April 2020

Committee Secretary
Department of the Senate
PO Box 6100
Canberra ACT 2600

By email: foreigninterference.sen@aph.gov.au

Dear Secretariat,

Thank you for the opportunity to participate in the consultation process led by the Select Committee on Foreign Interference through Social Media to inquire into and report on the risk posed to Australia's democracy by foreign interference through social media.

The purpose of Twitter is to serve the public conversation. 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. People from around the world come to Twitter to engage in a free exchange of ideas. We must be a trusted and healthy place that supports open democratic debate.

Protecting election integrity does not end with an election period. As the challenges evolve, so will our approach. We will continue to work with peers and partners to tackle issues as they arise, with collaborations across government institutions, civil society experts, political parties, candidates, industry, and media organisations as we move towards our common goal of a healthy and open democratic process.

We trust this written submission will provide a useful input to the Government's report. Twitter is committed to working with the Australian Government, our industry partners, non-government organisations, and civil society as we continue to build our shared understanding of the issues and find optimal ways to approach these together.

Kind regards,

Kara Hinesley
Director of Public Policy
Australia and New Zealand

Kathleen Reen
Senior Director of Public Policy
Asia Pacific

Introduction

Twitter shows the world what is happening, democratises access to information, and at its best, provides people with insights into a diversity of perspectives on critical issues in real time.

We work with commitment and passion to do right by the people who use Twitter and the broader public. Any attempts to undermine the integrity of our service are antithetical to our fundamental principles and erodes freedom of expression, a core value upon which our company is based. This issue affects all of us and is one that we care deeply about as individuals, both inside and outside the company.

Throughout this submission, we detail our undertakings as a company to address this evolving threat, and we endeavour to illustratively emphasise key areas of effort and proactivity on behalf of Twitter to address the issues as defined by this Committee's terms of reference. We also aim to frame and highlight additional contextual issues related to cyber interference operations that the Committee should take into consideration when evaluating this dynamic landscape and preparing its final report.

Twitter's understanding of the evolving landscape

Interference in the affairs of other nations by state and non-state actors is not new, however technology has changed the ways in which it can be conducted.

What is important is to approach the issue as a broad geopolitical challenge, not one of content moderation. Removal of content alone will not address this challenge and while it does play an important role in addressing the challenge, governments must address the broader landscape. We do not elevate our own values by seeking to silence those who do not share them. In fact, we undermine these principles and erode their global accessibility.

Framing the issues

The spectrum of state actions - from so-called "white propaganda" originating from self-declared agents of the state to the messaging of state-controlled media and covert activity, including the use of fake accounts - continues to evolve. The threat of hostile cyber activity with the intent to acquire information for distribution remains clear, and Twitter has a specific policy prohibiting the distribution of hacked materials.[1]

As we have seen in other policy areas, this issue is a challenge where domestic media actors distribute the contents of a hack through their own reporting, potentially achieving the aim of the hostile actor to amplify a desired message to large audiences in spite of Twitter's efforts to remove accounts distributing the hacked materials.

[1] https://help.twitter.com/en/rules-and-policies/hacked-materials

On a technical level, the tools that were once the domain of a small number of state-sponsored actors are now commercialised with companies offering services to enable manipulation of online discourse.

Outside of the Australian context, there is evidence that this commercialisation is being exploited by hostile actors. Individuals are being paid to mask the identity of who is behind activity - both foreign[2] and domestic[3] - while the broader risk of foreign sources of finance being used to interfere with domestic affairs are well documented. The monetisation of misinformation risks further obscuring 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 being exploited by hostile actors, either directly or indirectly through proxies, emphasises 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, through direct or indirect partnerships.

Furthermore, the growth in the number of private sector actors claiming to be able to identify the origins of information operations, based on little more than so-called 'bot' research and weak analysis, 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," and we continue to see attribution based on limited public information from actors who appear motivated more by their own press coverage than rigorous methodology.[4]

Recommendations

The foremost challenge is for governments to communicate to build public trust, directly engaging with the conflicting narratives propagated on and offline by foreign actors.

Through clear and concise electoral regulations, companies are able to navigate and address relevant interference concerns. Additionally, public trust can be built through clear communication and strong attribution to address transparency issues related to interference. Throughout the 2019 Federal Election, the Australian Electoral Commission undertook a leadership role engaging in the public conversation on Twitter to provide clear, credible information, running promoted campaigns attributed and documented in our Ads Transparency Centre, and helped direct voters to reliable resources.[5]

It is the nature of a democratic society to uphold freedom of political speech and to encourage citizens to develop informed opinions so they have their voices heard via the electoral process. Through transparent communication and enabling voters to access the information they need, the

[2] https://www.cnn.com/2020/03/12/world/russia-ghana-troll-farms-2020-ward/index.html
[3] https://www.buzzfeednews.com/article/craigsilverman/facebook-account-rental-ad-laundering-scam
[4] https://www.disinfo.eu/publications/being-cautious-with-attribution-foreign-interference-covid-19-disinformation
[5] https://ads.twitter.com/transparency

Australian Government can foster a sense of trust and encourage freedom of discussion reflective of the implied freedom of political communication embodied within the Australian Constitution.[6]

How we define spam and platform manipulation

We want Twitter to be a place where people can make human connections, find reliable information, and express themselves freely and safely. To make that possible, we do not allow spam or other types of platform manipulation. We define platform manipulation as using Twitter to engage in bulk, aggressive, or deceptive activity that misleads others and/or disrupts their experience.[7]

The Twitter Rules also prohibit people from using 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.[8]

In our most recent Transparency Report, we further clarified our original spam policies, which elucidate on spam tactics that are being used for an expanded range of motivations. Spam may be:

- Commercial — persistent, often automated content which tries to get you to click a link or to buy something;
- Artificial amplification — actions to make an account or concept more popular or controversial than it actually is through inauthentic engagements;
- Coordinated activity — efforts to artificially influence conversations through the use of multiple and/or deceptive fake accounts; or a
- Combination of any of the above — for instance, spammers may attempt to capitalise on a popular topic in order to sell something, or ideologically-motivated actors may use spammy amplification tactics to attempt to boost their narrative.

Additionally, "spam reports" reflected in our transparency reports are numbers that are an aggregate of reports from people who use Twitter after receiving an interaction (for example, a follow, mention, or Direct Message) from a suspected spam account.

To determine whether a human is in control of an account we suspect is engaging in platform manipulation, we require the account holder to pass an "anti-spam challenge." For example, we may require the account holder to verify a phone number or email address, or complete a reCAPTCHA test. While these challenges are simple for authentic account owners to solve, they are difficult (or costly) for spammy or malicious account owners to complete. Accounts which fail to complete a challenge within a specified period of time may be suspended.

Transparency about our actions

[6] http://www.austlii.edu.au/cgi-bin/viewdoc/au/cases/cth/HCA/1992/46.html
[7] https://transparency.twitter.com/en/platform-manipulation.html
[8] https://help.twitter.com/en/rules-and-policies/platform-manipulation

Twitter engages in intensive efforts to identify and combat state-sponsored and non-state sponsored hostile attempts to abuse our platform for manipulative and divisive purposes. Our efforts enable Twitter to fight this threat while maintaining the integrity of peoples' experience and supporting the health of conversation on our service.

In order to effectively combat malicious activity on our service and supplement our above efforts, transparency is key. That's why we maintain open, public APIs of our service, provide retrospective reviews of elections, publish bi-annual Transparency Reports, and routinely disclose datasets of information operations we can reliably link to state actors.

Retrospective reviews of elections

To provide additional transparency around recent elections, we published retrospective reviews for the European Union (EU) Elections last year[9] and the United States (US) 2018 Midterm Elections.[10]

With specific regard to the Australian elections, we also provided a publicly available submission to the Parliamentary Joint Standing Committee on Electoral Matters regarding our work to protect and support the 2019 Australian Federal Election.[11] As noted in our submission, ahead of the 2019 Federal Election, we launched the Ads Transparency Centre in Australia, which includes a repository of all advertisements served on Twitter within the last seven calendar days, as well as all of the political campaign ads paid for by certified political advertisers in Australia.[12] Our policies have since changed to ban political advertising, but the advertising information from the 2019 election is still publicly available.[13]

Bi-annual Transparency Report

For the past eight years, our biannual Twitter Transparency Report has highlighted trends in requests made to Twitter from around the globe.[14] Over time, we have significantly expanded the information we disclose, including adding metrics on platform manipulation.

In our most recent reporting period, we observed a nearly 50% drop in challenges issued to suspected spam accounts compared to the previous reporting period, demonstrating the progress

---

9 https://www.politico.eu/wp-content/uploads/2019/07/Retrospective-Review_-Twitter-and-the-2019-European-Elections-.pdf?utm_source=POLITICO.EU&

10 https://blog.twitter.com/en_us/topics/company/2019/18_midterm_review.html

11 https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Electoral_Matters/2019Federalelection/Submissions

12 https://ads.twitter.com/transparency

13 https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html

14 https://transparency.twitter.com/en.html

we're making in this area.[15] Actions taken in relation to spam tend to fluctuate for a variety of reasons, for example variations in the total number of attempted Twitter sign-ups for the same time period, as well as the volume of spam campaigns targeting our service at any point in time. We also believe our continued focus on deterring unhealthy accounts at the time of account creation may contribute to the overall drop in anti-spam challenges issued. Reports of spam interactions remained relatively steady since the previous reporting period, decreasing about 1% from the previous report.

Our transparency reports reflect not only the evolution of the public conversation on our service, but they also reflect the work we do every day to protect and support the people who use Twitter. We want people to feel safe on Twitter, and we will continue to make strides in creating a healthier service.

Information operations disclosures

In 2018, Twitter made the decision to provide greater clarity on state-backed foreign information operations that we removed from the service.[16] We did so by creating a unique and comprehensive archive of the Tweets and media that were connected to potentially-state backed operations.[17] The archive is the largest of its kind in the industry, and now includes more than 160 million Tweets and more than eight terabytes of media.

In line with our strong principles of transparency and with the goal of improving understanding of foreign influence and information campaigns, we release archives of Tweets and media associated with potential information operations that we had found on our service, including the 3,613 accounts we believe were associated with the activities of the Internet Research Agency on Twitter dating back to 2009.[18] We made this data available with the goal of encouraging open research and investigation of these behaviors from researchers and academics around the world.

Our work on this issue is not done, nor will it ever be. It is clear that information operations and coordinated inauthentic behavior will not cease. These types of tactics have been around for far longer than Twitter has existed. They will adapt and change as the geopolitical terrain evolves worldwide and as new technologies emerge. Given this, the threat we face requires extensive partnership and collaboration with government entities, civil society experts, and industry peers. We each possess information the other does not have, and our combined efforts are more powerful together in combating these threats.

[15] https://transparency.twitter.com/en/platform-manipulation.html

[16] https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html

[17] https://transparency.twitter.com/en/information-operations.html

[18] https://blog.twitter.com/official/en_us/topics/company/2017/Update-Russian-Interference-in-2016--Election-Bots-and-Misinformation.html

This is why we have shared, and will continue to share, information about this activity with our peer companies to enable investigations of related activity on their services, and the relevant law enforcement entities.

We are proud that thousands of researchers have made use of the datasets in our archive - the largest in its kind in the industry - to conduct their own investigations and shared their insights and independent analyses with the world.

We launched this unique initiative to improve academic and public understanding of these coordinated campaigns around the world, and to empower independent, third-party scrutiny of these tactics on our platform. As a private company, there are threats that we cannot understand and address alone. We must continue to work together with elected officials, government entities, industry peers, outside experts, and other stakeholders so that the American people and the global community can understand the full context in which these threats arise.

'Behaviour-first' approach

Twitter has strengthened how it deals with disruptive behaviors that do not violate our policies, but negatively impact and distort the health of the conversation. Our approach focuses on behaviour over content by integrating new behavioral signals into how Tweets are presented. For example:

- If the account holder has not confirmed an email address;
- If the same person signs up for multiple accounts simultaneously;
- Accounts that repeatedly Tweet and mention accounts that don't follow them; or
- Behavior that might indicate a coordinated attack.

Previously, our actions were predicated on people reporting accounts or content that violated the Twitter Rules before we could take action. By using new tools to address this conduct from a behavioral perspective, we're able to proactively identify violative accounts and content at scale, strategically and at source, while also reducing the burden on people who use Twitter.

This approach has enabled us to take aggressive action on more abusers, stop hundreds of thousands of accounts from rejoining after a suspension for abusive behavior, and reduce abuse in conversations.

Over the past 12 months, we have made substantial strides in tackling abusive content on our service globally, including:

- More than one in two Tweets we take action on for abuse are now proactively surfaced using technology rather than relying on reports to Twitter. This compares to one in five Tweets in 2018.[19]
- We have seen a 105% increase in accounts actioned by Twitter, including accounts that have been either locked or suspended for violating the Twitter Rules.[20]
- In November 2019, we launched the option for people to hide replies to their Tweets. Now anyone can choose to hide replies to their Tweets.[21]
- In December 2019, we expanded and diversified our Trust and Safety Council, which brings together experts and organisations from around the world to help advise us as we develop our products, programs, and the Twitter Rules.[22]
- We also recently announced a new rule to address synthetic and manipulated media. Since 5 March 2020, we have started labeling Tweets if we believe that media shared within them have been significantly and deceptively altered or fabricated. Labels will link to a Twitter Moment to provide additional context from reputable sources on the content in question.[23]

Going forward, we will continue to work with governments, civil society, and industry to address online safety. Across all areas, the investments Twitter has made to protect the health of the public conversation are now generating clear and tangible safety benefits for people that use our service.[24]

Conclusion

Twitter recognises that there is much to do and remains committed to understanding how bad-faith actors use our service. We will continue our efforts to proactively combat nefarious attempts to undermine the health of public conversation.

One critical component of the response must be an elevated role for public diplomacy and public information. Foreign actors exploit social tensions, amplify polarisation, and undermine trust and confidence in democratic norms, institutions, and values. As the Swedish MSB highlights, "safeguarding the democratic dialogue — the right to open debate, the right to arrive at one's own opinions freely, and the right to free expression — is paramount as we work to lay a solid foundation of social resilience to counter information influence activities."[25]

Foreign interference is an unavoidable part of a globalised communications landscape. Policy makers should seek to build resilience and digital literacy to protect against activity, while taking the necessary steps to inform the public of the facts on key public policy issues, defending domestic policy, and advocating against hostile actors where necessary. The question must be asked how these operations

[19] https://blog.twitter.com/en_us/topics/events/2020/safer-internet-day-2020-creating-a-better-internet-for-all.html
[20] *Ibid*.
[21] https://help.twitter.com/en/using-twitter/mentions-and-replies
[22] https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-trust-and-safety-council.html
[23] https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html
[24] https://blog.twitter.com/en_us/topics/company/2019/health-update.html
[25] https://www.msb.se/RibData/Filer/pdf/28698.pdf

can be deterred, using the full range of government policies available. While activity may be visible on social media, it is just one part of the information ecosystem, and a response that looks at what happens on social media in isolation risks neglecting the wider challenge posed by foreign information operations.

The threat we face requires extensive partnership and collaboration with government entities, civil society experts, and industry peers. We look forward to partnering with stakeholders in this space to improve our collective understanding of coordinated attempts to interfere in public conversations and what are effective holistic responses.