United States House of Representatives
Permanent Select Committee on Intelligence
"Emerging Trends in Online Foreign Influence Operations:
Social Media, COVID-19, and Election Security"
June 18, 2020

Nick Pickles, Twitter
Responses to Questions for the Record

From Chairman Schiff

For all witnesses

1. In the course of removing assessed networks engaged in CIB or foreign influence operations, does your company have standing policy or guidance with respect to proactively informing users who engaged with those removed accounts or the content? Why or why not?

Following the Intelligence Community Assessment findings in January 2017, Twitter conducted a retrospective review of Russian activity that occurred on the platform affiliated with the 2016 U.S. election. As we reported in January 2018, we found datasets comprising 3,841 accounts affiliated with the Internet Research Agency (IRA) originating in Russia and 50,258 automated accounts that were Russian-linked and Tweeting election-related content.

We provided notice to people on our service in the United States who followed, liked, or Retweeted IRA content on Twitter during the 2016 election time period. In addition to providing notice to those individuals on Twitter who interacted with IRA content, beginning in October 2018, we published the first comprehensive archive of Tweets and media associated with suspected state-backed information operations on Twitter. These public disclosures included the Russian activity tied to the 2016 U.S. election.

Since then, we have provided seven additional updates covering the activities of a range of state-backed actors and published all associated Tweets to a publicly available database. To date, it is the only public archive of its kind. The archive now spans operations across 15 countries, including more than nine terabytes of media and 200 million Tweets. Using our archive, thousands of researchers have conducted their own investigations and shared their insights and independent analyses with the world.

2. Can you please describe your company's relationships or engagements with the national political parties, state parties, and individual campaigns, generally, and in the event you discovered a covert foreign influence operation targeting a specific candidate or political party?

a) Are these interactions regular, or would they depend on identification of a specific threat?

We are in active and regular dialogue with candidates and campaign committees on a wide range of issues, including account security. Candidates and campaign committees are able to raise any content or activity they identify as problematic.

b) If an individual candidate suspects they are being subjected to malign online activity, do they know who and how to contact at your company?

Yes. We are in active and regular dialogue with candidates and campaign committees. We also have well-established relationships with law enforcement agencies active in this arena, including the Federal Bureau of Investigation Foreign Influence Task Force and the U.S. Department of Homeland Security's Election Security Task Force.

3. We've seen China in particular engage in overt use of its official diplomatic accounts and state-controlled media to shape the information space online and promote misleading or false narratives that advance its state strategic interests in an identifiably coordinated manner. Beyond mere labeling of state-controlled media or identification of official foreign or diplomatic account as such:

a) Can you please describe your company's approach to fact-checking or adding context to misleading or outright disinformation posted by these overt, foreign linked accounts in a coordinated manner, which might allow users to readily understand the broader context or be directed to authoritative, credible sources about the claims?

In 2019, we consulted with the public on our approach and conducted public research to guide our work. That research told us that people want to know if they are viewing manipulated content and they support Twitter labeling it. We heard:

- Twitter should not determine the truthfulness of Tweets.
- Twitter should provide context to help people make up their own minds in cases where the substance of a Tweet is disputed.

Hence, our focus is on providing context, and not fact-checking. When we label Tweets, we link to Twitter conversation that shows three things for context: (1) factual statements; (2) counterpoint opinions and perspectives; and (3) ongoing public conversation around the issue. We will only add descriptive text that is reflective of the existing public conversation to let people determine their own viewpoints. To date, we have applied these labels to thousands of Tweets around the world, primarily related to COVID-19 and manipulated media. We have been transparent that we are not attempting to address all types of misinformation.

Additionally, we are actively exploring opportunities to add further context to state-controlled media. We also note Twitter remains the only platform to fully prohibit advertising from state controlled media outlets and those associated with them.

b) If a Facebook post, Tweet, or YouTube video created by a state-controlled media outlet promotes misleading or probably false narratives in apparent coordinated manner reasonably assessed to be in the service of that state's interests, what steps might your respective platforms consider in terms of labeling, fact-checking, or providing context to users about such material?

We coordinate closely with our peers and share information relating to potential bad actors. Where material is removed from one service, where appropriate that information would be shared and we would review under our own Terms of Service. In many cases, this has led to removal of content, as is also the case when Twitter has shared information with other companies.

The Twitter Rules apply equally to all accounts and where necessary we will remove or apply a label and link to additional context. This includes state media.

In addition, Twitter does not allow news media entities controlled by state authorities to advertise. This decision was initially taken with regard to Russia Today and Sputnik based on the Intelligence Community Assessment of Russian Activities during the 2016 election, a report published in January 2017. In August 2019, we expanded this policy to cover all state controlled media entities globally, in addition to individuals who are affiliated with these organizations. Under this policy, news media entities controlled by state authorities may not purchase advertisements. This policy extends to individuals reporting on behalf of or who are directly affiliated with such entities.

4. Graphika's June 16, 2010 report about the so-called "Secondary Infektion" group assessed it as having links to Russia and attempted to use false stories and outright forged materials to advance narratives favorable to Moscow.

a) Does your company have a policy governing the removal of "genuine," probably hacked or stolen materials found on your platform, similar to the episode involving the hacked-and-dumped emails of Clinton Campaign Chair John Podesta in 2016? If so, please provide it in writing.

We have seen that sophisticated threat actors, including state-backed hacking groups, engage in the distribution of illegitimately obtained documents and private communications to try to influence global civic discourse. We have a zero-tolerance policy for this sharing of hacked materials on Twitter — one of the key policy changes introduced since 2016.

According to the Twitter Rules, we do not permit the use of our services to directly distribute content obtained through hacking that contains personally identifiable information, may put people in imminent harm or danger, or contains trade secrets. Direct distribution of hacked materials includes posting hacked content on Twitter (for instance, in the text of a Tweet or in an image), or directly linking to hacked content hosted on other websites.

We also will take enforcement action on accounts that claim responsibility for a hack, which includes threats and public incentives to hack specific people and accounts. We also may permanently suspend accounts in which Twitter is able to reliably attribute a hack to the account distributing that content. Commentary about a hack or hacked materials, such as news articles discussing a hack, are generally not considered a violation of this policy. This includes, for example, journalistic and editorial discussion of hacking and disclosures of legitimate public concern and which pose no physical harm.

As we have seen in other policy areas, this issue is a challenge when members of the media distribute the contents of a hack through their own reporting. These actions potentially achieve the aim of the hostile actor to amplify a desired message to large audiences in spite of Twitter's efforts to remove offending accounts.

b) Does this policy include or account for the posting of suspected or proven forgeries that were presented as genuine and was linked to a foreign influence operation? Or would your company otherwise prevent the sharing or re-posting of such forged content?

Synthetic and manipulated media, some forms of which are commonly referred to as "deep fakes," represent an emerging threat to the integrity and trustworthiness of conversations on Twitter. We have closely tracked the challenges associated with these new technologies and have introduced new policies and product features to help combat them.

Our policy in this area was built in the open and based on feedback from the people we serve. On November 11, 2019, we released a draft of our rules governing synthetic and manipulated media that purposely attempts to mislead or confuse people. We opened a public feedback period to get input from the public, providing a brief survey available in English, Hindi, Arabic, Spanish, Portuguese, and Japanese. Ultimately, we gathered more than 6,500 responses from people around the world. We also consulted with a diverse, global group of civil society and academic experts on our draft approach.

On February 4, 2020, we announced Twitter's policy on synthetic and manipulated media. Under our Rules, an individual may not deceptively share synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand the media's authenticity and to provide additional context.

We review a number of criteria when evaluating Tweets and media for labeling or removal under this rule. First, we determine whether media have been significantly and deceptively altered or fabricated. Some factors we consider include: (1) whether the content has been substantially edited in a manner that fundamentally alters its composition, sequence, timing, or framing; (2) any visual or auditory information (such as new video frames, overdubbed audio, or modified subtitles) that has been added or removed; and (3) whether media depicting a real person has been fabricated or simulated.

Second, we evaluate whether the media are shared in a deceptive manner. Under this review, we also consider whether the context in which media are shared could result in confusion or misunderstanding or suggests a deliberate intent to deceive people about the nature or origin of the content, for example by falsely claiming that it depicts reality.

Lastly, we assess the context provided alongside media, for example by reviewing the text of the Tweet accompanying or within the media; the metadata associated with the media; the information on the profile of the person sharing the media; and websites linked in the profile of the person sharing the media, or in the Tweet sharing the media.

Under our policy, we also review whether content is likely to impact public safety or cause serious harm. Tweets that share synthetic and manipulated media are subject to removal under this policy if they are likely to cause harm. Some specific harms we consider include threats to the physical safety of a person or group, risk of mass violence or widespread civil unrest, and threats to the privacy or ability of a person or group to freely express themselves or participate in civic events.

c) Do these or other policies cover content that might otherwise be illicitly obtained, e.g. a phone conversation that was recorded by a third party without the knowledge or consent of the calling or the called party, and then posted to Facebook, Twitter, or YouTube?

We recognize that we are operating in an increasingly complex dynamic. These are constantly evolving challenges and we will keep our policies and approach under advisement, particularly as we learn more about this type of content.

5. What changes has your company made to algorithms deployed on its internet platforms since 2017, especially with respect to limiting the reach or potential virality of extremist content and conspiracy theories?

a) How do you measure your success?

Individuals are prohibited from making specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people. This includes, but is not limited to, threatening or promoting terrorism.

Twitter's philosophy is to take a behavior-led approach, utilizing a combination of machine learning and human review to prioritize reports and improve the health of the public conversation. That is to say, we increasingly look at how accounts behave before we look at the content they are posting. This is how we seek to scale our efforts globally and leverage technology even where the language used is highly context specific. Twitter employs extensive content detection technology to identify potentially abusive content on the service, along with allowing users to report content to us either as an individual or a bystander.

As a result of this approach, we have now suspended more than 1.6 million accounts for violations related to the promotion of terrorism between August 1, 2015, and June 30, 2019. In 2018, a total of 371,669 accounts were suspended for violations related to

promotion of terrorism. We continue to see more than 90 percent of these accounts suspended through proactive measures.

The trend we are observing year-over-year is a steady decrease in terrorist organizations attempting to use our service. This is due to zero-tolerance policy enforcement that has allowed us to take swift action on ban evaders and other identified forms of behavior used by terrorist entities and their affiliates. In the majority of cases, we take action at the account creation stage — before the account even Tweets.

Our industry collaboration is critical in our efforts to address terrorist use of the internet. Through the Global Forum to Counter Terrorism, we collaborate to tackle the real-time distribution of content during a real-world incident, most recently in response to the shooting in Glendale, Arizona.

With regard to wider issues, we have been clear that we will take strong enforcement action on behavior that has the potential to lead to offline harm. In line with this approach, we have recently taken further action on so-called "QAnon" activity across the service.

We will permanently suspend accounts Tweeting about these topics that we know are engaged in violations of our multi-account policy, coordinating abuse around individual victims, or are attempting to evade a previous suspension — something we have seen more of in recent weeks.

In addition, we will no longer serve content and accounts associated with QAnon in trends and recommendations. We will also work to ensure we are not highlighting this activity in search and conversations and block URLs associated with QAnon from being shared on Twitter.

We will continue to review this activity across our service and update our rules and enforcement approach again if necessary.

b) Would you make public metrics so that we in Congress can judge these issues in a non-anecdotal fashion?

Twitter is committed to the open exchange of information. First published on July 2, 2012, 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 publishes data on the number of accounts actioned for violating

certain categories of the Twitter Rules (along with all incorporated policies), Privacy Policy, and Terms of Service.

The trend we are observing year-over-year is a steady decrease in terrorist organizations attempting to use our service. This is due to zero-tolerance policy enforcement that has allowed us to take swift action on ban evaders and other identified forms of behavior used by terrorist entities and their affiliates. In the majority of cases, we take action at the account creation stage — before the account even Tweets.

For Twitter

1) Is Twitter considering adopting a labeling system so its users can easily distinguish state controlled media? Why or why not?

Twitter does not allow news media entities controlled by state authorities to advertise. This decision was initially taken with regard to Russia Today and Sputnik based on the Intelligence Community Assessment of Russian Activities during the 2016 election, a report published in January 2017.

In August 2019, we expanded this policy to cover all state controlled media entities globally, in addition to individuals who are affiliated with these organizations. Under this policy, news media entities controlled by state authorities may not purchase advertisements. This policy extends to individuals reporting on behalf of or who are directly affiliated with such entities.

We are actively exploring opportunities to add further context to state-controlled media.

2) Twitter stands out for its allowing automated accounts — "bots" — as an integral part of the platform.

a. How do known examples of foreign malign use of automated accounts manipulate Twitter's platform affect the company's calculus about whether to keep automation as it is, or to place stricter limits on how "bots" operate?

People often refer to "bots" when describing everything from automated account activity to individuals who would prefer to be anonymous for personal or safety reasons, or avoid using their own photo because they have privacy concerns.

A bot is an automated account — nothing more or less. We acknowledge that in prior years, malicious automated accounts were a problem for Twitter. We focused on it, made the investments, and have seen significant gains in tackling them across all surfaces of Twitter. But that does not mean our work is done.

Today, we focus on the holistic behavior of an account, not just whether it is automated or not. That is why calls for bot labeling do not capture the problem we are trying to solve and the errors we could make to real people that need our service to make their voice heard. It is not just a binary question of bot or not — the gradients in between are what matter.

Oftentimes, the term “bot” is used to mischaracterize accounts with numerical usernames that are auto-generated when your preference is taken, and more worryingly, as a tool by those in positions of political power to tarnish the views of people who may disagree with them or online public opinion that’s not favorable.

There are also many commercial services that purport to offer insights on bots and their activity online, and frequently their focus is entirely on Twitter due to the free data we provide through our public APIs. Unfortunately, this research is rarely peer-reviewed and often does not hold up to scrutiny, further confusing the public's understanding of what is really happening.

b. Has the company made any changes to its policies or enforcement on automated accounts specifically since 2017 to combat attempted foreign CIB?

We do not permit malicious use of automation to undermine and disrupt the public conversation, like trying to get something to trend. We do not allow artificial amplification of conversations on Twitter, including through creating multiple or overlapping accounts An individual on Twitter is not permitted to generate, solicit, or purchase fake engagements. We do not allow people to engage in bulk or aggressive tweeting, engaging, or following. We also do not allow individuals to use hashtags in a spammy way, including using unrelated hashtags in a tweet (aka "hashtag cramming").

Our technological power to proactively identify and remove these behaviors across our service is more sophisticated than ever. We permanently suspend millions of accounts every month that are automated or spammy, and we do this before they ever reach an eyeball in a Twitter Timeline or Search.

If an individual sees suspicious activity, we ask for it to be reported to us. We add that signal to the hundreds of others we use to inform our technical approach. If we want to create a healthy information ecosystem, we all have a part to play. We are keenly aware of our responsibility in this space. That includes protecting the integrity of our service, continuing to keep platform manipulation off of Twitter, and leading with transparency by sharing regular updates on our progress and learnings.

Under changes to our developer policy launched in March 2020, our new policy requests that developers clearly indicate in their account bio or profile if they are operating a bot account, what the account is, and the entity behind it is. This ensures people using Twitter can identify bot accounts. For additional information, see Twitter's developer blog.

c. Does Twitter's algorithm give the same weight to retweets or posts by automated accounts in determining trending content?

Tweets which our systems identify as being created through automation are not counted towards Trends.

The number of Tweets that are related to the trends is just one of the factors the algorithm looks at when ranking and determining trends. Algorithmically, trends and hashtags are grouped together if they are related to the same topic. For instance, #MondayMotivation and #MotivationMonday may both be represented by #MondayMotivation.

The Twitter rules also prohibit posting multiple updates in an attempt to manipulate or undermine Twitter trends, along with using or promoting third-party services or apps that claim to get you more followers, Retweets, or likes; or that claim to be able to get topics to trend.

More broadly, behavioral signals are an important factor in how Twitter protects the public conversation. 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. Twitter is always working to improve our behavior-based ranking models such that their breadth and accuracy will improve over time. We use thousands of behavioral signals in our behavior-based ranking models—this ensures that no one signal drives the ranking outcomes and protects against malicious attempts to manipulate our ranking systems.

From Representative Carson

For all witnesses

1. Can you provide a brief update on the policies that you companies currently use to address the threat deepfakes or other sophisticated manipulated media pose to users? What is your current approach, and how confident are you that you can identify and stop a foreign-connected deepfake as part of an attempted online influence operation?

Synthetic and manipulated media, some forms of which are commonly referred to as "deep fakes," represent an emerging threat to the integrity and trustworthiness of conversations on Twitter. We have closely tracked the challenges associated with these new technologies and have introduced new policies and product features to help combat them.

Our policy in this area was built in the open and based on feedback from the people we serve. On November 11, 2019, we released a draft of our rules governing synthetic and manipulated media that purposely attempts to mislead or confuse people. We opened a public feedback period to get input from the public, providing a brief survey available in English, Hindi, Arabic, Spanish, Portuguese, and Japanese. Ultimately, we gathered more than 6,500 responses from people around the world. We also consulted with a diverse, global group of civil society and academic experts on our draft approach.

On February 4, 2020, we announced Twitter's policy on synthetic and manipulated media. Under our Rules, an individual may not deceptively share synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand the media's authenticity and to provide additional context.

We review a number of criteria when evaluating Tweets and media for labeling or removal under this rule. First, we determine whether media have been significantly and deceptively altered or fabricated. Some factors we consider include: (1) whether the content has been substantially edited in a manner that fundamentally alters its composition, sequence, timing, or framing; (2) any visual or auditory information (such as new video frames, overdubbed audio, or modified subtitles) that has been added or removed; and (3) whether media depicting a real person has been fabricated or simulated.

Second, we evaluate whether the media are shared in a deceptive manner. Under this review, we also consider whether the context in which media are shared could result in

confusion or misunderstanding or suggests a deliberate intent to deceive people about the nature or origin of the content, for example by falsely claiming that it depicts reality.

Lastly, we assess the context provided alongside media, for example by reviewing the text of the Tweet accompanying or within the media; the metadata associated with the media; the information on the profile of the person sharing the media; and websites linked in the profile of the person sharing the media, or in the Tweet sharing the media.

Under our policy, we also review whether content is likely to impact public safety or cause serious harm. Tweets that share synthetic and manipulated media are subject to removal under this policy if they are likely to cause harm. Some specific harms we consider include threats to the physical safety of a person or group, risk of mass violence or widespread civil unrest, and threats to the privacy or ability of a person or group to freely express themselves or participate in civic events.

2. I know that there was reporting in December about accounts associated with the Epoch Times media outlet as having used faked profile photos on Facebook. Has Facebook, or the other companies, identified any new deployments of deepfakes in a fashion such as this, particularly if linked to a state actor?

Using publicly available services, such as "This Person Does Not Exist," it is relatively straightforward for people to obtain algorithmically-generated images of a "person". The use of these images does not by itself constitute highly sophisticated activity, and we have seen a wide range of actors — from commercial spammers to coordinated groups sharing political information — employ algorithmically generated profile photos to create false personas on Twitter.

Our policies on this behavior are clear: Individuals are not permitted to use Twitter in a manner intended to artificially amplify, suppress information, or engage in behavior that manipulates or disrupts other people's experience on the service. We do not allow spam or platform manipulation, such as bulk, aggressive, or deceptive activity that misleads others and disrupts their experience on Twitter. We also prohibit the creation or use of fake accounts. Some of the factors that we take into account when determining whether an account is fake include the use of stock or stolen avatar photos; the use of stolen or copied profile bios; and the use of intentionally misleading profile information, including profile location.

3. Throughout the recent protests in the wake of George Floyd's murder, some white nationalist groups have pushed messages of hate and violence, in an attempt to undermine

the legitimacy of the protest movement. One such white nationalist group, Identity Evropa, actually created a fake Twitter account, impersonated a left-wing Antifa activist, and explicitly called for violence during some of the most tense moments of the protests. With this example in mind:

a. How do your companies assess and evaluate any attempts by foreign actors to manipulate the information environment or create chaos during such fast-moving and emotional charged events? Especially when weighed against social media's role as an engine for legitimate civic organizing and the airing of genuine political or social grievances, as we've seen nationwide this month?

We are acutely aware of the risk in such fast-moving situations from actors, both foreign and domestic, attempting to manipulate the public conversation.

Our Site Integrity team is dedicated to identifying and investigating suspected platform manipulation on Twitter, including activity associated with coordinated malicious activity that we are able to reliably associate with state-affiliated actors. In partnership with teams across the company, we employ a range of open-source and proprietary signals and tools to identify when attempted coordinated manipulation may be taking place, as well as the actors responsible for it. We also partner closely with governments, law enforcement, academics, researchers, and our peer companies to improve our understanding of the actors involved in information operations and develop a holistic strategy for addressing them.

For example, we typically challenge 3 to 5 million accounts per week for these behaviors, requesting additional details, like email addresses and phone numbers in order to authenticate the account. We also recently acquired a new business to augment our efforts in this regard. This strategic investment will be a key driver as we work to protect the public conversation and help all individuals on our service see relevant information.

Attempts to execute misinformation campaigns rely on tactics like coordinated account manipulation or malicious automation — all of which are against Twitter's Rules. We are continuing to explore ways at how we may take action — through both policy and product — on these types of issues in the future. We continue to critically examine additional safeguards we can implement to protect the conversation occurring on Twitter.

b. Can your company provide an update on the procedures that it currently uses to identify content that incites violence? Are those processes automated, or how does

that process currently work? What definitions are used, since I imagine that it's not always clear-cut?

The Twitter rules make clear individuals are prohibited from making specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people. This includes, but is not limited to, threatening or promoting terrorism.

In December 2017, we broadened our rules to encompass accounts affiliated with violent extremist groups. Our prohibition on the use of Twitter's services by violent extremist groups, identified groups subscribing to the use of violence as a means to advance their cause, applies irrespective of the cause of the group.

People on Twitter are 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 toward others on the basis of these categories.

Twitter's philosophy is to take a behavior-led approach, utilizing a combination of machine learning and human review to prioritize reports and improve the health of the public conversation. That is to say, we increasingly look at how accounts behave before we look at the content they are posting. This is how we seek to scale our efforts globally and leverage technology even where the language used is highly context-specific. Twitter employs extensive content detection technology to identify potentially abusive content on the service, along with allowing people on our service to report content to us either as an individual or a bystander.

<u>From Representative Swalwell:</u>

For all witnesses

1. Do your platforms have a policy to combat anti-vaccine misinformation in posts by users? Does that policy extend beyond demonetization, if relevant? If so, how?

At Twitter, we understand the importance of vaccines in preventing illness and disease and recognize the role that Twitter plays in disseminating important public health information. We think it is important to help people find reliable information that enhances their health and well-being.

On Twitter, when someone searches for certain keywords associated with vaccines, a prompt will direct individuals to a credible public health resource. In the United States, we partnered with the U.S. Department of Health & Human Services and point people to vaccines.gov. The search prompt is available on iOS, Android, and mobile.twitter.com in the United States (in English and Spanish), Canada (in English and French), UK, Brazil, Korea, Japan, Indonesia, Singapore, and in Spanish-speaking Latin American countries. If an individual searches on twitter.com, they will see a pinned Tweet with information from trusted partners.

Additionally, we do not auto-suggest queries that are likely to direct individuals to non-credible commentary and information about vaccines.

These steps build on our existing work to guard against the artificial amplification of non-credible content about the safety and effectiveness of vaccines. We already ensure that advertising content does not contain misleading claims about the cure, treatment, diagnosis or prevention of certain diseases and conditions, including vaccines.

2. Do your platforms have a policy to combat public health misinformation in posts by users? Does that policy extend beyond demonetization, if relevant? If so, how?

The power of a uniquely open service during a public health emergency is clear. The speed and borderless nature of Twitter presents an extraordinary opportunity to communicate in real time and ensure people have access to the latest information from expert sources around the world. Journalists, experts, and engaged citizens Tweet side-by-side correcting and challenging public discourse second by second. These vital interactions happen on Twitter every day, and we are working to ensure that we surface the highest quality and most relevant content and context first.

In order to counter misinformation, especially in the context of COVID-19, we broadened our definition of harm to address content that goes directly against guidance from authoritative sources of global and local public health information. Rather than reports, we are enforcing this in close coordination with trusted partners, including public health authorities and governments, and continue to use and consult with information from those sources when reviewing content.

We prioritize removing content when it has a clear call to action that could directly pose a risk to people's health or well-being. We may also apply the public interest notice in cases where world leaders violate the COVID-19 guidelines.

Twitter has a zero-tolerance approach to the artificial amplification of public health misinformation content on our platform, as well as any attempt to abuse or manipulate our service. We continue to invest in detection tools and technology to combat malicious automation and manipulation of our service. This investment has yielded positive results – we have seen a 50 percent drop in challenges to accounts for suspected breaches of our platform manipulation policy.

We remain vigilant and will continue to substantially invest in our abilities to ensure that trends, search, and other common areas of the service are protected from malicious behaviors. As ever, we also welcome constructive and open information sharing from governments and academics to further our work in these areas.

3. Has One American News Network (OANN) had videos or posts removed from your platform? If so, how many and for what reasons?

Twitter has not removed videos or posts from the official account of One American News Network.

4. Has Fox News had videos or posts taken removed from your platform? If so, how many and for what reasons?

Twitter has not removed videos or posts from the official account of Fox News.

5. Has The Epoch Times had videos or posts removed from your platform? If so, how many and for what reasons?

In 2019, we suspended approximately 700 accounts originating from Vietnam for violating our rules around platform manipulation, specifically fake accounts and spam. These accounts were remotely linked to affiliates of *The Epoch Times* in Vietnam.

6. On June 18, 2020, Facebook, Instagram, and Twitter removed a Trump campaign ad featuring a symbol (a red inverted triangle) used by Nazis to designate political prisoners in concentration camps. Facebook, which owns Instagram, stated, "We removed these posts and ads for violating our policy against organized hate. Our policy prohibits using a banned hate group's symbol to identify political prisoners without the context that condemns or discusses the symbol."

a. How many symbols of hate would a campaign or candidate have to run before the campaign's account or page would be taken down from your platform?

Firstly, Twitter does not allow political advertisements, so a candidate would not be able to run such an ad in the first place.

With regard to organic content, people on Twitter are 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 toward 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 toward 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.

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. How many false or partly false posts, videos, or ads would a campaign or candidate have to run before the campaign or candidate's account or page would be taken down from your platform? Or would consistent posting of false or partly false posts or ads go unenforced?

Firstly, Twitter does not allow political advertisements, so a candidate would not be able to run such an ad in the first place.

With regard to organic content, people on Twitter are 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 toward others on the basis of these categories.

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 toward 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.

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.

c. Have campaign or candidate accounts, pages, or channels associated with U.S. persons been taken down because of repeated posting or advertising of false or partly false information? If so, how many? And if not, have you taken other actions against said accounts, pages, or channels?

Twitter is an public service , where there is free exchange of ideas, on diverse topics. In developing and enforcing our Rules for the service, we seek to be impartial, and as a service we believe in impartiality strongly. We take action on content that violates the Twitter Rules. Our rules are not based on ideology or a particular set of beliefs. Instead, the Twitter Rules are based on behavior. We will take action on any content that violates our Rules.

d. Have campaign or candidate accounts, pages, or channels associated with U.S. persons been taken down because of repeated use – whether through advertising or not – of symbols of hate and/or violating anti-hate policies? If so, how many? And if not, have you taken other actions against said accounts, pages, or channels? e. Are

your platforms considering implementing new policies or revising existing ones to address the issues raised in questions 7a through 7d?

As stated in response 6(a), 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.

From Rep. Maloney

For all witnesses

1. Recognizing that strides have been made since 2016 through 2018:

a) Is it your company's stance that the current volume and types of indicators, data, and/or metadata about potential foreign influence activity shared both within the industry and between the industry and the U.S. government are sufficient for protecting our national conversation and elections from foreign influence or interference moving forward?

We have well-established relationships with law enforcement agencies, and we look forward to continued cooperation with them on these issues, as often they have access to information critical to our joint efforts to stop bad faith actors. The threat we face requires extensive partnership and collaboration with our government partners and industry peers. We each possess information the other does not have, and our combined information is more powerful in combating these threats together. We have continuous coverage to address reports from law enforcement around the world and have a portal to swiftly handle law enforcement requests rendered by appropriate legal process. The challenges we face are complex, varied, and constantly evolving, and cooperation and information sharing are our strongest tools moving forward.

b) What limits imposed by U.S. law or regulations might prevent your company from maximally sharing data or metadata associated with high-confidence foreign influence operations/CIB with U.S. law enforcement?

We work closely with the Federal Bureau of Investigation, along with law enforcement and numerous public safety around the world. As our partnerships deepen, we are able to better respond to the changing threats we all face, sharing valuable information and promptly responding to valid legal requests for information. We continue to urge careful consideration of all materials that can be declassified to further support industry efforts to combat these threats.

c) How might relevant changes to the Secure Communications Act (SCA), the Electronic Communications Privacy Act (ECPA), Cybersecurity Information Sharing Act (CISA), or the Section 230 Communications Decency Act (CDA) help or harm your companies' efforts to prevent foreign influence from infiltrating your platforms?

As policymakers and experts examine policies around content moderation, it is critical to examine the existing regulatory framework. Of particular importance is Section 230 of the Communications Decency Act (CDA § 230). When it enacted CDA § 230 more than twenty years ago as part of the Telecommunications Act of 1996, Congress made the judgment that companies like Twitter that host content provided by others should have the latitude to make editorial decisions without becoming legally responsible for that content. CDA § 230 is a foundational law that has enabled American leadership in the tech sector worldwide.

It is the protection that allows us to proactively moderate content around activities such as state-backed foreign influence, child sexual exploitation, terrorism, voter suppression, and illicit drug sales. Without these tools, platforms would either cease to moderate content, including content that could relate to offline harm, or over-moderate content, resulting in less speech. Eroding CDA § 230 creates risks of liability for companies that make good-faith efforts to moderate bad faith actors and could result in greater restrictions around free expression.

d) Would considerations such as creating a "safe harbor" provision, or clearly delineating that assessed foreign influence actors don't have claim to the same data privacy protections as genuine users, affect those stances?

While Twitter has taken the unique step of publishing an archive of information operations linked to state actors, we believe that there is still need to improve the legal framework to provide industry with greater confidence to take such steps and to expand our efforts. We would also urge policymakers in the U.S. to engage with peers in Europe to consider the impact of the General Data Protection Regulation on these efforts.

2. Would your company find valuable an Information Sharing and Analysis Center (ISAC) or equivalent formalized mechanism devoted specifically to data-sharing about potential foreign-linked influence operations? Would your company support a leading role in an ISAC or equivalent? Why or why not?

We currently have well-established relationships with law enforcement agencies active in this arena, including the Federal Bureau of Investigation Foreign Influence Task Force and the U.S. Department of Homeland Security's Election Security Task Force. We look forward to continued cooperation with federal, state, and local government agencies on election integrity issues because in certain circumstances, only they have access to information critical to our joint efforts to stop bad faith actors.