August 31, 2020

Twitter, Inc.
1355 Market St. #900
San Francisco, CA 94103

Dear Senator Lee,

Thank you for your letter dated July 30, 2020, regarding Twitter's policy enforcement and content moderation efforts.

Twitter's purpose is to serve the public conversation. We welcome perspectives and insights from diverse sources and embrace being a service where the open and free exchange of ideas can occur.

We serve our global audience by focusing on the needs of the people who use our service, and we put them first in every step we take. We have rules in place that are designed to ensure the safety and security of the people who come to our service. Safety and free expression go hand in hand, both online and in the real world. If people do not feel safe to speak, they very often will not.

These two guideposts, free expression for all perspectives and rules of the road to promote safety, are not only in the interests of the people who use Twitter, but also paramount to sustaining our business.

Twitter does not use political viewpoints, perspectives, ideology or party affiliation to make any decisions, whether related to automatically ranking content or how we enforce our rules.

Please see our answers below to the questions you posed.

1. **What content-moderation standards do you employ when you remove content from your platform, where the content does not violate state or federal laws. Specifically, please explain your standards for removing content related to:**

    a. **COVID-19**

To address the global pandemic, on March 16, 2020, we announced new enforcement guidance, broadening our definition of harm. This guidance addresses content related to COVID-19 that goes directly against health information from domestic and global authoritative sources. We require individuals to remove violative Tweets in a variety of contexts with the goal of preventing offline harm.

Our new enforcement guidelines do not permit individuals to deny global or local health authority recommendations to decrease

someone's likelihood of exposure to COVID-19 or to Tweet content with the intent to influence people into acting against recommended guidance.

Individuals are also not permitted to describe treatments or protective measures which are not immediately harmful but are known to be ineffective, are not applicable to the COVID-19 context, or are being shared with the intent to mislead others, even if made in jest.

We also do not allow descriptions of harmful treatments or protection measures which are known to be ineffective, do not apply to COVID-19, or are being shared out of context to mislead people, even if made in jest. It is also a violation of our new enforcement guidelines to deny established scientific facts about transmission during the incubation period or transmission guidance from global and local health authorities.

People cannot make specific claims around COVID-19 information that intend to manipulate people into certain behavior for the gain of a third party with a call to action within the claim. The new enforcement guidelines do not allow specific and unverified claims that incite people to action and cause widespread panic, social unrest or large-scale disorder.

Specific and unverified claims made by people impersonating a government or health official or organization (such as a parody account of a government health official stating that the country's quarantine is over) are also not allowed under our Rules.

Individuals are not permitted to propagate false or misleading information around COVID-19 diagnostic criteria or procedures. We also do not allow false or misleading claims on how to differentiate between COVID-19 and a different disease, and if that information attempts to definitively diagnose someone. We also will not permit claims that specific groups or nationalities are never susceptible to COVID-19 or claims that specific groups or nationalities are more susceptible to COVID-19. For more information about our enforcement guidance on this topic, please see here.

b. Violent riots, and how you distinguish them from peaceful protests

Twitter prohibits individuals to use Twitter to make violent threats. We define violent threats as statements of an intent to kill or inflict serious physical harm on a specific person or group of people. Under this policy, an individual cannot state an intention to inflict violence on a specific person or group of people.

Violations of this policy include, but are not limited to threatening to kill someone; threatening to sexually assault someone; threatening to seriously hurt someone and/or commit a other violent act that could lead to someone's death or serious physical injury; and asking for or offering a financial reward in exchange for inflicting violence on a specific person or group of people.

This policy is enforced in tandem with our policies on abusive behavior and hateful conduct. Statements that express a wish or hope that someone experiences physical harm, making vague or indirect threats, or threatening actions that are unlikely to cause serious or lasting injury are not actionable under this policy, but may be reviewed and actioned under those policies. We also have a policy prohibiting the glorification of violence.

c. Hateful conduct

Twitter has policies prohibiting hateful conduct on the service as our Rules focus on behavior. 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.

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.

d. Protections for the unborn

Twitter does not have policies prohibiting people's right to discuss protections for the unborn. However, Twitter's advertising policies prohibit cause-based ads, which often encompasses content around this issue across all sides of the debate.

e. Misinformation

Twitter continues its zero-tolerance approach to platform manipulation. Individuals are not permitted to use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behavior that manipulates or disrupts people's experience on Twitter.

Platform manipulation can take many forms, including inauthentic engagements that attempt to make accounts or content appear more popular or active than they are; and coordinated activity that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation, and scripting. We have substantially invested in our proactive abilities to ensure Trends, Search, and other common areas of the service are protected from malicious behaviors.

f. Terrorist influence

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. A person on Twitter also may not affiliate with organizations that — whether by their own statements or activity both on and off the platform — use or promote violence against civilians to further their causes.

We also do not tolerate groups or individuals associated with them who engage in and promote violence against civilians both on and off the platform. Accounts affiliated with groups in which violence is a component of advancing their cause risk having a chilling effect on opponents and bystanders. The violence that such groups promote could also have dangerous consequences offline, jeopardizing their targets' physical safety.

We prohibit the use of Twitter's services by violent extremist groups. We consider violent extremist groups to be those which identify through their stated purpose, publications, or actions, as an extremist group;

have engaged in, or currently engage in, violence (and/or the promotion of violence) as a means to further their cause; and target civilians in their acts (and/or promotion) of violence. An individual on Twitter may not 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. From July through December 2019, action was taken on 86,799 unique accounts for violating our policies prohibiting terrorism and violent extremism. Seventy-four percent of those accounts were proactively identified and actioned.

2. **How did you formulate standards for the above categories of content? What sources did you look to in forming your content policies?**

We have an internal team dedicated to this work that brings a wide range of perspectives and expertise to developing our policies. 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. We also gather external feedback from a variety of partners -- including government, civil society, and nonprofits -- to inform our work.

3. **What are the prerequisites for a content-moderator position at your company? Do you inquire about the political or other beliefs of a candidate before making a hiring decision? Where you use contractors to serve in these roles, how do you ensure that they follow your internal guidelines and standards?**

Twitter does not use political ideology as a factor when hiring 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.

4. **What is the internal process that your content moderators follow to remove content that violates your standards?**

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. We do not take ideology into account when making content decisions. 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.

5. **How do you ensure that a content-moderation decision is not influenced by the personal beliefs or political views of the moderator?**

As noted above, our policies and enforcement are focused on behavior. 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. We believe strongly in being impartial, and we strive to enforce our rules dispassionately. We work to make sure our algorithms are fair and endeavor to be transparent and fix issues when we make mistakes.

We also have a process in which people can appeal content decisions. You can find more information about this here.

6. **If your content-moderation standards rely on guidance from a government entity, please explain your policy for allowing on your platform speech that disagrees with the government. If CDC guidance is the basis for removing content regarding COVID-19, how is that standard applied consistently? For example, since the CDC says that it is safe for schools to open, do you remove content from your platform that claims it is unsafe to reopen schools?**

We are closely working with a wide-range of trusted partners across the globe, including public health authorities, organizations, and government agencies to inform our approach. We will continue to review the Twitter Rules in the context of COVID-19 and are considering ways in which they may need to evolve to account for new behaviors. We are keeping our enforcement guidance under close review and are taking into account the views of medical professionals on any updates we may need to make in the future.

We will continue to prioritize removing content when it has a clear call to action that could directly pose a risk to people's health or well-being, but we want to make it clear that we do not take enforcement action on every Tweet that contains incomplete or disputed information about COVID-19. Our policies are not meant to limit good faith discussion or expressing hope about ongoing studies related to potential medical interventions that show promise.

7. **Where do you clearly articulate your content-moderation standards? How do you convey your moderation standards**

to consumers? Do you regularly update your users on changes made to your policies?

Individuals who create Twitter accounts must abide by our Twitter Terms of Service, Twitter Privacy Policy, and Twitter Rules. We make these publicly available on our website and utilize the Twitter Blog and Twitter owned and operate handles to communicate updates.

8. **Are your users required to provide knowing consent to the standard before giving you their personal information, data, and content, which gives value to your platform?**

Twitter believes individuals should know, and have meaningful control over, what data is being collected about them, how it is used, and when it is shared. Twitter is always working to improve transparency into what data is collected and how it is used. Twitter designs its services so that individuals can control the personal data that is shared through our services. People who use our services have tools to help them control their data. For example, if an individual has registered an account, through their account settings they can access, correct, delete or modify the personal data associated with their account.

Twitter recently updated our Privacy Policy to include callouts, graphics, and animations designed to enable people to better understand the data we receive, how it is used, and when it is shared.

Twitter also provides a toolset called Your Twitter Data. Your Twitter Data tools provide individuals accessible insights into the type of data stored by Twitter, such as username, email address, and phone numbers associated with the account and account creation details. The birthdays and locations of individuals are also shown in the tool if they have previously been provided to Twitter.

Individuals using the Your Twitter Data tool can also see and modify certain information that Twitter has inferred about the account and device such as gender, age range, languages, and interests. People on Twitter can review inference information, advertisers who have included them in tailored audiences, and demographic and interest data from external advertising partners. The Your Twitter Data tool also allows people with a Twitter account to download a copy of their relevant data from Twitter. We recently updated the download feature of the Your Twitter Data tool to include additional information. There is a version of this tool available to individuals who do not have a Twitter account, or for those logged out of the account.

9. **Do you coordinate your content moderation standards with other online platforms or competitors? Have you ever**

discussed or reached an agreement regarding these standards -- or the removal of content generally -- with any other online platform or competitor? If so, please [sic] each discussion or agreement, its subject, and the parties thereto.

Twitter develops its own content moderation standards. As a matter of public policy, we believe there is value in individual companies maintaining unique content moderation standards as a matter of consumer choice and competition within our industry.

10. **Do you coordinate the removal of specific content with other online platforms or competitors? If so, explain the process and what content has been subject to coordinate removal.**

We voluntarily collaborate with industry peers and civil society in five critical areas: terrorism, child sexual exploitation, election security, state-backed information operations, and COVID-19 misinformation.

In all of these areas, we use the signals from industry collaboration to inform and support our efforts, but we do not remove content automatically solely on the basis a peer company has done so.

In June 2017, we launched the Global Internet Forum to Counter Terrorism (the “GIFCT”), a partnership among Twitter, YouTube, Facebook, and Microsoft. The GIFCT facilitates, among other things: information sharing; technical cooperation; and, research collaboration, including with academic institutions. Technological collaboration is a key part of GIFCT’s work. In the first two years of GIFCT, two projects have provided technical resources to support the work of members and smaller companies to remove terrorist content.

Because Twitter does not allow files other than photos or short videos to be uploaded, one of the behaviors we saw from those seeking to promote terrorism was to post links to other services where people could access files, longer videos, PDFs, and other materials. Our pilot system allows us to alert other companies when we remove an account or Tweet that links to material that promotes terrorism hosted on their service. This information sharing ensures the hosting companies can monitor and track similar behavior, taking enforcement action pursuant with their individual policies. This is not a high-tech approach, but it is simple and effective, recognizing the resource constraints of smaller companies.

Regarding content depicting or promoting child sexual exploitation, the material is removed without further notice and reported to the National

Center for Missing & Exploited Children (NCMEC). We participate in NCMEC's hash sharing database for industry and non-governmental organizations which consists of image and video hashes of known child sexual abuse material.

When appropriate, we work with law enforcement and numerous public safety authorities around the world to combat child sexual exploitation material. We look forward to continued cooperation with them on this important issue. We also partner with nonprofits dedicated to child protection across the globe. In addition to our important relationship with NCMEC, Twitter is an active member of the Technology Coalition. This industry-led non-profit organization strives to eradicate child sexual exploitation by mentoring emerging or established companies, sharing trends and best-practices across industry, and facilitating technological solutions across the ecosystem.

The Technology Coalition serves as an effective model because it gives companies the flexibility to create, test, and iterate across our diverse products and models. As this threat changes and evolves, our playbook must also change and evolve in order to be effective. The flexibility of the model advanced by the Technology Coalition enables rapid evolution of best practices.

Additionally, a number of technology companies — including Twitter — established a dedicated, formal communications channel to facilitate real-time information sharing regarding election integrity. We also share information with our industry peers when we identify accounts acting in coordination in state-backed information operation campaigns.

At the onset of the recent COVID-19 pandemic, the White House requested that the technology industry come together to combat misinformation about the virus. In March 2020, we announced an informal industry effort to share information about this emerging threat.

11. **Some of you have removed or threatened to remove the ability of third parties to monetize their content through your advertising platform on the basis of content found in the third party's comments section. What is your policy or standard for such demonetization? Do you believe the same standard should be applied to platforms currently protected by Section 230?**

Individuals cannot monetize Tweets posted organically on Twitter. Twitter does not currently provide a service to monetize organic Tweets.

* * *

We strive to be as transparent as possible as we know it is key to earning and maintaining trust with the people who use our service. On August 19, we reimagined and rebuilt our biannual Twitter Transparency Report to become a comprehensive Twitter Transparency Center. Our goal with this evolution is make our content moderation practices more easily understood and accessible to the general public. We will continue to make investments and improvements in this area.

We welcome the opportunity to engage with you and your staff on these important topics. Please reach out if you would like further background, and we would also be pleased to set up a follow-up meeting with you and your staff.

Sincerely,

Lauren Culbertson
Head of U.S. Federal Policy
Twitter