20 August 2021

Committee Secretary
Australian Senate Select Committee on
Foreign Interference through Social Media
PO Box 6021
Parliament House
Canberra ACT 2600

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

Dear Chair,

Thank you for the opportunity to provide responses to questions taken on notice following Twitter's appearance as part of the Australian Senate Select Committee on Foreign Interference through Social Media inquiry into the risk posed to Australia's democracy by foreign interference through social media.

We have endeavoured to provide answers to the best of our ability for questions outlined by the Committee below.

Twitter is committed to working with the Australian Government, our industry partners, non-government organisations and wider civil society as we build our shared understanding of the issues and find optimal ways to approach these together.

Please don't hesitate to let us know if there is any additional information we can provide to assist the Committee. Thank you again for the opportunity to provide input as part of this important legislative process.

Kind regards,

[redacted]

[redacted]

Kara Hinesley
Director of Public Policy
Australia and New Zealand

Kathleen Reen
Senior Director of Public Policy
Asia Pacific

Australian Senate Select Committee on Foreign Interference through Social Media

Inquiry into the risk posed to Australia's democracy by foreign interference through social media

Answers to Questions on Notice
Twitter Inc.

1. **You have just talked a lot about transparency. It seems to me that one of the things that allows that interference….is the fact that it allows people to have anonymous accounts and there is no verification. Because of that, people are able to build bot farms and things like that in order to react. It is on Twitter, rather than on the other platforms, that you see the biggest pile-ons, the biggest divisions, the biggest brawls. Would you care to talk about your policy of not having verified accounts and not being able to verify where people are tweeting from et cetera? — Senator David Van**

For Twitter to endure, it needs to provide an environment for users to feel safe in communicating on the platform. To provide this environment, Twitter needs to have the ability to remove bad faith actors on the platform who intend to use it to divide, threaten, or manipulate.

Platform manipulation

People 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 prohibit the creation or use of fake accounts. We also do not allow spam or platform manipulation, such as bulk, aggressive, or deceptive activity that misleads others and disrupts their experience on Twitter.

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.

Twitter relies on behavioural signals – such as how accounts behave and react to one another – to identify accounts that detract from a healthy public conversation, such as spam and abuse. This includes building new proprietary systems to identify and remove ban evaders at speed and scale.

We prioritise 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 our most recent Transparency Report, we challenged over 143 million accounts for engaging in suspected spammy behaviour, including those engaged in suspected platform manipulation.[1] We have also implemented mandatory email or phone verification for all new accounts.

Since 2018, we also introduced a registration process for developers requesting access to our application programming interfaces (APIs) to prevent the registration of spammy and low quality apps, and we are continuing to roll out improvements to our proactive enforcements against common policy violations.[2]

Automation and automated accounts

People 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 a photo because they have strong privacy concerns.

[1] https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jul-dec

[2] https://blog.twitter.com/developer/en_us/topics/tips/2018/automation-and-the-use-of-multiple-accounts

In sum, a bot is an automated account. With regards to automation, our rules specifically state that platform manipulation and spam are prohibited on Twitter. People cannot use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behaviour that manipulates or disrupts people's experience on Twitter.

It's important to note, however, that not all forms of automation are violations of the Twitter Rules. We've seen innovative and creative uses of automation to enrich the Twitter experience. For example, accounts that track air quality, earthquakes, or general reminders to drink your water like @tinycarebot.

Automation can also be a powerful tool. For example, a conversational bot can help people find information about orders or voting information, like the Twitter Direct Message Chatbot we set up for the 2019 Australian Election.[3] This kind of innovative tooling has proved safe and efficient for a myriad of civic and corporate functions, especially at a time of social distancing.

The presence of a bot account on Twitter is not an indication that content from that user is shown or distributed in the same way as organic content, and our actions to limit the spread of spammy or automated content are not available to developers or researchers using the public APIs. The end user experience of someone using the Twitter app or website is not replicated by looking at an unfiltered stream of content obtained via our public API.

We see a lot of non-peer reviewed and commercially-driven research that makes sweeping assessments about automated accounts that are deeply flawed.

This means when groups of bots or incidents of malicious automation are identified by researchers, they are unable to factor in defensive measures taken by Twitter. The actions we take – such as challenging, filtering, and removing accounts – are not reflected in research. This being the case, it's reasonable to assume that a great many of the accounts presented in a data set may have already been addressed by our proactive measures and detection systems in some way.

Abuse and harassment

We're committed to enabling safe and healthy conversations on the service. We work with safety advocates, academics, researchers, and community groups that support our work to prevent abuse, harassment, and bullying. By providing continuous feedback on our safety mechanisms, these partners help us maintain a safe environment.

As part of that commitment, we have introduced a number of recent updates and policies to reduce misinformation, abuse, and harassment, including:

- Impressions for rule-violating Tweets: Our impressions metric captures the number of views a violative Tweet received prior to removal. From July - December 2020, Twitter removed rule-violating 3.5 million Tweets. Out of those Tweets, 77% received fewer than 100 impressions prior to removal.[4]
- In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets during that time period.[5]

We also stepped up the level of proactive enforcement across the service and invested in technological solutions to respond to ever-evolving malicious online activity. Today, by using technology, 65% of the abusive content we action is surfaced proactively for human review, instead of relying on reports from people using Twitter.[6]

Pseudonymous accounts

At Twitter, we are guided by our values, and never more so than when it comes to fundamental issues like identity. We believe everyone has the right to share their voice without requiring a government ID to do so. Our approach in this space has been developed in consultation with leading NGOs. While pseudonymity has been a vital tool for speaking out in oppressive regimes, it is no less critical in democratic societies.

Pseudonymity may be used to explore your identity, to find support as victims of crimes, or to highlight issues faced by vulnerable communities. Indeed, many of the first voices to speak out on societal wrongdoings, have

[3] https://blog.twitter.com/en_au/topics/company/2019/get--ausvotes2019-election-information-through-twitter-
[4] https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec
[5] https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center
[6] *Ibid.*

done so behind some degree of pseudonymity. Once they do, their experience can encourage others to do the same, knowing they don't have to put their name to their experience if they're not comfortable doing so.

Perhaps most fundamentally of all, some of the communities who may lack access to government IDs are exactly those who we strive to give a voice to on Twitter.

Pseudonymity is not a shield against Terms of Service violations, and Twitter will take action against pseudonymous accounts that are in violation of the Twitter Rules. It is against the rules to have a fake account on Twitter, and we have implemented mandatory email or phone verification for all new accounts.

Additionally, empirical evidence overwhelmingly points to anonymity bans as ineffective.[7] Currently, there is not conclusive evidence that requiring the display of names and identities will reliably reduce social problems, and many studies have documented the problems it creates, like posing real threats to vulnerable communities online that rely on anonymity — victims of domestic violence, members of the LGBTQIA+ community, political and human rights activists, journalists, and informants, to name a few.[8]

Additionally, there is also a deeper question regarding rights to privacy, expression, and association. Companies that store personal information for business purposes also expose people to potentially serious risks, especially when that information is leaked or a data breach occurs. Restrictions on anonymity or pseudonymity online risk deeply undermining trust in public debate and conversation and placing vulnerable communities at heightened likelihood of targeted abuse and harm.

Commitment to privacy

It is also critical to protect the privacy of the people who use online services. We offer a range of ways for people to control their privacy experience on Twitter, from offering pseudonymous accounts to letting people control who sees their Tweets to providing a wide array of granular privacy controls. Our privacy efforts have enabled people around the world to use Twitter to protect their own data.

That same philosophy guides how we work to protect the data people share with Twitter. We empower the people who use our service to make informed decisions about the data they share with us. We believe individuals should know, and have meaningful control over, what data is being collected about them, how it is used, and when it is shared.

We believe that individuals should control the personal data that is shared with companies and provide them with the tools to help them control their data. Through the account settings on Twitter, we give people the ability to make a variety of choices about their data privacy, including limiting the data we collect, determining whether they see interest-based advertising, and controlling how we personalise their experience.[9] In addition, we provide them with the ability to access information about advertisers that have included them in tailored audiences to serve them ads, demographic, and interest data about their account from ad partners, and information Twitter has inferred about them.

2. **In cases where you have provided researchers with access to platform data, have you made the results available to others—for example, research institutions, regulators such as the AEC, and government agencies—so that findings can be utilised in the public interest? Can you do this? — Senator Jim Molan**

In line with our principles of transparency and to improve understanding of the public conversation, Twitter makes available real-time access to the global conversation through our free, open application programming interfaces (APIs).[10]

At a high level, APIs are the way computer programs "talk" to each other so that they can request and deliver information. This is done by allowing a software application to call what's known as an endpoint: an address that corresponds with a specific type of information we provide (endpoints are generally unique like phone numbers).[11]

[7] Michigan State University (Huang, G & Li, K 2016), 'The effect of anonymity on conformity to group norms in online contexts: A meta-analysis' International Journal of Communication, vol. 10, no. 1, pp. 398–415. <https://ijoc.org/index.php/ijoc/article/view/4037>.
[8] Matias, J. N. (2017). Why Real Names Don't Fix Trolling. <https://guides.coralproject.net/real-names-dont-fix-trolling/>.
[9] https://help.twitter.com/en/privacy
[10] https://developer.twitter.com/en
[11] https://help.twitter.com/en/rules-and-policies/twitter-api

Twitter allows access to parts of our service via APIs to allow people to build software that integrates with Twitter, like a solution that helps a company respond to customer feedback on Twitter.

Our API platform provides broad access to Twitter data that users have chosen to share with the world. Twitter data is unique from data shared by most other social platforms because it reflects information that users choose to share publicly. We also support APIs that allow users to manage their own non-public Twitter information (e.g. Direct Messages) and provide this information to developers whom users have authorised to do so.

Academic partnerships

At Twitter, we believe that we do not have all the answers and have to work together for better outcomes. We are inspired by the first-in-the-field research by our academic partners, and humbled by the ongoing work to address the challenges at hand. Most of the new studies in this space are working with Twitter data, which is a reflection of the open nature of our service. However, we are conscious that Twitter data is only a subset of all the available information online.

With issues that are new, complex, and rapidly evolving, a lot of work is also being done to better define the scope and extent of online behaviours to be studied so as to make research methodologies more robust. Due to our public nature, Twitter is frequently scrutinised; however, we encourage robust study and academic partnerships through recent expansion of our APIs coupled with our own public disclosures, like our state-backed information operations database.[12]

State-backed information operations

We draw a clear line between the use of our service to engage in political discourse versus attempts to manipulate the public conversation by inauthentic means, the latter of which is strictly prohibited.

Our data disclosures on such operations are part of Twitter's efforts to contribute to such independent analysis, which now constitutes the largest archive of state-backed information operations on Twitter. We also believe that independent analysis of this activity by researchers is a key step toward promoting shared understanding of these threats.

In line with this commitment, we have proactively shared 37 datasets from 17 countries, which contain over 200 million Tweets and 9 terabytes of media.

Twitter is the only company to make its archive – which is now the largest of its kind in the industry – fully available to the public and has been accessed thousands of times by researchers all over the world.

For example, in our data disclosure in June 2020, we shared relevant data with research partner the Australian Strategic Policy Institute (ASPI), who published the report Retweeting through the Great Firewall utilising these new data sets.[13]

Their analysis looked at a dataset of over 23,000 Twitter accounts and almost 350,000 Tweets that occurred from January 2018 to 17 April 2020, which we attributed to Chinese state-linked actors and took the accounts offline.

This activity largely targeted Chinese-speaking audiences outside of the Chinese mainland (where Twitter is blocked) with the intention of influencing perceptions on key issues, including the Hong Kong protests.

The main vector of dissemination was through images, many of which contained embedded Chinese-language text. The linguistic traits within the dataset suggest that audiences in Hong Kong were a primary target for this campaign, with the broader Chinese diaspora as a secondary audience.

Based on the data in the takedown dataset, ASPI found that these efforts are sufficiently technically sophisticated to persist, but they lacked the linguistic and cultural refinement to drive engagement on Twitter through high-follower networks, and thus far have had relatively low impact on the platform.

The operation's targeting of higher value aged accounts as vehicles for amplifying reach, potentially through the influence-for-hire marketplace, is likely to have been a strategy to obfuscate the campaign's state-sponsorship.

[12] https://transparency.twitter.com/en/reports/information-operations.html
[13] https://www.aspi.org.au/report/retweeting-through-great-firewall

This suggests that the operators lacked the confidence, capability, and credibility to develop high-value personas on the platform. This mode of operation highlights the emerging nexus between state-linked propaganda and the internet's public relations shadow economy, which offers state actors opportunities for outsourcing their disinformation propagation.

Similar studies support ASPI report's findings. Graphika has undertaken two studies of a persistent campaign targeting the Hong Kong protests, Guo Wengui (Miles Kwok) and other critics of the Chinese Government.[14] Researchers at Bellingcat have also previously reported on networks targeting Guo Wengui and the Hong Kong protest movement.[15]

Together, we are at the beginning of a journey to understand the issues at hand. In order to make real progress in this space we would encourage more services and more stakeholders to play a greater role in supporting more research in these fields. We believe that partnerships with the Government, including law enforcement, academics, and the wider community will further improve our understanding of coordinated attempts to interfere with the public conversation and the best ways to combat it.

[14] https://public-assets.graphika.com/reports/Graphika_Report_Spamouflage_Returns.pdf

[15] https://www.bellingcat.com/news/2020/05/05/uncovering-a-pro-chinese-government-information-operation-on-twitter-and-facebook-analysis-of-the-milesguo-bot-network/