U.S. Senate Committee on the Judiciary
Subcommittee on the Constitution

Responses to Questions for the Record for Mr. Carlos Monje, Jr.
Director of Public Policy and Philanthropy, U.S. & Canada

SENATOR TED CRUZ

1. Please state the number of users or advertisement purchasers elected to any political office or standing as a candidate for any political office in the United States (including any state, local, or municipal office) that have been banned, shadowbanned, or in any other way had posts, content, or advertisements demoted, downgraded, restricted, or blocked (whether permanently or temporarily) by Twitter, any of its employees or contractors, or any algorithm designed by Twitter or any of its employees or contractors. In providing this answer, please include all incidents involving any restriction on content or advertising, even if Twitter subsequently reversed or altered its decision.

Twitter does not use political viewpoints, perspectives, or party affiliation to make any decisions, whether related to automatically ranking content on our service or how we develop or enforce our rules. Our rules are not based on ideology or a particular set of beliefs. Instead, the Twitter Rules are based on behavior. We do not shadowban anyone based on political ideology. In fact, from a simple business perspective and to serve the public conversation, Twitter is incentivized to keep all voices on the platform.

a. Please provide a complete list of the above-described incidents, naming each user or advertisement purchaser affected, the post(s), content or advertisement(s) that lead to Twitter's decision, and the political affiliation of the user or advertisement purchaser elected to or standing for political office. If Twitter is unable to provide a complete list, please provide the most complete list possible after a reasonable and thorough investigation, including, without limitation, all such incidents that are already a matter of public record.

Twitter does not use political viewpoints, perspectives, or party affiliation to make any decisions. In regard to the removal of accounts, our biannual Twitter Transparency Report highlights trends in enforcement of our Rules, legal requests, intellectual property-related requests, and email privacy best practices.
We provide aggregate numbers of accounts we have actioned across six categories of terms of service violations in this report, which can be found at

transparency.twitter.com. Due to security and privacy concerns, we often cannot discuss individual incidents, but as I noted in my testimony, we action accounts across the world and across the political spectrum.

b. Does Twitter take the political affiliation of any user or advertisement purchaser that is elected to political office or standing for political office in the United States into account when determining whether to take any adverse action regarding that user or advertisement purchaser? For purposes of this question, please disclose instances when any individual moderator has ever taken such factors into account in making the decision to restrict any content or advertising in any way on behalf of Twitter, even if such consideration was contrary to Twitter policy.

No.

c. Does Twitter require or provide any internal training or education to moderators or administrators of its platform regarding how to enforce Twitter's policies in a politically neutral manner? If so, please indicate whether this training is mandatory or optional, what positions at Twitter may or must attend such training, the frequency with which these positions are required or able to attend such training, and the nature, extent, and duration of the training.

Yes. Twitter moderators are provided ongoing training on how to enforce our rules impartially. Notably, because 79 percent of our users are outside of the United States, much of the training involves being sensitive to cultural and language differences across the world. The executive leadership at Twitter recently reorganized the structure of the company to allow our valued employees greater durability, agility, invention, and entrepreneurial drive. The reorganization simplified the way we work, and enabled all of us to focus on health of our platform. Twitter uses a combination of machine learning and human review to adjudicate abuse reports and whether they violate our rules.

One of the underlying features of our approach is that we take a behavior-first approach. That is to say, we look at how accounts behave before we look at the content they are posting. This is how we were able to scale our efforts globally. Twitter employs extensive content detection technology to identify and police harmful and abusive content embedded in various forms of media on the platform. We currently employ approximately 1,500 full time employees and contractors across the globe who are directly involved in enforcing our rules on Twitter, all of whom work in different ways to improve the health of the platform. We have made

the health of Twitter our top priority, and our efforts will be measured by how we help encourage more healthy debate, conversations, and critical thinking on the platform. Conversely, abuse, automation, hateful conduct, terrorism, and manipulation will detract from the health of our platform. The behavioral ranking that Twitter utilizes does not consider in any way political views or ideology. It focuses solely on the behavior of all accounts.

d. Does Twitter take the stance on any political issue—for example, abortion—that a user or advertisement purchaser that is elected to political office or standing for political office has adopted into account when determining whether to take any adverse action regarding that user or advertisement purchaser? For purposes of this question, please disclose whether any individual moderator has ever taken such factors into account in making the decision to restrict any content or advertising in any way on behalf of Twitter, even if such consideration was contrary to Twitter policy.

No.

e. Conversely, does Twitter take such adopted stances into account when providing advertisement rates, coverage, duration, or any other factor affecting the cost or quality of an advertisement on Twitter? Again, for purposes of this question, please disclose whether any of Twitter's employees has ever taken such factors into account, even if such consideration was contrary to Twitter policy.

No. For more information about how Twitter's ads marketplace works, please see: https://business.twitter.com/en/help/troubleshooting/bidding-and-auctions-faqs.html

2. Has Twitter ever conducted any investigation, whether formal, informal, or otherwise, to determine whether its content moderation polices or advertising rules have a disparate impact on users or advertisers based on partisan identity (e.g. Republican) or issue positions (e.g. pro-life)?

Yes.

a. If so, please provide the results of such investigation.

In preparation for this hearing and to better inform the members of the Subcommittee, our data scientists analyzed Tweets sent by all members of the House and Senate that have Twitter accounts for a five-week period spanning February 7, 2019, until March 17, 2019. We learned that, during that period, Democratic members sent 8,665 Tweets and Republican members sent 4,757. Democrats on

average have more followers per account and have more active followers. As a result, Democratic members in the aggregate receive more impressions or views than Republicans.

We conducted a similar analysis in September 2018 with the same results. Our data scientists analyzed Tweets sent by all members of the House and Senate that have Twitter accounts for a 30 day period spanning July 23, 2018 until August 13, 2018. We learned that, during that period, Democratic members sent 10,272 Tweets and Republican members sent 7,981. Democrats on average have more followers per account and have more active followers. As a result, Democratic members in the aggregate receive more impressions or views than Republicans.

Despite this greater number of impressions, after controlling for various factors such as the number of Tweets and the number of followers, and normalizing the followers' activity, we observed that there is no statistically significant difference between the number of times a Tweet by a Democrat is viewed versus a Tweet by a Republican. In the aggregate, controlling for the same number of followers, a single Tweet by a Republican will be viewed as many times as a single Tweet by a Democrat, even after all filtering and algorithms have been applied by Twitter. Our quality filtering and ranking algorithm does not result in Tweets by Democrats or Tweets by Republicans being viewed any differently. Their performance is the same because the Twitter platform itself does not take sides.

b. If not, why not? Will Twitter conduct such an investigation and provide the results of that investigation?

Not applicable.

c. Has Twitter ever conducted any investigation, whether formal, informal, or otherwise, to determine whether its content moderation policies or advertising rules have a disparate impact on users who advocate for or against certain political or issue positions (e.g. abortion)?

Yes.

i. If so, please provide the results of such investigation.

Despite this greater number of impressions, after controlling for various factors such as the number of Tweets and the number of followers, and normalizing the

followers' activity, we observed that there is no statistically significant difference between the number of times a Tweet by a Democrat is viewed versus a Tweet by a Republican. In the aggregate, controlling for the same number of followers, a single Tweet by a Republican will be viewed as many times as a single Tweet by a Democrat, even after all filtering and algorithms have been applied by Twitter. Our quality filtering and ranking algorithm does not result in Tweets by Democrats or Tweets by Republicans being viewed any differently. Their performance is the same because the Twitter platform itself does not take sides.

In preparation for this hearing and to better inform the members of the Subcommittee, our data scientists analyzed Tweets sent by all members of the House and Senate that have Twitter accounts for a five-week period spanning February 7, 2019, until March 17, 2019. We learned that, during that period, Democratic members sent 8,665 Tweets and Republican members sent 4,757. Democrats on average have more followers per account and have more active followers. As a result, Democratic members in the aggregate receive more impressions or views than Republicans.

We conducted a similar analysis in September 2018 with the same results. Our data scientists analyzed Tweets sent by all members of the House and Senate that have Twitter accounts for a 30 day period spanning July 23, 2018 until August 13, 2018. We learned that, during that period, Democratic members sent 10,272 Tweets and Republican members sent 7,981. Democrats on average have more followers per account and have more active followers. As a result, Democratic members in the aggregate receive more impressions or views than Republicans.

ii. If not, is Twitter willing to conduct such an investigation and provide its results?

Please see the two analyses described above that demonstrated our quality filtering and ranking algorithm does not result in Tweets by Democrats or Tweets by Republicans being viewed any differently.

3. Yes or No: Does Twitter consider itself a platform that is open to all ideas and all forms of expression that are protected by the First Amendment?

Like any other private citizen, Twitter has a First Amendment right to free speech. Twitter speaks when it enforces our rules about what content should and should not be allowed on our platform. For example, by prohibiting members of violent

extremist groups from using the platform, Twitter is exercising its First Amendment right to declare its disapproval of content and messages from such groups.

a. Yes or No: Does Twitter consider itself to be a modern equivalent to the historical public square?

Twitter is used at times by many people around the world like a public square. It is not the only place sometimes used like a public square on the Internet, and not everyone uses it that way, but it can be a place where people from around the world come together in an open and free exchange of ideas.

b. Yes or no: Does Twitter consider itself to be a neutral public forum?

Twitter is an open communications platform, where there is free exchange of ideas, on topics as diverse as sporting events, award shows, natural disasters, political movements, and the latest music. In developing and enforcing our rules for the service we seek to be impartial, and as a service we believe in impartiality strongly.

c. When Twitter crafts its content moderation policies and advertising rules, does it seek to craft rules that are viewpoint neutral?

In developing and enforcing our rules for the service we seek to be impartial, and as a service we believe in impartiality strongly.

d. In practice, does Twitter moderate content and enforce its advertising rules on a viewpoint-neutral basis?

In developing and enforcing our rules for the service we seek to be impartial, and as a service we believe in impartiality strongly.

e. Has Twitter ever made any moderating decision or enforced its advertising rules in a non-viewpoint-neutral manner? Please describe all such incidents, even if they were contrary to Twitter policy.

We are committed to combating abuse motivated by hatred, prejudice, or intolerance. We also remove terrorism and violent extremism content from our platform. We enforce our Terms of Service and Twitter Rules.

4. Has Twitter ever dismissed, demoted, fired, or otherwise taken adverse employment action against an employee on the basis of political speech that the employee undertook within the company, on Twitter, or elsewhere?

No.

a. If so, please list each such incident. If federal law requires Twitter to keep any of these incidents or their details confidential, please disclose as much information as federal law permits and anonymize the instances through appropriate pseudonyms and redactions. If Twitter does so, please note the legal basis for such redaction or confidentiality.

Not applicable.

5. Has Twitter ever dismissed, demoted, fired, or otherwise taken adverse employment action against an employee on the basis of that employee's discrimination against content or viewpoint within the company or on Twitter's platform?

No.

a. If so, please list each such incident. If federal law requires Twitter to keep any of these incidents or their details confidential, please disclose as much information as federal law permits and anonymize the instances through appropriate pseudonyms and redactions. If Twitter does so, please note the legal basis for such redaction or confidentiality.

Not applicable.

6. Does Twitter provide access to its services on a viewpoint-neutral basis? For this question and its subparts, please construe "access to its services" and similar phrases broadly, including the position or order in which content is displayed on its services, the position or order in which users or content appear in searches (or whether they appear at all), whether users or content are permitted to purchase advertisements (or be advertised), the rates charged for those advertisements, and so on.

Twitter provides access to our service to anyone abiding by our Terms of Service and Twitter Rules.

a. Yes or no: Has Twitter ever discriminated among users on the basis of viewpoint when determining whether to permit a user to access its services? If so, please list each instance in which Twitter has done so.

No.

i. If so, does Twitter continue to do so today, or when did Twitter stop doing so?

Twitter does not discriminate among users on the basis of viewpoint when determining whether to permit a user to access our service.

ii. If so, what viewpoint(s) has Twitter discriminated against or in favor of? In what way(s) has Twitter done so?

Twitter does not discriminate among users on the basis of viewpoint when determining whether to permit a user to access our service.

iii. If so, does Twitter consider only on viewpoints expressed on Twitter, or does it discriminate among users based on viewpoints expressed elsewhere? Has Twitter ever based its decision to permit or deny a user access to its services on viewpoints expressed off Twitter?

Twitter does not discriminate among users on the basis of viewpoint when determining whether to permit a user to access our service.

iv. Yes or no: Excluding content encouraging physical self-harm, threats of physical violence, terrorism, and other content relating to the credible and imminent physical harm of specific individuals, has Twitter ever discriminated against users or their communications on the basis of viewpoint in its services? If so, please list each instance in which Twitter has done so.

No.

v. Excluding content encouraging physical self-harm, threats of physical violence, terrorism, and other content relating to the credible and imminent physical harm of specific individuals, has Twitter ever discriminated against users or their communications on the basis of content in its services? If so, please list each instance in which Twitter has done so.

Our terms of service have a number of prohibitions that extend beyond the limited exclusions described above. Please see the following information here. https://help.twitter.com/en/rules-and-policies/twitter-rules

vi. Yes or no: Has Twitter ever discriminated against American users or content on the basis of an affiliation with a religion or political party? If so, please list each instance in which Twitter has done so and describe the group or affiliation against which (or in favor of which) Twitter was discriminating.

No.

b. Yes or no: Has Twitter ever discriminated against any American users or content on its services on the basis of partisan affiliation with the Republican or Democratic parties? This question includes advocacy for or against a party or specific candidate or official. If so, please list each instance and the party affiliation discriminated against.

No.

c. Yes or no: Has Twitter ever discriminated against any American users or content on its services on the basis of the user's or content's advocacy for a political position on any issue in local, State, or national politics? This question includes but is not limited to advocacy for or against abortion, gun control, immigration, criminal justice reform, and net neutrality.

No.

d. Yes or no: Has Twitter ever discriminated against any American users or content on its services on the basis of the user's or content's religion, including advocacy for one or more tenets of that religion? If so, please list each such instance in which Twitter has done so and identify the religion, religious group, or tenet against which Twitter discriminated.

No.

7. Yes or no: Has Twitter ever discriminated between users in how their content is published, viewed, received, displayed in "trending" or similar lists, or otherwise in any function or feature, based on the user's political affinity, religion, religious tenets, ideological positions, or any ideological or philosophical position asserted? This includes either the insertion of a topic or individual into the "trending" topics feature or the prohibition of a topic's or individual's display in the "trending" topics feature. If so, please list each such incident as well as the basis on which Twitter discriminated against that user or content.

No. Please see more information about how trends work here: https://help.twitter.com/en/using-twitter/twitter-trending-faqs

8. How does Twitter moderate, prohibit, ban, or in any way otherwise restrict content or advertising that it considers to be "hate speech?"

a. How does Twitter define the term "hate speech?"

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

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

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

b. What objective metrics, if any, does Twitter use to determine whether a statement constitutes "hate speech?"

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

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

c. To what extent does whether a statement constitutes "hate speech" depend on the subjective judgment of the moderator reviewing the content?

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

d. What training, if any, does Twitter provide moderators in restricting content or users' access to the platform on the basis of "hate speech" in a way that does not otherwise discriminate on the basis of viewpoint or partisan affiliation?

Our moderators engage in frequent training on our Twitter Terms of Service, Twitter Privacy Policy, and Twitter Rules.

e. Has Twitter ever changed its definition of "hate speech" or how it applies its hate speech policies? If so, please describe those changes.

We continually update our policies and how we implement them as threats to the health of the platform evolve. For more information about the process, please see: https://help.twitter.com/en/rules-and-policies/enforcement-philosophy

f. Does Twitter moderate, prohibit, ban, or in any way otherwise restrict content or advertising now on the basis of that ad or content being hate speech that it would have permitted at some previous time?

Not applicable. Twitter imposes higher standards for our promoted products and you can read more about them here:

https://business.twitter.com/en/help/ads-policies/introduction-to-twitter-ads/twitter-ads-policies.html

9. Did or does Twitter collaborate with or defer to any outside individuals or organizations in determining whether to classify a particular statement as "hate speech?" If so, please list the individuals and organizations.

No, we do not defer to any outside individuals or organizations in establishing our rules. The Twitter Trust and Safety Council provides input on our safety products, policies, and programs. Twitter works with safety advocates, academics, and researchers; grassroots advocacy organizations around the globe that rely on Twitter to build movements; and community groups working to prevent abuse. We have over 60 partners focused on specific issues, including mental health, child protection and online safety.

10. Did or does Twitter collaborate with or defer to any outside individuals or organizations in determining whether a given speaker has committed acts of "hate speech" in the past? If so, please list the individuals and organizations.

No. Our rules are largely focused on activity that is currently on the platform, as opposed to offline activity, with a notable exception being our policies against terrorist organizations and violent extremist groups.

a. Does Twitter review these groups' internal procedures in determining whether an entity is a "hate group" or committing acts of "hate speech" to determine that these determinations are not made on a partisan basis?

Not applicable.

11. Under what circumstances does Twitter ban or otherwise limit the content of individuals or organizations who have spoken "hate speech" on its platform aside from the offending content?

Our enforcement options have expanded significantly over the years. We originally had only one enforcement option: account suspension. Since then, we've added a range of enforcement actions and now have the ability to take action at the Tweet, Direct Message, and account levels. Additionally, we take measures to educate individuals that have violated our rules about the specific Tweets in violation and which policy has been violated. We also continue to improve the technology we use

to prioritize reports that are most likely to violate our rules and last year we introduced smarter, more aggressive witness reporting to augment our approach.

12. Twitter is not subject to the First Amendment's limitations against government censorship, and is free to moderate content as it sees fit in the same way that the New York Times or Wall Street Journal do.

a. As Twitter defines "hate speech," does Twitter believe that its hate speech policy affects content that would be protected from government censorship by the First Amendment?

Like any other private citizen, Twitter has a First Amendment right to free speech. Twitter speaks when it enforces our rules about what content should and should not be allowed on our platform. For example, by prohibiting members of violent extremist groups from using the platform, Twitter is exercising its First Amendment right.

b. If so, please describe what content would be subject to Twitter's policy that is nonetheless protected from government censorship by the First Amendment.

Like any other private citizen, Twitter has a First Amendment right to free speech. Twitter speaks when it enforces our rules about what content should and should not be allowed on our platform. For example, by prohibiting members of violent extremist groups from using the platform, Twitter is exercising its First Amendment right

13. Twitter states on its website that, per its Hateful Conduct Policy, Twitter will remove hate speech, which it describes as content that "promote[s] violence against or directly attack[s] or threaten[s] other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." Yes or no: Does Twitter limit its definition of hate speech only to content that "directly attacks or threatens" people based on the aforementioned characteristics?

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

14. What standard or procedure has Twitter applied now and in the past in determining whether content “directly attacks or threatens” an individual or group based on a protected characteristic under Twitter’s community standards?

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

15. Yes or no: Has Twitter ever removed content for “hate speech” that did not directly attack or threaten a person on the basis of his or her race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, or gender identity, or serious disabilities or diseases? If so, what criteria did Twitter use to determine that the content violated Twitter’s policy?

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

16. Can expressing a controversial opinion itself—when not transmitted to a particular user or indicated as directed at a particular individual, given the circumstances—count as a “direct attack or threat” that violates Twitter’s “hate speech” policy?

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

17. If an individual posted any of the following statements, standing alone and not directed to any Twitter user in particular, would that statement violate Twitter’s “hate speech” policy? To the extent that the decision would depend on additional facts, please describe whether the

statement would prompt an investigation to determine whether it constitutes "hate speech," and whether the decision would involve algorithmic or human decision making.

a. There are only two sexes or two genders, male and female.
b. Men cannot become women.
c. A person's sex or gender are immutable characteristics.
d. Caitlin Jenner, f/k/a Bruce Jenner, is a man.
e. Sex reassignment surgery is a form of bodily mutilation.
f. The abortion of an unborn child is murder.
g. It should be a crime to perform or facilitate an abortion.
h. Same-sex marriage is wrong.
i. No person of faith should be required to assist a same-sex wedding by providing goods or services to a same-sex marrying couple.
j. When an individual enters the marketplace, he gives up the right to choose whether to support a same-sex marriage.
k. Islam is a religion of war.
l. All white people are inherently racist.
m. All black people are inherently racist.
n. Donating to the NRA funds the murder of children, such as those slain in Parkland, Florida.
o. Donating to Planned Parenthood funds the murder of children, such as those dismembered by Kermit Gosnell.
p. The U.S. should build a wall at its southern border.
q. Illegal aliens need to be sent back to their home countries.

Twitter's mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right – we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives.

We recognize that if people experience abuse on Twitter, it can jeopardize their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. This includes; women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities. For those who identity with multiple underrepresented groups, abuse may be more common, more severe in nature and have a higher impact on those targeted.

We are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized. For this reason, we prohibit behavior that targets individuals with abuse based on protected category.

We may suspend an account if it has been reported to us as violating our Rules surrounding abuse. When an account engages in abusive behavior, like sending threats to others, we may suspend it temporarily or, in some cases, permanently. An individual may be able to unsuspend his or her own account by providing a phone number or confirming an email address.

An account may also be temporarily disabled in response to reports of automated or abusive behavior. For example, an individual may be prevented from Tweeting from his or her account for a specific period of time or may be asked to verify certain information before proceeding.

If an account was suspended or locked in error, an individual can appeal. First, the individual must log in to the account that is suspended and file an appeal. The individual must describe the nature of the appeal and provide an explanation of why the account is not in violation of the Twitter Rules. Twitter employees will engage with the account holder via email to resolve the suspension.

The examples you list above may be violations of different Twitter rules, in addition to our hateful conduct policy, based on the context of the discussion, which is never as simple as it may appear externally.

18. Recently, on a podcast with Sam Harris, Mr. Dorsey said that Twitter "can't afford to be neutral anymore." Does that statement reflect Twitter's official policy?

This quote is taken out of context, despite testimony in the course of the hearing indicating that it is an inaccurate representation.

Mr. Dorsey stated on the podcast titled "Making Sense with Sam Harris" episode 148 dated February 5, 2019:

"I don't believe that we can afford to take a neutral stance anymore. I don't believe that we should optimize for neutrality. I do believe that we should optimize for impartiality. And I do think there is a difference there. To me, neutrality is a lot more passive, a lot more hands off. …

But also I don't think we can be this neutral passive platform anymore because of the threats of violence, because of doxxing, because of troll armies intending to silence someone especially more marginalized members of society. We have to take on an approach of impartiality meaning that we need very crisp and clear rules, we need case studies and case law for how we take action on those rules, and any evolutions of that we're transparent and up front about that."

a. In what ways does Twitter intend to be non-neutral?

Not applicable.

b. Why has Mr. Dorsey decided that Twitter cannot afford to be neutral anymore?

Not applicable.

c. Insofar as Twitter believes that Twitter may act in a non-neutral but nonetheless impartial matter, please explain in detail the difference between neutrality and impartiality.

Not applicable.

19. Has Twitter ever studied or examined, whether formally, informally, or otherwise, the political beliefs or affiliations of its users? If so please disclose the results of those studies or examinations.

Twitter does not ask about or collect political affiliation data about U.S. users directly through the platform. Twitter has conducted surveys on the opinions of the general population in the United States through external research agencies, contracted by Twitter, which comply with Twitter's data privacy and protection policies. The surveys were used for research purposes only and results were not shared externally.

Respondents were randomly selected, and include both Twitter users and non Twitter users. The political affiliation question in our surveys was always optional and self-reported. There was no guaranteed cash compensation associated with participation.

a. How has Twitter used this information? Please explain each use of this information. If any uses contain information that would be protected by law as proprietary or trade secrets, please inform us so that we may arrange for procedures to keep this information appropriately confidential.

The data collected was used for research purposes to better understand perceptions of Twitter. Results were not shared externally. As per Twitter's policy, we do not collect or receive any individualized or personally identifying information (PII) about survey respondents without consent. Research findings and data files for each study are always delivered by the external agency to Twitter in a format that ensures that there is no way in which responses can be traced back to any individual.

b. Has Twitter ever reviewed or made use of third-party studies or examinations of the political affiliations of its users? If so, please explain when and how, including in what ways these conclusions affected Twitter's policies or how Twitter enforces its policies.

Twitter has conducted surveys on the opinions of the general population in the United States through external research agencies, contracted by Twitter, which comply with Twitter's data privacy and protection policies. The surveys were used for research purposes only and results were not shared externally.

20. Has Twitter ever conducted any study or investigation, whether formal, informal, or otherwise, the level of engagement that Twitter users have with accounts held by individuals who are elected to or standing for any political office in the United States? If so, please provide the results of such investigation.

In preparation for this hearing and to better inform the members of the Subcommittee, our data scientists analyzed Tweets sent by all members of the House and Senate that have Twitter accounts for a five-week period spanning February 7, 2019, until March 17, 2019. We learned that, during that period, Democratic members sent 8,665 Tweets and Republican members sent 4,757. Democrats on average have more followers per account and have more active followers. As a result, Democratic members in the aggregate receive more impressions or views than Republicans.

We conducted a similar analysis in September 2018 with the same results. Our data scientists analyzed Tweets sent by all members of the House and Senate that have Twitter accounts for a 30 day period spanning July 23, 2018 until August 13, 2018.

We learned that, during that period, Democratic members sent 10,272 Tweets and Republican members sent 7,981. Democrats on average have more followers per account and have more active followers. As a result, Democratic members in the aggregate receive more impressions or views than Republicans.

Despite this greater number of impressions, after controlling for various factors such as the number of Tweets and the number of followers, and normalizing the followers' activity, we observed that there is no statistically significant difference between the number of times a Tweet by a Democrat is viewed versus a Tweet by a Republican. In the aggregate, controlling for the same number of followers, a single Tweet by a Republican will be viewed as many times as a single Tweet by a Democrat, even after all filtering and algorithms have been applied by Twitter. Our quality filtering and ranking algorithm does not result in Tweets by Democrats or Tweets by Republicans being viewed any differently. Their performance is the same because the Twitter platform itself does not take sides.

21. Under what circumstances does Twitter either ban content criticizing Twitter's decision to restrict content or users, or otherwise require users to remove such content critical of Twitter as a condition of using the platform?

Twitter does not remove content that is critical of Twitter, unless it is a violation of our Twitter Terms of Service, Twitter Privacy Policy, and Twitter Rules.

22. How many individuals at Twitter have the ability to moderate, remove, downgrade, conceal, or otherwise censor content, ban, suspend, warn, or otherwise discipline users, or approve, price, review, or refuse advertisements on the platform? (For this question only, we refer to these individuals as moderators.) This question includes individuals with the power to alter search results and similar mechanisms that suggest additional content to users in order to promote or demote content, whether individually or routinely through an algorithm or by altering any of the platform's search functions. (Please include all employees, independent contractors, or others with such ability at Twitter.)

a. How many moderators work for Twitter? This includes individuals who serve in moderating functions part-time or as independent contractors. This question includes individuals with the power to alter search results and similar mechanisms that suggest additional content to users in order to promote or demote content, whether individually or routinely through an algorithm or by altering any of the platform's search functions.

Twitter uses a combination of machine learning and human review to adjudicate abuse reports and whether they violate our rules. One of the underlying features of our approach is that we take a behavior-first approach. That is to say, we look at how accounts behave before we look at the content they are posting. This is how we were able to scale our efforts globally. Twitter employs extensive content detection technology to identify and police harmful and abusive content embedded in various forms of media on the platform. We currently employ approximately 1,500 full time employees and contractors across the globe who are directly involved in enforcing our rules on Twitter, all of whom work in different ways to improve the health of the platform. We have made the health of Twitter our top priority, and our efforts will be measured by how we help encourage more healthy debate, conversations, and critical thinking on the platform. Conversely, abuse, automation, hateful conduct, terrorism, and manipulation will detract from the health of our platform.

b. Who are the individuals responsible for supervising these moderators as their conduct relates to American citizens, nationals, businesses, and groups?

Twitter uses a combination of machine learning and human review to adjudicate abuse reports and whether they violate our rules. Twitter employs extensive content detection technology to identify and police harmful and abusive content embedded in various forms of media on the platform. We use PhotoDNA and hash matching technology, particularly in the context of terrorism or child sexual exploitation. We use these technologies to identify previously identified content in order to surface it for agent review, however, if it is the first time that an image has been seen, it would not necessarily be subject to our technology. It is important to note that we continually expand our databases of known violative content.

c. How many moderators has Twitter had on its platform for each of the calendar years 2006 to 2019? Please provide approximations if exact numbers are impossible to obtain.

Our chief executive officer reorganized the structure of the company to allow our valued employees greater durability, agility, invention, and entrepreneurial drive. The reorganization simplified the way we work, and enabled all of us to focus on health of our platform.

d. How many moderators does Twitter intend to retain for the years 2020 and 2021?

We plan to continue to invest in key areas of our business and ensure that we have the right level of general and administrative support on our key priorities and

objectives. We expect that general and administrative expenses will increase in absolute dollar amounts and vary as a percentage of revenue.

e. On average, how many pieces of content does a moderator remove a day?

In regard to the removal of accounts, our biannual Twitter Transparency Report highlights trends in enforcement of our Rules, legal requests, intellectual property-related requests, and email privacy best practices. The report also provides insight into whether or not we take action on these requests. The Transparency Report includes information requests from governments worldwide and non-government legal requests we have received for account information. Removal requests are also included in the Transparency Report and include worldwide legal demands from governments and other authorized reporters, as well as reports based on local laws from trusted reporters and non-governmental organizations, to remove or withhold content.

f. On average, how many users does a moderator discipline a day?

According to our latest Twitter Transparency Report covering the second half of 2018, across the six Twitter Rules policy categories included, we received reports regarding 11,000,257 unique accounts for possible violations of those Twitter Rules, amounting to a 19% increase compared to the prior reporting period. Of those accounts, 6,388 accounts were reported by known government entities compared to 5,461 reported during the last reported period, an increase of 17%. We have a global team that manages enforcement of our Rules with continuous coverage, in every supported language on the service.

It is worth noting that the raw number of reported accounts is not a consistent indicator of the validity of the reports we receive. During our review process, we may consider whether reported content violates aspects of the Twitter Rules beyond what was initially reported. For example, content reported as a violation of our private information policy may also be a violation of our policies for hateful conduct. If the content is determined to violate any Twitter Rule, it is actioned accordingly. Not all reported accounts are found to violate the Twitter Rules, and reported accounts may be found to violate a different rule than was initially reported. We may also determine that reported content does not violate the rules at all. The volumes often fluctuate significantly based on world events, including elections, national and international media stories, and large conversational moments in social and political culture.

g. On average, how many advertisements does a moderator approve, disapprove, price, consult on, review, or refuse a day?

Twitter takes violations of our Twitter Ads policies, the Twitter Rules, and Terms of Service seriously. We will examine reported violations and take appropriate action, which may include removal of offending advertisements and advertisers from the Twitter Ads platform. The volume of ads on the platform is very small relative to overall Tweet volume.

23. As Twitter has previously acknowledged, Silicon Valley is predominantly politically liberal, and Twitter's employees are likewise predominantly liberal. Republicans and conservatives are concerned that such a political monoculture leads to disproportionate sanctions against conservatives and conservative views, such as those researchers find prevail in academia. To that end, please answer the following questions based on any information Twitter has, whether formal or informal, as to the political beliefs or political involvement of Twitter's personnel.

a. What percentage of Twitter's Board of Directors self-identify as "liberal" or Democrats versus "conservative" or Republicans?

It is the principal duty of the Board to exercise its powers in accordance with its fiduciary duties to the Company and in a manner it reasonably believes to be in the best interests of the Company and its stockholders. It is also the Board's duty to oversee senior management in the competent and ethical operation of the Company. Directors bring to the Company a wide range of experience, knowledge and judgment, and will use their skills and competencies in the exercise of their duties as directors of the Company.

The Nominating Committee works with the Board to determine periodically, as appropriate, the desired Board qualifications, expertise and characteristics, including such factors as business experience and diversity; and with respect to diversity, the Nominating Committee may consider such factors as differences in professional background, education, skill, and other individual qualities and attributes that contribute to the total mix of viewpoints and experience represented on the Board.

The Nominating Committee and the Board evaluate each individual in the context of the membership of the Board as a group, with the objective of having a group that can best perpetuate the success of the business and represent stockholder interests

through the exercise of sound judgment using its diversity of background and experience in the various areas. Each director should be an individual of high character and integrity. In determining whether to recommend a director for reelection, the Nominating Committee also considers the director's past attendance at meetings, participation in and contributions to the activities of the Board and the Company and other qualifications and characteristics set forth in the charter of the Nominating Committee.

b. How many of Twitter's Board of Directors have donated or raised money for Democrats, the Democratic National Committee, or political action committees primarily supporting Democrats? For Republicans and their counterparts?

Our Board of Directors are individuals who can best perpetuate the success of the business and represent stockholder interests through the exercise of sound judgment using its diversity of background and experience in the various areas.

c. What percentage of Twitter's senior management have worked in Democratic administrations? In Republican administrations?

Twitter does not use political ideology as a factor in its hiring decisions.

d. What percentage of Twitter's senior management self-identify as "liberal" or Democrats versus "conservative" or Republicans?

Twitter does not use political ideology as a factor in its hiring decisions.

e. How many of Twitter's senior management have donated or raised money for Democrats, the Democratic National Committee, or political action committees primarily supporting Democrats? For Republicans and their counterparts?

Twitter does not use political ideology as a factor in its hiring decisions.

f. What percentage of Twitter's senior management have worked in Democratic administrations? In Republican administrations?

Twitter does not use political ideology as a factor in its hiring decisions.

g. What percentage of Twitter's moderators self-identify as "liberal" or Democrats versus "conservative" or Republicans?

Twitter does not use political ideology as a factor in its hiring decisions.

h. How many of Twitter’s moderators have donated or raised money for Democrats, the Democratic National Committee, or political action committees primarily supporting Democrats? For Republicans and their counterparts?

Twitter does not use political ideology as a factor in its hiring decisions.

i. What percentage of Twitter’s moderators have worked in Democratic administrations? In Republican administrations?

Twitter does not use political ideology as a factor in its hiring decisions.

j. To the extent that Twitter does not have and cannot reasonably ascertain the information called for above, what steps does Twitter intend to take to gather that information?

Our responses to question 23, subparts (a) through (i) are complete. We do not ask these questions of our employees and believe it would be inappropriate to do so.

24. Does Twitter conduct any voter outreach, for example encouraging users to vote in an election or register to vote in elections?

Yes.

a. If so, does Twitter consider the political party of those reached by its voter outreach efforts when designing or engaging in those efforts?

Yes. We seek only nonpartisan and bipartisan partners.

b. If so, do Twitter’s voter outreach efforts disparately reach registered Democrats or Republicans?

No.

c. Please list each such voter outreach effort that Twitter has conducted, including the year, the election, and the candidates in that election, and the means and extent to which Twitter engaged in voter outreach.

In addition to the strong discussion we hosted on Twitter about the 2018 U.S. midterm election, we also collaborated with a number of non-governmental organizations to promote voter registration, civic engagement, and media literacy, including RockTheVote, Democracy Works, TurboVote Challenge, HeadCount, DoSomething, and Ballotpedia.

In May 2018, we participated in the TurboVote Challenge Summit along with other industry peers and election nonprofits and presented to over 100 attendees. In July and August, Twitter participated in the leadership committees of TurboVote and National Voter Registration Day for voter engagement efforts.

We deployed three #BeAVoter voter assistance prompts displayed in the home timeline of individuals on the service located in the United States aged 18 and older. Each prompt linked to a nonpartisan, nonprofit managed site for voter assistance that facilitated voter registration and identification of polling locations. The secondary link drove individuals to a Tweet compose window to share with their followers how to register to vote.

Twitter further bolstered the visibility of National Voter Registration Day, a nonpartisan day of action supported by many corporate and nonprofit partners. We saw an increase in Tweets with the primary event hashtag for #NationalVoterRegistrationDay, with a two-fold increase over the number of similar Tweets in 2018, a presidential election year. Of those who Tweeted about National Voter Registration Day, 37.9% had not previously Tweeted about the midterm elections in the six months prior.

We also developed Twitter Emoji to drive meaningful, healthy conversation around civic participation.

Twitter also engaged social influencers to mount a video campaign across @TwitterMovies, @TwitterMusic, @TwitterTV, @TwitterSports, @TwitterWomen. Having a range of social influences for our #BeAVoter video series enabled us to reach a wide variety of people based on interest group. The Tweet from Ariana Grande “thank u, vote” is @Twitter’s second most engaged ever, demonstrating a broad interest in consumer political content.

d. Has Twitter or any employees, contractors, or subsidiaries ever engaged in any voter outreach in order to influence the outcome of any election?

The company does not attempt to influence elections.

e. Has Twitter ever conducted any investigation, whether formal, informal, or otherwise, to determine the political leanings or party affiliation of the users it reaches with voter outreach efforts?

No.

f. Has Twitter ever conducted any investigation, whether formal, informal, or otherwise, to determine the political leanings or party affiliation of the users that respond to or interact with voter outreach efforts?

No.

g. Has Twitter ever conducted specific voter outreach efforts to reach any identifiable demographic, including by race, sex, nationality, sexual orientation or gender identity, marital status, geography, language, or age?

No.

h. Has Twitter ever used its platform to influence public debate or the outcome of an election, either through direct communications or through the enforcement of its policies?

No.

25. Jack Dorsey praised an article by two Democrats calling for a "new civil war" against the Republican Party, in which "the entire Republican Party, and the entire conservative movement that has controlled it for the past four decades" will be given a "final takedown that will cast them out" to the "political wilderness" "for a generation or two."

a. Do you or Twitter agree with the premise of this article? It is located here: https://medium.com/s/state-of-the-future/the-great-lesson-of-california-in-americas-new-civil-war-e52e2861f30

Mr. Dorsey has previously explained that he believes this article to be a compelling narrative of competing economic systems in the past, such as labor, and the present, primarily one based on energy. He does not support a civil war or the banishment of any political parties.

b. Do you or Twitter believe it is appropriate for its platform or company to call for a “new civil war?”

As stated above, Mr. Dorsey has previously explained that he believes this article to be a compelling narrative of competing economic systems in the past, such as labor, and the present, primarily one based on energy.

c. Do you or Twitter believe it is appropriate for its platform or company to call for an end to one of the Nation’s two major political parties?

Mr. Dorsey has previously explained that he believes this article has lessons that apply across the political spectrum.

d. Do you or Twitter believe it is appropriate for its platform or company to call for an end to the conservative movement?

As stated above, Mr. Dorsey has previously explained that he believes this article has lessons that apply across the political spectrum.

U.S. Senate Committee on the Judiciary
Subcommittee on the Constitution

Responses to Questions for the Record for Mr. Carlos Monje, Jr.
Director of Public Policy and Philanthropy, U.S. & Canada

SENATOR MAZIE K. HIRONO

1. Last month, Facebook announced that it was banning white nationalist and white separatist content from its platform. Unfortunately, Twitter has not followed suit.

a. What is Twitter doing to prevent the spread of white nationalism on its platform?

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

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

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

In addition, we do not allow people who affiliate with organizations that – whether by their own statements or activity both on and off the platform – use or promote violence against civilians to further their causes to use our platform

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

primarily in abusive behavior, or is deemed to have shared a violent threat, we will permanently suspend the account upon initial review.

b. Why hasn't Twitter banned white nationalist and white separatist content?

We do not allow people who affiliate with organizations that – whether by their own statements or activity both on and off the platform – use or promote violence against civilians to further their causes to use our platform

One of the underlying features of our approach to enforcing our rules is that we take a behavior-first approach. That is to say, we look at how accounts behave before we look at the content they are posting. This is how we were able to scale our efforts to combat ISIS content far faster than we would have been if we had relied on experts reviewing the content produced by ISIS. As we developed a greater understanding of how ISIS-linked accounts operated, we could further invest in technology to identify accounts, ultimately allowing us to take action on a large number of accounts before they had even Tweeted.

Our response to the challenges of terrorism and violent extremism on our platform are a company-wide effort and not viewed in isolation from our wider mission to improve the health of the public conversation. We cannot solve the problems posed by violent extremist ideologies by removing content alone, particularly given the clear migration of these bad actors to smaller platforms who do not share their peers commitment to solving this problem.

2. With regard to Twitter's content moderation practices:

a. How many content moderators does Twitter employ worldwide? Please provide the total number content moderators along with a breakdown by country of residence, by state of residence (if country of residence is the United States), and by employment status (i.e., how many content moderators are Twitter employees v. contractors).

Twitter uses a combination of machine learning and human review to adjudicate abuse reports and whether they violate our rules. One of the underlying features of our approach is that we take a behavior-first approach. That is to say, we look at how accounts behave before we look at the content they are posting. This is how we were able to scale our efforts globally. Twitter employs extensive content detection technology to identify and police harmful and abusive content embedded in various forms of media on the platform. We currently employ approximately 1,500 full time

employees and contractors across the globe who are directly involved in enforcing our rules on Twitter, all of whom work in different ways to improve the health of the platform. We have made the health of Twitter our top priority, and our efforts will be measured by how we help encourage more healthy debate, conversations, and critical thinking on the platform. Conversely, abuse, automation, hateful conduct, terrorism, and manipulation will detract from the health of our platform.

b. Please describe the training provided to content moderators.

Twitter moderators are provided ongoing training on how to enforce our rules impartially. Notably, because 79 percent of our users are outside of the United States, much of the training involves being sensitive to cultural and language differences across the world. One of the underlying features of our approach to enforcing our rules is that we take a behavior-first approach. This behavior-first approach means that we look at how accounts behave before we look at the content they are posting. Because our service operates in dozens of languages and hundreds of cultural contexts around the globe, we have found that behavior is a strong signal that helps us identify bad faith actors on our platform. The behavioral ranking that Twitter utilizes does not consider in any way political views or ideology. It focuses solely on the behavior of all accounts.

c. What is the average salary of a content moderator?

We depend on highly skilled personnel to grow and operate our business. Our future success and strategy will depend upon our continued ability to identify, hire, develop, motivate and retain highly skilled personnel, including content moderators. To attract and retain highly skilled personnel, we have had to offer, and believe we will need to continue to offer, highly competitive compensation packages.

d. On average, how many hours per week does a content moderator work?

Hours are variable.

e. On average, how many tweets, retweets, likes, replies, etc. does a content moderator review per week?

In regard to the removal of accounts, our biannual Twitter Transparency Report highlights trends in enforcement of our Rules, legal requests, intellectual

property-related requests, and email privacy best practices. The report also provides insight into whether or not we take action on these requests. The Transparency Report includes information requests from governments worldwide and non-government legal requests we have received for account information. Removal requests are also included in the Transparency Report and include worldwide legal demands from governments and other authorized reporters, as well as reports based on local laws from trusted reporters and non-governmental organizations, to remove or withhold content.

f. On average, how much time does a content moderator have to determine if a tweet, retweet, like, reply, etc. violates the Twitter Rules and Terms of Service?

Workflows are variable, depending on the types of violations and how clear the context is. Moderators are assisted by automated tools in completing their work. According to our latest Twitter Transparency Report, we reported that 6,388 accounts were reported by known government entities compared to 5,461 reported during the last reported period, an increase of 17%. We have a global team that manages enforcement of our Rules with continuous coverage, in every supported language on the service. It is worth noting that the raw number of reported accounts is not a consistent indicator of the validity of the reports we receive. During our review process, we may consider whether reported content violates aspects of the Twitter Rules beyond what was initially reported. For example, content reported as a violation of our private information policy may also be a violation of our policies for hateful conduct. If the content is determined to violate any Twitter Rule, it is actioned accordingly. Not all reported accounts are found to violate the Twitter Rules, and reported accounts may be found to violate a different rule than was initially reported. We may also determine that reported content does not violate the rules at all. The volumes often fluctuate significantly based on world events, including elections, national and international media stories, and large conversational moments in social and political culture.

g. What percentage of content moderators have reported a diagnosis of or symptoms of post-traumatic stress disorder (PTSD), drug abuse, anxiety, and/or another psychological disorder as a result of their work?

We have a full suite of support services available for our employees, including content moderators. In addition to an increased investment in machine learning, our efforts to improve the health of the public conversation do include global content review teams made up of agency partners. These teams are sometimes

exposed to material that is sensitive and potentially distressing in nature. Our highest priority is to ensure they are treated with compassion, care, and respect. We are continually evaluating our partners' standards and remain committed to protecting the well-being of the teams tasked with this important and challenging role.

U.S. Senate Committee on the Judiciary
Subcommittee on the Constitution

Responses to Questions for the Record for Mr. Carlos Monje, Jr.
Director of Public Policy and Philanthropy, U.S. & Canada

SENATOR RICHARD BLUMENTHAL

1. In September, Twitter CEO, Jack Dorsey, committed at a House hearing to undergo a civil rights audit but has not provided any more information of its assessment or its status.

a. Who is conducting the audit, what is the scope of this audit, and what is the status of the audit?

We agree that third-party feedback and metrics can be valuable resources to inform our work. In 2018, we announced an open Request for Proposal asking academics and researchers to develop methods to measure the health of our service. As a result of our request for proposal, we partners with experts at Leiden University and other academic institutions to better measure the health of Twitter, focusing on informational echo chambers and unhealthy discourse on Twitter.

This collaboration will also enable us to study how exposure to a variety of perspectives and opinions serves to reduce overall prejudice and discrimination. While looking at political discussions, these projects do not focus on any particular ideological group and the outcomes will be published in full in due course for further discussion. We believe this work is aligned with the request for this external audit. For more information, please see: https://blog.twitter.com/official/en_us/topics/company/2018/measuring_healthy_conversation.html.

b. What changes has Twitter made specifically as a result of the audit?

Ensuring we have thoughtful, comprehensive metrics to measure the health of public conversation on Twitter is crucial to guiding our work and making progress, and our partners will help us continue to think critically and inclusively so we can get this right. We know this is a very ambitious task, and look forward to working with the team, challenging ourselves to better support a thriving, healthy public conversation.

c. What issues have been identified by the audit that Twitter plans to address?

Twitter continues to collaborate with the non-profit research center Cortico and the Massachusetts Institute of Technology Media Lab on exploring how to measure aspects of the health of the public sphere. As a starting point, Cortico proposed an initial set of health indicators for the United States (with the potential to expand to other nations), which are aligned with four principles of a healthy public sphere. Those include:

- **Shared Attention: Is there overlap in what we are talking about?**
- **Shared Reality: Are we using the same facts?**
- **Variety: Are we exposed to different opinions grounded in shared reality?**
- **Receptivity: Are we open, civil, and listening to different opinions?**

d. Will Twitter commit to making the outcomes of the audit, issues identified by the audit, and changes made as a result of the audit available to the public?

We believe transparency is critical and have committed to sharing the results of our work on a regular basis.

2. Tech companies occasionally remove voices that spread hatred, lies, and bigotry. Mr. Parker testified that people used the internet to "regurgitate demonstrably and undeniably false information about [the Sandy Hook shooting] while simultaneously attacking victims' families for profit." Those lies were a coordinated campaign of harassment, not a mere incident, and had real consequences.

a. What criteria is used by Twitter to assess content that provokes or facilitates the online and offline harassment of crime victims?

Twitter uses a combination of machine learning and human review to adjudicate abuse reports and whether they violate our rules. In order to maintain a safe environment for individuals on Twitter, we may suspend accounts that violate the Twitter Rules. Most of the accounts we suspend are suspended because they are automated or fake accounts, and they introduce security risks for Twitter and all of the individuals who use our service. These types of accounts are contrary to our Twitter Rules. When an account engages in abusive behavior and it has been reported to us as violating our Rules surrounding abuse, like sending threats to others, we may suspend it temporarily or, in some cases, permanently. Reducing the

possibility of offline harm is one of our guiding principles in developing and implementing our rules.

b. What staff and resources has Twitter made available to proactively monitor for and address the online and offline harassment of crime victims?

Twitter uses a combination of machine learning and human review to adjudicate abuse reports and whether they violate our rules. One of the underlying features of our approach is that we take a behavior-first approach. That is to say, we look at how accounts behave before we look at the content they are posting. This is how we were able to scale our efforts globally. Twitter employs extensive content detection technology to identify and police harmful and abusive content embedded in various forms of media on the platform. We currently employ approximately 1,500 full time employees and contractors across the globe who are directly involved in enforcing our rules on Twitter, all of whom work in different ways to improve the health of the platform. We have made the health of Twitter our top priority, and our efforts will be measured by how we help encourage more healthy debate, conversations, and critical thinking on the platform. Conversely, abuse, automation, hateful conduct, terrorism, and manipulation will detract from the health of our platform.

c. What are the specific steps that crime victims such as Mr. Parker should take in order to elevate their cases to Twitter to receive specialized assistance?

In the cases of violent threats, Twitter recommends that in addition to reporting the abusive content to the platform, the individual considers contacting local law enforcement and we provide a tool that allows the individual to generate an email with all of the relevant necessary information to submit a law enforcement report. Local law enforcement agencies can accurately assess the validity of the threat, investigate the source of the threat, and respond to concerns about physical safety. If Twitter is contacted by law enforcement directly, we can work with them and provide the necessary information for their investigation of the threat. We continuously deploy new technological tools to identify Tweets and accounts that violate our Terms of Service.

3. Twitter does not publish specific information about enforcement of its content moderation policies, such as how many pieces of content it takes down, what rules were violated, the demographics of people targeted, and rates of appeals. This is different from its current transparency reports, which include little substantive information in them about hateful or abusive activities against individuals. Facebook has provided such information in its transparency reports. These data are important to researchers and civil society organizations

trying to study the problems and target resources to affected communities including crime victims.

a. Does Twitter statistically track its enforcement of content moderation policies?

Yes. In regard to the removal of accounts, our biannual Twitter Transparency Report highlights trends in enforcement of our Rules, legal requests, intellectual property-related requests, and email privacy best practices. The report also provides
insight into whether or not we take action on these requests. The Transparency Report includes information requests from governments worldwide and non-government legal requests we have received for account information. Removal requests are also included in the Transparency Report and include worldwide legal demands from governments and other authorized reporters, as well as reports based on local laws from trusted reporters and non-governmental organizations, to remove or withhold content.

b. Will Twitter commit to including quantitative information on enforcement of content moderation policies?

Yes. According to our latest Twitter Transparency Report covering the second half of 2018, across the six Twitter Rules policy categories included, we received reports regarding 11,000,257 unique accounts were reported for possible violations of those Twitter Rules, amounting to a 19% increase compared to the prior reporting period. Of those accounts, 6,388 accounts were reported by known government entities compared to 5,461 reported during the last reported period, an increase of 17%. We have a global team that manages enforcement of our Rules with continuous coverage, in every supported language on the service.

It is worth noting that the raw number of reported accounts is not a consistent indicator of the validity of the reports we receive. During our review process, we may consider whether reported content violates aspects of the Twitter Rules beyond what was initially reported. For example, content reported as a violation of our private information policy may also be a violation of our policies for hateful conduct. If the content is determined to violate any Twitter Rule, it is actioned accordingly. Not all reported accounts are found to violate the Twitter Rules, and reported accounts may be found to violate a different rule than was initially reported. We may also determine that reported content does not violate the rules at all. The volumes often fluctuate significantly based on world events, including

elections, national and international media stories, and large conversational moments in social and political culture.

c. Please provide specific information on the following:

i. How many reports of hateful activities do you receive quarterly? How many of these initial reports are generated by automated systems vs users or other human flaggers?

ii. What percentage of these reports are found to violate your rules?

iii. Approximately how many violations of each category of abuse or rule occurred?

iv. What are the demographics of the users engaging in hateful activities?

v. What are the demographics of the users receiving hateful content?

vi. What are the demographics of the users reporting hateful activities?

vii. What percentage of hateful activities come from repeat offenders?

viii. What is the average review time of reported hateful activities?

ix. How many appeals of rules violations do you receive? What percentage of appeals are granted?

As reported in the latest Twitter Transparency Report covering the enforcement of hateful conduct reports from July through December 2018, we received 3,518,898 reported accounts. We took action on 250,806 individual accounts based on hateful conduct violations. For additional information on the enforcement of our Rules, please see: https://transparency.twitter.com/en/twitter-rules-enforcement.html.