such as WikiLeaks, Julian Assange, and Edward Snowden to maintain Twitter accounts?

One of Twitter's core purposes is to help advance the global, public conversation on various topics of interest to our users. As we stated in connection with the Committee hearing, Twitter's values include defending and respecting the user's voice—a two-part commitment to freedom of expression and privacy. Twitter has a history of facilitating civic engagement and political freedom, and we intend for Twitter to remain a vital avenue for free expression here and abroad.

Consistent with our values and commitment to fostering an open exchange of ideas, unless the activity or posted content violates our Terms of Service or Twitter Rules, we do not bar controversial figures from our platform or prohibit accounts from posting controversial content. We believe that barring controversial figures from our platform or removing their controversial Tweets would hide important information that our users should be able to see and debate and would detract from the public dialogue that our platform is intended to promote.

We take action against accounts for Terms of Service and Twitter Rules violations, and we apply those rules consistently to all accounts. So long as those accounts remain in compliance with our policies, we do not take action against their Tweets or suspend them from the platform.

23. Describe how Twitter came to acquire DataMinr.

Twitter did not acquire Dataminr. As has been publicly reported, Twitter owns a 5% stake in Dataminr and maintains an ongoing commercial relationship with that company.

24. Is there any portion of the DataMinr platform that shows non-public facing tweets or messages?

Twitter provides Dataminr only with Tweets that users choose to make public. We do not share non-public facing Tweets or messages with Dataminr.

25. What is Twitter's justification for labeling U.S. Intelligence Community access to information sold by DataMinr as "surveillance"?

We prohibit the use of Twitter data for surveillance purposes by any entity whose primary function is surveillance or the collection of intelligence. This is a longstanding Twitter policy. Twitter is proud to work with a range of government and law enforcement agencies, including the FBI, and to provide Twitter data for public safety, news alerting purposes, or in response to valid legal process.

26. Which Russian and Chinese entities currently have access to information sold by DataMinr?

Dataminr and its business relationships with its end users are managed independently of Twitter. Dataminr has notified Twitter that it does not have existing customers in China or Russia.

[From Senator Heinrich]

27. On November 8, CNN reported that a Twitter account — MAGA Mike King — tweeted more than a dozen times on Election Day, November 7, 2017, a graphic purportedly instructing Virginians how to vote by text. This graphic included the logos of the Democratic Party and its gubernatorial candidate, Ralph Northam. According to CNN, the account remained active for almost three hours out of the 13 hours that the polls were open in Virginia.

- **Did Twitter devote additional resources to monitoring misinformation on Election Day, November 7, 2017?**

- **Why did Twitter take so long to suspend the account?**

When we become aware of malicious uses of our platform, we take immediate action to enforce our rules. We also recognize the importance of events such as elections, and we commit additional resources to respond to law enforcement inquiries concerning abuse, spam, and other malicious uses of our platform during campaign and election periods. We regularly reexamine staffing and resources and adjust as needed, and we are launching a process to ensure that during election periods we are positioned to respond in the most effective and efficient way.

Due to the high volume of content posted to Twitter and the real-time nature of the platform, Twitter is not able to proactively monitor all Tweets posted in relation to those events. Our main line of defense against malicious activity on our platform remains our technological tools and systems, which are capable of detecting malicious automated and spam accounts across their lifecycles, including at the account creation and login phase, as well as when those accounts exhibit 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, as well as by information we obtain through third-party security vendors. Those efforts are further supplemented by user reports, on which we rely not only to address the content at issue but also to calibrate our detection tools to identify similar content as spam.

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. Our efforts have already allowed us to respond more quickly to automation, spam, and malicious activity during high-profile events.

When Twitter received reports about the referenced account, we permanently suspended the user. Our action against this account is consistent with the approach we took against illegal voter suppression Tweets during the 2016 election. Here, however, and well before our manual review of the account's activity resulted in its permanent suspension, Twitter's automated spam detection systems identified malicious behavior originating from this account and took action to hide that user's Tweets from appearing in searches and counting toward trends. Those

automated systems, which we continue to invest in as part of our Information Quality initiative, help us address emerging malicious behavior even before a human reviewer can assess the content.

28. What percentage of Twitter content reviews are conducted by an actual human being rather than via automated review?

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.

Critical to the continued success of our efforts is our ability to leverage our technological advancements and improvements to tackle this problem; given the scale, this is not something that we or anyone else can review and address by hand.

But we do not depend on automated systems alone to address malicious automation, bots, and other abusive content or activity on our platform. 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 activity at issue but also to calibrate our detection tools to identify similar content as spam and to enforce the Twitter Terms of Service and Twitter Rules.

While automation provides significant opportunities to scale enforcement activity, it frequently performs better when it supplements, but does not supplant, human review of content. For example, our automated systems flag accounts that engage in suspicious activity for further manual review by Twitter personnel. Those systems also assist our staff in prioritizing manual reviews of user reports. For accurate and consistent enforcement of the Twitter Rules, many of the most sensitive and nuanced types of Twitter Rules violations are typically conducted by Twitter employees. Sensitive reports of illegal activities, such as voter suppression, are subject to review by a Twitter employee and in coordination with the legal department.

We also recognize that computer algorithms alone are not sufficient to address the problem. Context matters, and Twitter's commitment to protecting users' opinions and expression require that tooling alone does not try and solve these complex issues. 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 resources, information, knowledge, and effort with industry partners, civil society, 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.

29. Are Twitter's content review processes the same now as they were during the 2016 election? If not, how have they changed?

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 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 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 that are likely to be 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 deny 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 450,000 suspicious logins each day that we believe to be generated through automation. In September 2017, our systems identified and challenged an average of four million suspicious

accounts per week, which represents more than double our rate of detection at this time last year. Over three million of those accounts were challenged upon signup, before their content or engagements could impact other users. Between June and September 2017, we have also suspended more than 117,000 malicious applications for API abuse. Those applications were collectively responsible for more than 1.5 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. 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. 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.

As noted above, ability to detect malicious activity and spam on our platform is bolstered by internal, manual reviews conducted by Twitter employees. Those efforts are further supplemented by user reports, which allow us to refine and calibrate our detection tools and carefully review content potentially in violation of the Twitter Rules and Twitter Terms of Service. Twitter has also devoted resources to improving our process for reviewing user reporting, including adding better technology to improve how we rank reports for review and adopting policies to allow more reports filed by observers of abuse to be actioned.

30. In hiring more content reviewers, are your companies simply throwing bodies at a specific problem, or are you fundamentally rethinking how to prioritize which user interactions require additional human oversight and review; if so, how? What other changes have you made in this regard?

The answer to question 30 is provided in response to question 28.

[From Senator Manchin]

31. Does Twitter or any Twitter affiliate use the information security products or services of Kaspersky Lab or any Kaspersky Lab affiliate?

Neither Twitter nor any entity controlled by Twitter uses any Kaspersky Lab or Kaspersky Lab-affiliated products or services.

32. Does Twitter or any Twitter affiliate sell network space to Russia Today or Sputnik news agencies?

Twitter recently off-boarded Russia Today ("RT") and Sputnik and will no longer allow those companies to purchase ad campaigns and promote Tweets on our platform. As we announced in October 2017, Twitter will donate the $1.9 million that RT had spent globally on advertising on Twitter to academic research into elections and civic engagement.

33. If you recently terminated any agreements with Russia Today or Sputnik, on what date did the termination become effective?

Twitter off-boarded RT and Sputnik as advertisers on October 26, 2017.

34. Do Russia Today or Sputnik need to purchase advertising space on your platform, or can they freely maintain a Page or distribute web content via their own or affiliated Twitter accounts?

As noted above, Twitter recently off-boarded RT and Sputnik as advertisers on the platform. With respect to organic (non-paid) content, in contrast, unless those accounts violate the Twitter Rules or the Twitter Terms of Service—including by engaging in abusive behavior, promoting violence, harassing individuals, posting prohibited content, or using the platform for unlawful purposes—we do not restrict the type of content they choose to share with their followers. As a global platform, we believe that our users benefit from the exchange of ideas with other users around the world.

Twitter places greater limitations on the type of content that can be promoted with Twitter ads compared to organic content that our users generate. We draw this distinction because organic Tweets are generally shown to people who choose to follow the user that creates it, while ads—Promoted Tweets—are served to a broader audience, including users who have not chosen to follow the user or account that generated the ad.

35. Does Twitter prohibit, or have any concern about, foreign state-sponsored news organizations posting content via the Twitter platform?

Twitter does not categorically prohibit state-sponsored news organizations from posting organic content on our platform. Twitter is a global company with hundreds of millions of users accessing and engaging with information on the platform from around the world. Access to news and real-time media reports is an essential feature of our platform, regardless of location. As with any other account, we permit news organizations to post content on our platform that is

accessible to their followers, so long as they do not engage in illegal activity or otherwise violate our Terms of Service.

[From Senator Harris]

36. Your company has produced information about Russian propaganda advertisements. Your company has also produced information about Russian propaganda that appeared as ordinary user content. You have not, however, provided information about the legitimate advertisements that accompanied Russian content.

- **How long do you retain placement and billing records for advertisements on your services?**

- **Have you instructed your relevant business units to retain the records of advertisements that accompanied Russian propaganda? If you have not, will you immediately issue that instruction?**

Twitter maintains advertisers' billing records in the ordinary course of business. Those records are and will be retained.

- **How much revenue do you estimate that you earned from the advertising that accompanied Russian propaganda?**

- **Have you notified the advertisers whose advertisements accompanied Russian propaganda?**

- **What do you plan to do with the revenue that you earned from the advertisements that accompanied Russian propaganda?**

Twitter's advertising revenue is primarily driven by our Promoted Products. The way in which advertisers use our platform, and the nature of those Promoted Products, supports an estimate that very little revenue was generated from advertising that "accompanied" Russian propaganda on Twitter, as described below. Our Promoted Products are designed to be incorporated into our platform as native advertising, ideally to be as compelling and useful to our users as organic content on our platform. Given this design, Twitter's advertising differs from other platforms and most Twitter advertising does not accompany particular content. For example, Twitter does not display banner ads that accompany a news story.

Twitter's Promoted Products include Promoted Accounts, Promoted Trends, and Promoted Tweets. Promoted Accounts appear in the same format and place as accounts suggested by our Who to Follow recommendation engine, or in some cases, in Tweets in a user's timeline. Promoted Accounts provide a way for our advertisers to grow a community of users who are interested in their business, products or services.

Promoted Trends appear at the top of the list of trending topics for an entire day in a particular country or on a global basis. When a user clicks on a Promoted Trend, search results for that trend are shown in a timeline and a Promoted Tweet created by our advertisers is displayed to the user at the top of those search results.

Promoted Tweets, in the vast majority of cases, appear within a user's timeline or search results just like an ordinary Tweet, and the advertisement is the Tweet itself. These Promoted Tweets do not "accompany" any specific Tweet, nor are they otherwise linked to a particular account.

A small percentage of Promoted Tweets or Accounts (approximately 7% during the 2016 election period) are served to user profiles and may appear within the profile page of an account that a user chooses to visit. As with other Promoted Product placements, to protect user experience, only a very limited number of Promoted Product impressions will render in a profile when another user views it.

The results of our retrospective review allow reasonable estimates of how much revenue could have been generated by Promoted Tweets served to profiles of accounts linked to the Internet Research Agency ("IRA"). Our review has found that a very small fraction of the total content on Twitter during the pre-2016 election period originated from IRA accounts. Even in the unlikely scenario in which ads appeared in each of the account profiles identified as linked to the IRA, the maximum amount of revenue Twitter would have earned from those ads would be very small, as described below.

The analysis we have conducted supports that estimate. We reviewed data from the election time period concerning the number of impressions that were generally served on user profiles available for ad impressions as well as the average revenue collected from those profiles. Extrapolating from that data to a set of 2,752 accounts such as those that Twitter previously identified as linked to the IRA, would yield an estimate of approximately $400 total revenue during the late 2016 election period for Promoted Products that would have accompanied that number of accounts by appearing in those user profiles. While this is a rough estimate based upon aggregate data, it provides a sense of scale for revenue from the types of ads that could accompany this content given the nature of Twitter's advertising offerings during the relevant time period.

Twitter has recently off-boarded Russia Today ("RT") and Sputnik from running ad campaigns on our platform 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. Twitter subsequently announced that we will donate the $1.9 million that RT had spent globally on advertising on Twitter to academic research into elections and civic engagement.

37. The problems of inauthentic, false, and hyper-partisan content are much broader than Russian propaganda.

- **How many of the accounts on your service do you estimate are inauthentic?**

- **If you are aware of independent estimates of inauthentic or false content on your platforms, please provide those estimates. If you disagree with the estimates, please explain why.**

- **How much of the activity on your service do you estimate is inauthentic or false?**

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.

A study conducted by the University of Southern California and Indiana University estimated that as much as 15% of Twitter accounts are automated, spam accounts. That estimate, however, 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.

Note that Twitter does not prohibit the use of pseudonymous accounts (accounts used by real people for non-spam purposes under any name they choose) provided they comply with the Twitter Rules. Many common and powerful uses of Twitter are enabled by this policy, including allowing religious and/or political dissidents and others to engage in free expression without fear of retribution from oppressive governments. It also allows users to have accounts focused on specific interests and to keep those interests separate from a professional or other account associated with their real name. Other users make creative use of this ability to Tweet in the name of a pet or to engage in parody or satire. Such uses of Twitter do not violate our Rules.

- **If the independent estimates were accurate, how much of your annual revenue would be attributable to inauthentic or false content?**

- **How much of your annual revenue do you estimate is attributable to inauthentic or false content?**

- **Do you have a policy of notifying advertisers when their advertisements accompany inauthentic or false content?**

- **What do you do with the revenue that you earn from advertisements that accompany inauthentic or false content?**

- **How much of the news content that is shared on your services do you estimate is false?**

- **How much of the news content that is shared on your services do you estimate is hyper-partisan?**

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.

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 deny bad automation, improve machine learning to spot spam, and increase the precision of our tools designed to prevent such content from contaminating our platform.

The spread of misinformation online is neither a new phenomenon nor unique to our platform. Twitter takes this issue seriously, but we also recognize that our ability to monitor or control the veracity of the content our users choose to share on the platform is limited. We cannot prevent individuals from lying or exaggerating. And given the scale of activity on our platform—where over 330 million users are Tweeting nearly half a billion Tweets per day in scores of languages—we are not able to assess whether each of those Tweet contains arguably inaccurate information (or assess revenue using that criterion).

We are open to examining new solutions to addressing this problem. But we also recognize that the Twitter community itself remains one of the most powerful tools to addressing the spread of misinformation. While it is true that false information and rumors can spread quickly, accurate information—particular information directed at contesting untruths—propagates in a similarly high velocity.

We have observed our users engage with false information by refuting it: they Retweet it, reply to it, and Tweet original content contradicting it. As we noted in connection with the Committee hearing, in response to the attempted "vote-by-text" effort and similar voter suppression attempts during the 2016 election, Twitter restricted as inaccessible, pending deletion, 918 Tweets from 529 users who proliferated that content. Twitter also 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, because we determined that those accounts were seeking to refute the "text-to-vote" message and alert other users that the information was false and misleading. Notably, those refuting Retweets generated significantly greater engagement across the platform compared to the Tweets spreading the misinformation—8 times as many impressions, engagement by 10 times as many users, and twice as many replies.

- **Have you conducted any studies of how false content performs on your services? If yes, please describe those studies and provide copies.**

- **Have you conducted any studies of how hyper-partisan content performs on your services? If yes, please describe those studies and provide copies.**

Twitter has not conducted such studies. As noted above, our efforts to address malicious activity on our platform focus on behavior rather than content.

38. In the area of state-sponsored hacking, each of your companies has a responsible senior executive and dedicated technical experts.

- **Who is the senior executive responsible for countering state-sponsored information operations? When did that executive assume that responsibility, and what is the scope of the responsibility?**

The threats posed by state-sponsored misinformation operations have the potential to impact many parts of our company, including consumer product, advertising, and information security teams. Twitter's General Counsel and Head of Consumer Product, along with our Head of Trust & Safety and Chief Information Security Officer, are generally responsible for ensuring that the platform remains safe. That responsibility includes overseeing and directing Twitter's response to state-sponsored misinformation and malicious human coordinated and automated activity. Twitter's Information Quality team, formed in in early 2017, reports to Twitter's Head of Consumer Product and is intensively focused on 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.

- **As of November 2016, how many of your technical employees had the primary day-to-day task of countering state-sponsored information operations?**

- **As of today, how many of your technical employees have the primary day-to-day task of countering state-sponsored information operations?**

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.

Critical to the continued success of our efforts is our ability to leverage our technological advancements and improvements to tackle this problem; given the scale, this is not something that we or anyone else can review and address by hand alone. We have been successful at addressing other challenges in other contexts and we believe we can meet this challenge as well. 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. We are confident that the combination of our dedicated teams and our ability to use our detection tools and other technological advancements at our disposal equip us well to confront this ongoing threat.

We also recognize that, at this time, computer algorithms alone are not 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 resources, information, knowledge, and effort with industry partners, civil society, 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.

39. Much of what we now know about Russian propaganda is because of academic researchers and investigative journalists. These groups do not currently have access to the data that they need to inform the public and to build tools for detecting state-sponsored information operations. For example, these groups generally cannot assess the full set of public user activity associated with a specific topic, nor can they analyze the behavior of accounts associated with state-sponsored information operations. Providing access to this data need not come at the expense of user privacy, since these groups could be bound by non-disclosure agreements and use privacy-preserving algorithms to conduct their studies.

- **Will you commit to, by the end of the year, providing five or more independent, non-profit entities with access to the data they need to understand and counter state-sponsored information operations? If you will, please provide specifics and a timeline for how you plan to honor the commitment. If you will not, please explain why.**

Twitter is working to deepen our partnership with independent researchers. An example of this commitment is the significant resources we are dedicating to the effort. Twitter recently off-boarded Russia Today ("RT") and Sputnik from running ad campaigns on our platform 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. Twitter subsequently announced that we will donate the $1.9 million that RT had spent globally on advertising on Twitter to academic research into elections and civic engagement.

We are also implementing changes to our data services that will make our public data more accessible than before for research purposes. These changes include, for example, new services which offer developers greater historical search access than was previously accessible. Twitter's data services are unique in the industry, offering insights into the Twitter platform that other companies do not provide.

In addition, Twitter is launching an industry-leading Transparency Center that will offer the public better visibility into who is advertising on Twitter and how those ads are targeted. That information will better enable users and outside parties to conduct their own research or evaluation regarding particular ads.

40. Similarly, much of what we now know about inauthentic, false, or hyper-partisan content is because of independent groups.

- **Will you commit to, by the end of the year, providing five or more independent, non-profit entities with access to the data they need to understand the prevalence and performance of inauthentic, false, or hyper-partisan content on your services? If you will, please provide specifics and a timeline for how you plan to honor the commitment. If you will not, please explain why.**

The answer to question 40 has been provided in response to question 39.

41. Addressing state-sponsored information operations will continue to require cooperation among private sector entities and with the government.

- **Have you established a formal mechanism for promptly sharing actionable information about state-sponsored information operations with other online services, similar to the mechanisms that already exist for sharing information about state-sponsored cybersecurity threats? If not, will you commit to developing such a mechanism?**

- **The FBI is the federal agency responsible for countering foreign propaganda. Do you have a written policy of promptly sharing what you learn about state-sponsored information operations with the FBI? If not, will you commit to developing such a policy?**

Twitter agrees that cooperation with other private-sector entities and the government is necessary to address state-sponsored information operations. Twitter engages in information sharing with its industry counterparts on a variety of threats and is committed to maintaining such cooperative efforts.

Twitter has partnered with other platforms to make progress against common threats. In June 2017, for example, we launched the Global Internet Forum to Counter Terrorism (the "GIFCT"), a partnership among Twitter, YouTube, Facebook, and Microsoft. The GIFCT will facilitate, among other things, information sharing, technical cooperation, and research collaboration, including with academic institutions.

The GIFCT 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, Oath, and Snap have also joined this joint initiative, and we are working to add several additional companies in 2018.

Twitter also engages in regular discussions with law enforcement agencies, including the FBI. We respond promptly to properly scoped legal process and valid requests for information from those agencies. Given the difficulty of identifying and labeling activity as state-sponsored, we also recognize the important role of government information sharing efforts, such as the Intelligence Community's Report about the 2016 U.S. election.

42. You currently have automated systems in place to detect spam and abuse.

- **Do you have an automated system in place to detect state-sponsored information operations? If yes, will you provide this Committee with a private briefing on the system's design and performance? If no, why not?**

Twitter's approach to addressing the spread of malicious, inauthentic automation 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 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 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 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 450,000 suspicious logins each day that we believe to be generated through automation. In September 2017, our systems identified and challenged an average of four million suspicious accounts per week, which represents more than double our rate of detection at this time last year. Over three million of those accounts were challenged upon signup, before their content or engagements could impact other users. Between June and September 2017, we have also suspended more than 117,000 malicious applications for API abuse. Those applications were collectively responsible for more than 1.5 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. 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. 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.

43. You have promised to adopt additional transparency and verification requirements for political advertising.

- **Please detail the new requirements and your timeline for implementing those requirements.**

- **How do you define the political advertisements that are covered by the new requirements? Why did you adopt the definition that you did?**

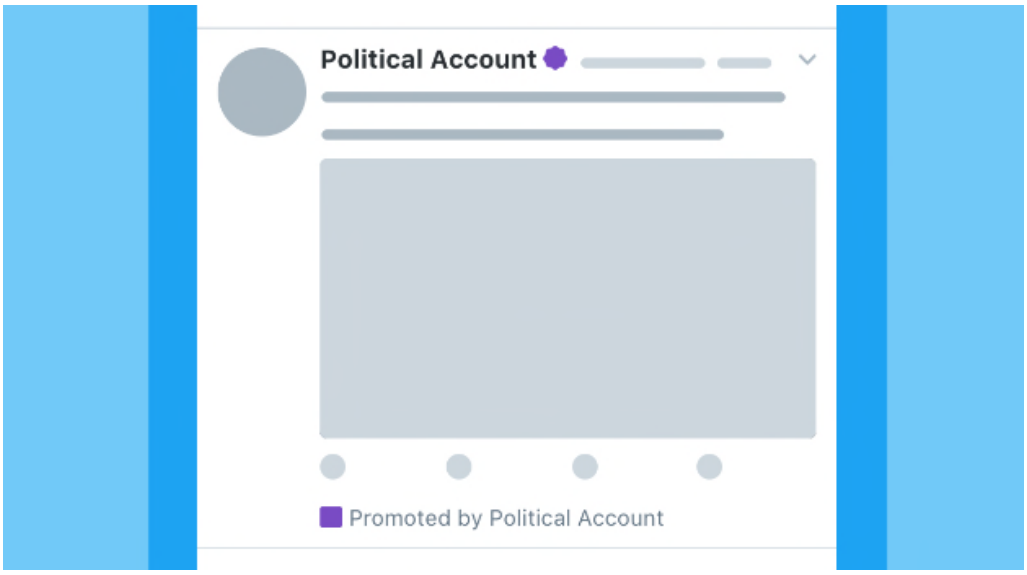
- **Will you commit to including within your definition, at a minimum, advertisements that advocate for or against a specific candidate, political party, piece of legislation, regulatory action, or ballot referendum? If not, why not?**

Twitter's approach to greater transparency in political advertising centers on two components: a new electioneering policy and an industry-leading Transparency Center. We expect to roll out the new policy in the U.S. during the first quarter of 2018.

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. 2 below).

Fig. 2: Template for New Electioneering Ad



Twitter's definition of electioneering ads will be derived from the FEC regulations' definition of that term, which includes any broadcast, cable, or satellite communication that refers clearly to a candidate for federal office, is published 60 days before a general election or 30 days before a primary, convention, or caucus, and is targeted to the relevant electorate (if the candidate is running for Congress).

The goal of the Transparency Center is to 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.

44. Your platform offers a range of advertisement targeting criteria.

- **Which types of targeting criteria, such as demographic, behavioral, lookalike, or email matching, did Russia use for its information operations?**

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.

45. Have you seen any evidence of state-sponsored information operations associated with American elections in 2017, including the gubernatorial elections in Virginia and New Jersey?

Twitter is not aware of any specific state-sponsored attempts to interfere in any American elections in 2017, including the Virginia and New Jersey gubernatorial elections. However, our automated systems for detecting and preventing abuse of our services (including our spam and malicious automation) continually operate with the goal of ensuring that all conversations on Twitter—including those surrounding elections—are spam- and abuse-free. As was publicly reported, we were made aware of a surge in automated followers for a candidate in a recent Senate election, immediately took action, and do not have any indication that the activity was state-sponsored.

46. User reports are an important signal of when an account is not authentic.

- **How frequently do you receive user reports about inauthentic accounts?**

- **What is your process for responding to those reports? How often does that process usually take?**

- **What proportion of those reports result in an account restriction, suspension, or removal?**

- **Among the reports that you decline to take action on, what proportion involve reported accounts that you subsequently identify as inauthentic?**

Twitter receives approximately 5 million reports of spam content or malicious automated Tweets, accounts, or interactions (*e.g.*, follows) per month. The majority of those reports automatically trigger a signal of potential inauthentic behavior by an account seeking to manipulate our platform.

To prevent the misuse of reporting to trigger enforcement actions against users who are not necessarily in violation of the Twitter Rules, user reports are not the sole factor that Twitter considers when taking action against an account. Rather, we use a variety of signals—including overall account behavior and interaction history—to determine whether a report of inauthentic behavior warrants further review or action. Given the broad range of signals Twitter relies on in determining whether an account should be restricted, suspended, or removed, we cannot with any precision assess how many of the 5 million monthly reports directly result in action against the reported account. In addition, Twitter accounts may be reported under different policies, for different Tweets or content, and at different points in the account lifecycle. All of these factors influence Twitter's review and enforcement decisions.

It is not uncommon for an account to be subject to enforcement action later even if Twitter does not initially suspend the account based on the initial report. For example, if Twitter does not take action in response to a user report, the reporting user may be able to submit further information for Twitter to consider during its review. Such supplemental reports may result in Twitter taking action where it may not have previously. In some instances, such as in reports of impersonation or intellectual property infringement, Twitter may require that the report include specific types of information in order to take action against the account. Twitter may not be able to take action until it has received that information. Twitter may also continue to receive reports from the same or other reporters regarding an account or specific content, which may result in Twitter taking action on those new reports at some point after the initial report.

- **How many of the accounts that you have identified as associated with Russian information operations were the subject of a user report? Please provide all the user reports associated with these accounts and the actions that you took in response, including the specific time for the report and each action.**

We received user reports for a small minority of the IRA accounts previously identified to the Committee, prior to the suspension of these accounts from Twitter. A large percentage of those reports related to content from two accounts: @TEN_GOP and @SouthLoneStar. While

many of these reports were not actioned at the time, Twitter has since made substantial changes to its operations and policies to respond more effectively to user reports. For example, in July of 2017, we announced that we are now taking action on 10 times the number of abusive accounts every day compared to the same time last year. We also announced that we now limit account functionality or place suspensions on thousands more abusive accounts each day.

47. Much of the public discussion about state-sponsored information operations on your platforms has centered on the Internet Research Agency. That is not the only group surreptitiously spreading state-sponsored propaganda.

- **What other groups are you tracking that are affiliated with the Russian government?**

- **What other countries do you believe are conducting state-sponsored information operations on your platforms? Please describe the groups that you are tracking for each country, including both government agencies and affiliates.**

As we noted in connection with the Committee hearing, there are technological limits to what we can determine based on the information we can detect regarding a user's origin. Based in part on work conducted by our Information Quality team, we are aware of the fact that, among other things, a high concentration of automated engagement and content originates from data centers and users accessing Twitter via Virtual Private Networks ("VPNs") and proxy servers, which obscure the user's location and affiliation. Twitter's abuse and spam detection and prevention systems and enforcements mechanisms operate without regard to the specific country of origin of an offending Tweet or malicious account. Users who violate the Twitter Rules against abuse, spam, malicious automation, or other forms of prohibited behavior are subject to enforcement, regardless whether they are affiliated with specific state actors.

Information we receive from third parties, including government agencies, security research firms, and NGOs, may allow us to reliably associate certain accounts with particular groups (such as the IRA). And we will continue working with the Committee and other groups to help identify further state-sponsored actors that seek to abuse our services and manipulate activity on our platform.

48. You have confirmed that you have systems for assessing whether a specific account is automated (i.e. a "bot") and whether a specific piece of content is being amplified through automated means.

- **Twitter allows bots to operate on its social network. When and why did Twitter make that decision?**

Automation has not been categorically prohibited on Twitter for years, primarily because we recognize that it often serves a useful and important purpose. Automation is essential for certain informational content, particularly when time is of the essence, including for law enforcement or public safety notifications. Examples include Amber Alerts, earthquake and other storm warnings, and notices to "shelter in place" during active emergency situations. Automation is also used to provide customer service for a range of companies. For example, as of April 11, 2017, users are able to Tweet @TwitterSupport to request assistance from Twitter.

If a user reports a forgotten password or has a question about our rules, the initial triage of those messages is performed by our automated system—a Twitter-developed program to assist users in troubleshooting account issues.

To maintain the integrity of our platform and to ensure a positive user experience, we focus on addressing the spread of *malicious* automation, abusive content, and accounts that engage in such behavior. To that end, 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 automated or acting in a malicious automated and coordinated fashion.

- **How does Twitter differentiate between permissible automated activity (i.e. "benign bots") and impermissible automated activity (i.e. "bad bots")?**

Twitter distinguishes between "good" and "bad" automation based on the behavior of the account, using both algorithm-driven behavior detections and manual reviews. Many of our spam enforcement targeting malicious automation take place automatically and look for signals such as high-volume Tweeting, inhuman response times, and coordinated activities across accounts.

Thus, 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. An example of automation that typically does not trigger our detection tools involves automatic customer service responses to user's Tweets that include a company's handle.

- **When a user visits a profile page, Twitter does not currently indicate whether it believes the profile belongs to a bot. Will you commit to providing a visual indication to users that Twitter believes that the account is a bot, so that users can better understand and evaluate the content that they see? If not, why not?**

- **When a user encounters a piece of content, you do not currently indicate whether the content is being amplified through automated means. Will you commit to providing a visual indication to users of whether Twitter believes that the content has been amplified through automation? If not, why not?**

Because automated content can originate with any type of account, such disclosure would necessarily be both over-inclusive and under-inclusive. For example, certain automation tools that do not violate our Rules enable users to schedule their Tweets to post automatically at a particular time or in response to specific activity. Since they may be utilized by individual users, applying a categorical designation to accounts that employ those tools would not accurately characterize such accounts.

We are committed to keeping Twitter a safe environment, and we continue to invest in improving our systems for detecting and preventing malicious uses of our platform to amplify content using automation. We prohibit the use of automation to artificially amplify content.

Once our systems detect an account as generating automated content or spam, we can take action against that account at either the account level or the Tweet level (*e.g.*, hiding the Tweet, revoking a user's ability to post content on the platform until that user deletes the Tweet in question, or temporarily or permanently suspending the account).

49. Inauthentic accounts can be disabled subsequent to automated or manual review.

- **What role do automated and human employee review play in your decision to disable a suspected inauthentic account?**

Our systems are built to detect malicious automation 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 malicious automated content or 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, and 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 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 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 they delete 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.

- **Do you require that a human employee review a suspected inauthentic account before it is disabled?**

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 activity at issue but also to calibrate our detection tools to identify similar content as spam and to enforce the Twitter Terms of Service and Twitter Rules.

While our automated systems provide significant opportunities to scale enforcement activity on spam (though presenting far greater challenges in other areas where account level signals are less direct and user content itself is the focus), they frequently perform better when supplemented by human review of content. For example, our automated systems flag accounts that engage in suspicious activity for further manual review by Twitter personnel. Those systems also assist our staff in prioritizing manual reviews of user reports. However, the majority of enforcement actions against spam and malicious automated accounts are assisted by automated systems. We continue to invest in improving these systems while ensuring that we maintain a low rate of false positives to protect our users. And where we do not have sufficient confidence in an automated assessment to take immediate action against a suspicious account, that account may be reviewed by a Twitter employee.

- **If so, given the rate at which inauthentic accounts can be regenerated, how do you anticipate remaining ahead of the problem?**

- **What are you doing to improve automation in the process of detecting and disabling inauthentic accounts?**

- **What are you doing to make it more difficult to establish inauthentic 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 450,000 suspicious logins each day that we believe to be generated through automation. In September 2017, our systems identified and challenged an average of four million suspicious accounts per week, which represents more than double our rate of detection at this time last year. Over three million of those accounts were challenged upon signup, before their content or engagements could impact other users. Between June and September 2017, we have also suspended more than 117,000 malicious applications for API abuse. Those applications were collectively responsible for more than 1.5 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.

50. According to recent news reports, Twitter hosted content that was intended to disenfranchise voters in the Virginia gubernatorial election by misleading them about how to vote.

- **What automated and manual processes does Twitter have in place to identify content that is intended to disenfranchise voters?**

- **How quickly does Twitter remove content that is intended to disenfranchise voters? Please provide a histogram or quantile data.**

When Twitter receives reports of illegal voter suppression content, we review the Tweets and accounts in question and, where appropriate, take action to remove the Tweets or suspend the accounts for violating the Twitter Rules. We also employ the use of automated systems to identify "lookalike" posts (*i.e.*, posts which share a known misleading image) which were not directly reported but which should be reviewed.

For example, during the period leading up to the 2016 election, in response to the attempted "vote-by-text" effort and similar voter suppression attempts, Twitter restricted as inaccessible, pending deletion, 918 Tweets from 529 users who proliferated that content. Twitter also permanently suspended 106 accounts that were collectively responsible for 734 "vote-by-text" Tweets. Twitter took action against the first reports of those Tweets within a day or two; once we calibrated our systems to detect the vote-by-text content, our response time decreased and we were able to hide the content rapidly. Because such content removal takes place on a case-by-case basis and involves a fact-specific inquiry and human review, there is not a uniform frequency or pattern to such enforcement actions that we can depict with graphs.

[From Senator McCain]

51. Current campaign finance law establishes disclosure standards for television, radio, and print media. The Pew Research Center recently found that 65 percent of Americans identified an Internet-based source as their leading source of information in the 2016 election.

- **Under current law, to what extent is Twitter responsible for providing a similar quality of disclosure to the public?**

Current campaign finance laws require advertisers to include disclosure language on certain "public communications" and "electioneering communications." Compliance with FEC regulations and guidance from advisory opinions rests with the advertiser, rather than on Twitter or the television, radio, print or digital provider on which an advertiser runs an ad. The Federal Election Commission—through its regulations and advisory opinions—has advised advertisers who disseminate paid communications on digital platforms that they may be able to rely on regulatory exceptions to the disclaimer requirements for "small items" or for communications where including a disclaimer is "impracticable."

Twitter will take further steps to promote transparency and public understanding through the Transparency Center that we publicly announced last year. The goal of the Transparency Center is to offer the public increased visibility into 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 that will be 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 with launching the Transparency Center to ensure the best experience for our users.

52. Your platform hosts thousands of tweets per second, or billions of tweets every year.

- **Given the challenge of monitoring such a vast amount of content, to what extent is the monitoring of campaign advertisements automated?**

- **Do you feel that this automation is sufficient in capturing bad actors?**

Twitter relies on two methods to prevent prohibited promoted content from appearing on the platform: a proactive method and a reactive method. Proactively, Twitter relies on custom-built algorithms and models for detecting Tweets or accounts that might violate its advertising policies. Reactively, Twitter takes user feedback through a "Report Ad" process, which flags an ad for manual human review.

On the proactive side, an advertisement and advertiser account that is subject to manual review is first reviewed by a set of machine classifiers that are built to detect Twitter Policy violations; any suspicious ads that those models flag are subsequently reviewed by Twitter

personnel. On the reactive side, when we receive reports through our "Report an Ad" service, those ads are similarly subject to manual review.

We believe the balance of proactive and reactive review allows us to actively enforce our policies through an effective process that incorporates user feedback. We continue to invest in improving our detection tools and developing new machine-learning models to improve detection accuracy and remain up-to-date with new trends and new Twitter policies.

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. 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. While it is possible that foreign governments may attempt to purchase ads through consultants or management companies, Twitter's upcoming Transparency Center is intended provide identifying information about any such companies and their other advertising activities on Twitter. That information will better enable users and outside parties to conduct their own research or evaluation regarding particular ads.

With respect to organic content, Twitter dedicates significant resources to addressing and blocking the use of malicious automation, bots, and other coordinated activities on our platform. 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.

Critical to the continued success of our efforts is our ability to leverage our technological advancements and improvements to tackle this problem; given the scale, this is not something that we or anyone else can review and address by hand. We have been successful at addressing other challenges in other contexts and we believe we can meet this challenge as well. 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.

We also recognize that computer algorithms alone are not 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 resources, information, knowledge, and effort with industry partners, civil society, 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.

- **What sort of appeals process is in place in order to prevent faulty sorting?**

A user suspended from Twitter can file an appeal directly from the Twitter mobile application or from our website. Twitter's Help Center, accessible at http://help.twitter.com, provides additional information to users about how to an appeal of a suspension.

Suspended advertisers are also able to file appeals. In many cases, an advertiser will work with their account representative in order to do so. If an advertiser does not have a Twitter account representative, or if the advertiser prefers to proceed independently, the advertiser can file an appeal through Twitter's advertiser support form.

53. Twitter recently announced efforts to make campaign advertising more transparent, including the development of a "transparency center," and harsher penalties for electioneering advertisers that violate policies.

- **How do you plan to detect false disclosure information and enforce your policies?**

Twitter is developing mechanisms and processes for verifying the information we will ask our advertisers to disclose, and policies governing enforcement of and compliance with the applicable ads policies. 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.