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

Testimony of Nick Pickles
Twitter, Inc.

Chairman Schiff, Ranking Member Nunes, and Members of the Committee:

Thank you for the opportunity to appear.

The purpose of Twitter is to serve the public conversation and that conversation is never more important than during elections and civic events, the cornerstone of democracies across the globe.

Our service gives people the ability to share what is happening and provides people insights into a diversity of perspectives on critical issues; all in real time. We are humbled by the way our platform is used by those seeking to speak out against injustice, to hold those in power accountable, and to build movements for change.

Twitter is committed to furthering the health, openness, and civility of the public conversation on our service. We measure our success in these areas by how we help encourage more healthy debate, conversations, and critical thinking. Abuse, malicious automation, and platform manipulation detract from this goal and undermine our success.

As America has responded to the death of George Floyd and cities across the country have seen protestors take to the streets, the public conversation on Twitter has highlighted the deep-rooted nature of issues related to race, justice, and equality. While we have not seen evidence of concerted foreign state-backed efforts to manipulate the public conversation in recent weeks, we remain vigilant.

The threat of interference in elections from foreign and domestic actors is real and evolving. Since 2016, we have made significant investments to study and address these threats, taking lessons from the 2018 U.S. midterms, as well as elections around the world. I am grateful for the opportunity to discuss the policies, product changes, and partnerships Twitter has put in place to better protect the public conversation.

I. RELEVANT TWITTER POLICIES

A. *Civic Integrity Policy*

Twitter has a responsibility to protect the integrity of conversations involving elections and civic processes. Since 2016, we have made a sustained investment in expanding and clarifying our policies on this subject based on what we have learned over time.

Twitter has a clear policy prohibiting the use of our service for the purpose of manipulation or interfering in elections. Last month, we expanded this policy to include other forms of global civic engagement, including critical activities such as the Census.

We do not allow the posting or sharing of content that may suppress participation or mislead people about when, where, or how to participate in a civic process. We prohibit attempts to use our service to manipulate or disrupt civic engagement, including through the distribution of false or misleading information about the procedures or circumstances around participation in civic processes.

We prohibit individuals from sharing false or misleading information about how to participate in an election or other civic process in three categories:

1. Misleading information about procedures on how to participate in a civic process;
2. Misleading information about requirements for participation, including identification or citizenship requirements; or
3. Misleading statements or information about the official, announced date or time of a civic process.

We also do not permit individuals to share false or misleading information intended to intimidate or dissuade people from participating in an election or other civic process. This includes but is not limited to: misleading claims that polling places are closed, that voting has ended, or other misleading information relating to votes not being counted; misleading claims about police or law enforcement activity related to voting in an election, polling places, or collecting census information; misleading claims about long lines, equipment problems, or other disruptions at voting locations during election periods; misleading claims about process procedures or techniques which could dissuade people from participating; and threats regarding voting locations or other key places or events, among others.

Finally, we do not allow individuals to create fake accounts which misrepresent their affiliation, or share content that falsely represents its affiliation, to a candidate, elected official, political party, electoral authority, or government entity.

B. *Platform Manipulation Policy*

As platform manipulation tactics evolve, we are continuously updating and expanding our rules to better reflect what types of inauthentic activity violate our guidelines. We continue to develop and acquire sophisticated detection tools and systems to combat malicious automation on our service.

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.

We prioritize identifying suspicious account activity, such as exceptionally high-volume Tweeting with the same hashtag or mentioning the same @handle without a reply from the account being addressed. When we identify such activity, we require an individual using the service to confirm human control of the account or their identity.

We have increased our use of challenges intended to catch automated accounts, such as reCAPTCHAs (that require individuals to identify portions of an image or type words displayed on screen), and password reset requests that protect potentially compromised accounts. In the first six months of 2019, we challenged more than 97 million accounts which showed signs of engaging in some form of platform manipulation. We have also implemented mandatory email or phone verification for all new accounts.

C. *Rules Prohibiting Attributed Activity*

We know that certain groups and individuals engage in persistent, organized efforts to manipulate and interfere with the conversation on Twitter. Therefore, when we are able to reliably attribute an account on Twitter to an entity known to violate the Twitter Rules, we will remove additional accounts associated with that entity. For instance, if we are able to identify activity associated with the Russian Internet Research Agency, all accounts tied to that entity will be removed, regardless of the content they share. We likewise will remove accounts that deliberately mimic or are intended to replace accounts we have previously suspended for violating our rules. These steps allow us to take more aggressive action against known malicious actors.

D. *Distribution of Hacked Materials Policy*

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 behavior on Twitter — one of the key 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.

E. *Political Advertising Policy*

We believe political message reach should be earned, not bought. On October 30, 2019, Twitter's chief executive officer Jack Dorsey announced the decision to stop all political advertising on Twitter globally.

While Internet advertising is incredibly powerful and effective for commercial advertisers, that power brings significant risks to politics, where it can be used to influence votes to affect the lives of millions.

Online political advertising presents entirely new challenges to civic discourse that today's democratic infrastructure may not be prepared to handle, particularly the machine learning-based optimization of messaging and microtargeting.

F. *State Media Advertising Policy*

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.

II. PRODUCTS THAT SAFEGUARD THE CONVERSATION

A. *Approach to Misinformation*

At Twitter, we prioritize healthy public conversation through our product, policies, and enforcement. The health principles that guide our work include decreasing potential for likely harm; harmful bias and incentives; and reliance on content removal. Our principles also push us to increase diverse perspectives and public accountability. These principles connect to everything for us — from our decision to ban all political ads, to our policy around public-interest notices, and even a product test that allows people to choose who can reply to their Tweets.

These principles also shape our work on misleading information. In this area, too, we are using feedback from the people on our service. In 2019, we consulted with the public on our approach and that has guided our work since. Our initial review shows 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.

We are not attempting to address all misinformation. We are focused on where we can make the biggest impact and add context in a way that dovetails with the fundamental nature of our service: open, real-time, and conversational.

We prioritize based on the highest potential for harm, focusing on manipulated media, civic integrity, and COVID-19. Likelihood, severity, and type of potential harm — along with reach and scale — factor into this. Due to the large potential reach and persuasive impact of media content, we started with a policy on manipulated media. We have since expanded to issues of civic integrity and COVID-19 given the critical importance of elections and the global pandemic.

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.

B. *Synthetic and Manipulated Media*

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. *State-Backed Information Operations*

Combatting attempts to interfere in conversations on Twitter remains a top priority for the company, and we continue to invest heavily in our detection, disruption, and transparency efforts related to state-backed information operations. Our goal is to remove bad faith actors and to advance public understanding of these critical topics.

Twitter defines state-backed information operations as coordinated platform manipulation efforts that can be attributed with a high degree of confidence to state-affiliated actors. State-backed information operations are typically associated with misleading, deceptive, and spammy behavior. These behaviors differentiate coordinated manipulative behavior from legitimate speech on behalf of individuals and political parties.

Whenever we identify inauthentic activity on Twitter that meets our definition of an information operation, and which we are able to confidently attribute to actors associated with a government, we share comprehensive data about this activity.

In October 2018, we published the first comprehensive archive of Tweets and media associated with suspected state-backed information operations on Twitter and since then we have provided seven additional updates covering a range of state-backed actors. 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.

By making this data open and accessible, we empower researchers, journalists, governments, and members of the public to deepen their understanding of critical issues impacting the integrity of public conversations online. This transparency is core to Twitter's mission.

III. PARTNERSHIPS WITH KEY STAKEHOLDERS

Information sharing and collaboration are critical to Twitter's success in preventing hostile foreign actors from disrupting meaningful political conversations on the service. We 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.

On Election Day for the 2018 U.S. midterms, Twitter virtually participated in an operations center convened by the U.S. Department of Homeland Security. The operations center also convened officials from the U.S. Department of Justice, the Federal Bureau of Investigation, and the Office of the Director of National Intelligence, in addition to federal, state, local, and private sector partners. In the lead up to Election Day, and throughout the course of the day itself, Twitter remained in constant contact with officials throughout all levels of government. We plan to participate in a similar operations center in November 2020.

We also worked in close collaboration with the National Association of Secretaries of State (NASS) and the National Association of State Election Directors (NASED). Founded in 1904, NASS is the nation's oldest, nonpartisan professional organization for public officials, and is open to secretaries of states and lieutenant governors in the 50 states, D.C. and territories. We also partner with Civic Alliance, Vote Early Day and National Voter Registration Day to amplify credible election-related content.

Finally, we have significantly deepened our partnership with industry peers, including Facebook and Google, establishing formal processes for information sharing and a regular cadence of discussion about shared threats. We routinely collaborate to prepare for the upcoming U.S. election, building on our engagement in prior elections around the world. These collaborations are a powerful way to identify and mitigate malicious activity that is not restricted to a single platform or service.

IV. CONCLUSION

The threat of foreign and domestic interference is one that has evolved since 2016, with new tactics, tools, and vulnerabilities.

The issue is a broad geopolitical challenge, not one solely of content moderation. Removal of content alone will not address this challenge and while it does play an important role, particularly in tackling platform manipulation, governments must play a part in addressing the broader issue. While policy proposals may differ, it is clear that this is not the time to curtail online public conversation and the values that underpin the open Internet.

Like all Internet services, we are a constant target of opportunistic, spam-like activity. Some of this may be captured in "bot" analyses. While this activity may violate our rules, it generally is not indicative of larger, coordinated efforts — and especially, efforts backed by nation-state actors. Some of these actors may be commercially motivated, while others recycle existing narratives to fit current events to try to build their audience.

Indeed, the growth in the number of private sector actors claiming to be able to identify the origins of information operations risks unwittingly making the problem worse. As the European Union (EU) Disinformation Lab warned recently, "attributing disinformation without sufficient evidence, or with a restricted lens, can cause more harm than good." We continue to see attribution based on limited public information from actors who appear motivated more by their own press coverage than rigorous methodology.[1]

Second, and to a more limited extent, we continue to see inauthentic behavior originating from domestic individuals and groups on Twitter. This activity is not new — in 2018 we saw memes originating domestically, attempting to spread incorrect voting information, often following off-platform discussion and coordination. We removed nearly 6,000 Tweets that violated our policies during the 2018 U.S. midterms, primarily originating within the United States.

Furthermore, we recently suspended a series of fake accounts operating outside the U.S. that our team uncovered proactively purporting to be Antifa groups, but which were found to be linked to the white supremacist group Identity Evropa (also referred to as the American Identity Movement). These accounts were primarily engaged in hateful conduct, focused on issues of race, religion, and sexual orientation. We suspended the accounts and, as we regularly do for identified inauthentic behavior, immediately shared information with industry peers to enable their investigations.

[1] EU Disinformation Lab, *Being cautious with attribution: Foreign interference & COVID-19 disinformation,* April 10, 2020 (online at: https://www.disinfo.eu/publications/being-cautious-with-attribution-foreign-interference-covid-19-disinformation).

As we have invested in better defensive mechanisms, hostile actors have changed their behavior, either to amplify existing domestic content or increasingly to focus on vulnerabilities in the wider information ecosystem in the hope that domestic audiences will then distribute and amplify their message on social platforms.

Reports have also suggested individuals are being paid to mask the identity of who is behind activity — both foreign[2] and domestic[3]. The monetization of misinformation risks further obscure the commercially-motivated domestic actors from foreign-supported ones, highlighting the need for a broad approach to tackling this issue.

While Twitter has taken steps to remove paid political advertising, the wider risk of online advertising is being exploited by hostile actors, either directly or indirectly through proxies. It emphasizes the need for a root-and-branch risk assessment of the vulnerabilities of modern political financing and the different avenues foreign actors can use to influence domestic opinion.

More broadly, the way that official government accounts and state-controlled media engage in U.S.-focused discussion has evolved and the geopolitical debate surrounding COVID-19 has made this change clear.

Taken together, these are a diverse range of challenges and we continue to emphasize the need for a whole-of-society response. Twitter has a central role to play in that response and we take our responsibilities seriously, but a wide range of stakeholders need to play their part, too. Twitter offers a unique forum to engage in the global geopolitical conversation, challenge narratives, and debate policy choices.

Now is the time to unleash the best of democratic values and the opening of courageous communication in every sense. A well-informed public at home is still the strongest response to censorship abroad.

[2] CNN, *Russian election meddling is back -- via Ghana and Nigeria -- and in your feeds,* April 11, 2020 (online at: https://www.cnn.com/2020/03/12/world/russia-ghana-troll-farms-2020-ward/index.html).

[3] Buzzfeed News, People Are Renting Out Their Facebook Accounts In Exchange For Cash And Free Laptops, January 18, 2019 (online at: https://www.buzzfeednews.com/article/craigsilverman/facebook-account-rental-ad-laundering-scam).

The people who use Twitter must have confidence in the integrity of the service we provide: a place for participatory, open, and democratic public conversation. This is particularly critical with respect to information relevant to elections and the democratic process. From #BlackLivesMatter to Hong Kong protests and the upcoming U.S. elections, at its best Twitter builds movements, drives social change, and holds powerful voices to account. We continue to invest in our efforts to address threats that attempt to undermine this potential.

We have never been more focused on taking enforcement action on hostile actors and fostering an environment conducive to healthy, meaningful dialogue, debate, and connection on our service. We look forward to working with the Committee on these vital issues.