31 August 2018

Committee Secretary
Joint Standing Committee on Electoral Matters
PO Box 6021
Parliament House
Canberra ACT 2600

By email: em@aph.gov.au

Dear Committee Secretary,

Thank you for the opportunity to make this submission and participate in the Australian Parliamentary Inquiry into the conduct of the 2016 Federal Election, specifically:

- the extent to which maliciously automated accounts may have targeted Australian voters and political discourse in the past;
- the likely sources of malicious automation on social media platforms within Australia and internationally;
- ways to address the spread of disinformation (i.e. deliberately false news) online during elections; and
- measures to improve the media literacy of Australian voters.

Twitter deeply respects the integrity of the election process, which is a cornerstone for all democracies around the world, and we are committed to providing a platform that fosters healthy civic discourse and democratic debate in Australia.

Everyday people come to Twitter to see what's happening. One of the key areas we are focusing on as a company is improving the health of conversations on Twitter. To help move toward this goal, in the past sixteen months we've introduced over thirty new measures to fight abuse and trolls, new policies on hateful conduct and violent extremism[1] and are introducing new technology and staff to fight spam and abuse.[2]

Beginning last year, we publicly outlined some of our work to combat bots and networks of manipulation on Twitter.[3] Since then we have received a number of questions about how

[1] Twitter Safety (18 December 2017). Enforcing New Rules to Reduce Hateful Conduct and Abusive Behavior [Blog post]. Retrieved from https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html
[2] Twitter Safety (21 June 2018). Continuing our commitment to health [Blog post]. Retrieved from https://blog.twitter.com/official/en_us/topics/company/2018/CommitmentToHealth.html
[3] Twitter (14 June 2017). Our Approach to Bots & Misinformation [Blog post]. Retrieved from (https://blog.twitter.com/official/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html;

maliciously automated accounts and disinformation networks on Twitter may have been used in the context of federal elections.

It's important to note our work to fight both maliciously automated accounts and disinformation goes beyond any one specific election, event, or time period. We've spent years working to identify and remove spammy or malicious accounts and applications on Twitter. We continue to improve our internal systems to detect and prevent new forms of spam and malicious automation in real time while also expanding our efforts to educate the public on how to identify and use quality content on Twitter.

With hundreds of millions of Tweets sent globally every day, scaling these efforts continues to be a challenge. We will continue to look into these matters on an ongoing basis.

## Russia Today

The United States intelligence community released a report in January 2017 highlighting the role that Russia Today (hereinafter referred to as 'RT'), which has strong links to the Russian government, allegedly played in seeking to interfere in the 2016 United States election.[4] RT has accounts on Twitter and Tweets regularly. The open nature of the Twitter platform means this activity was public.

In September 2017, we determined that three RT accounts (@RT_com, @RT_America, and @ActualidadRT) spent $274,100 USD on ads in 2016 and promoted 1,823 Tweets that potentially targeted the United States market. These campaigns were directed at followers of mainstream media, and primarily promoted RT Tweets regarding news stories.

In October 2017, Twitter made the policy decision to off-board advertising from all accounts owned by Russia Today (RT) and Sputnik effective immediately. This decision was based on the retrospective work we've been doing around the 2016 United States election and the United States intelligence community's conclusion that both RT and Sputnik attempted to interfere with the election on behalf of the Russian government.[5]

We took this step as part of our ongoing commitment to help protect the integrity of the user experience on Twitter. This decision is restricted to these two entities based on our internal investigation of their behavior, as well as their inclusion in the January 2017 DNI report. RT and Sputnik may remain organic users on our platform in accordance with the Twitter Rules.

Twitter took the $1.9 million we were projected to have earned from RT global advertising since they became an advertiser in 2011 and donated those funds to support external research. This

[4] ICA (January 2018). 'Background to "Assessing Russian Activities and Intentions
in Recent US Elections": The Analytic Process and Cyber
Incident Attribution. Retrieved from https://www.dni.gov/files/documents/ICA_2017_01.pdf
[5] *Ibid*.

research is focused on the use of Twitter in civic engagement and elections, including use of malicious automation and misinformation with an initial focus on elections and automation.

Additionally, we identified and suspended a number of accounts on Twitter that were potentially connected to a propaganda effort by a Russian government-linked organization known as the Internet Research Agency (IRA). We identified an additional 1,062 accounts associated with the IRA. We suspended all of these accounts for Terms of Service violations, which were primarily violations of our spam policy, and all but a few accounts, which were restored to legitimate users, remain suspended. Because we suspended these accounts, the relevant content on Twitter is no longer publicly available in accordance with our Privacy Policy.

## Election Vote Issues

We take violations of our Terms of Service very seriously and respect local Australian law concerning interference in the exercise of voting rights. When we become aware of activity that violates our TOS, we take appropriate and timely action.

During the 2016 election, we were not made aware of any activity related to the suppression or interference with the exercise of voting rights in Australia.

## Political Advertising Policies

We note recent calls for increased public disclosure with respect to political advertisements on social media, including Twitter. Twitter supports making political advertising more transparent to our users and the public.

Internally, we already have stricter policies for advertising campaigns on Twitter than we do for organic content. We also have existing specific policies and review mechanisms for campaign advertising.[6]

Twitter's current advertising policies permit political campaigning advertising, but we maintain additional country level restrictions. In addition to Twitter advertising policies, all political campaigning advertisers must comply with applicable laws regarding disclosure and content requirements, eligibility restrictions, and blackout dates for the countries where they advertise, including Australia.

We welcome the opportunity to work with the Australian Electoral Commission and leaders in Parliament to review and strengthen guidelines for political advertising on social media.

## Automated Traffic and Spam

[6] "Twitter Advertising Policy." Twitter. https://business.twitter.com/en/help/ads-policies/restricted-content-policies/political-campaigning.html. Web (30 August 2018).

Every online platform has to deal with spam, and there is no silver bullet to fix this ongoing issue. For example, the Internet Society estimated in October 2015 that up to 85 percent of all global email is spam -- and that's after decades of every email platform in the world tackling this challenge.[7] Email is very different from Tweets, but it's important to understand the scale of this phenomenon and that it is a global issue for all platforms.

## New Processes for Fighting  Malicious Automation and Spam

Twitter fights spam and malicious automation strategically and on a global scale. Our focus is increasingly on proactively identifying problematic accounts and behavior rather than waiting until we receive a report. We focus on developing machine learning tools that identify and act on networks of spammy or automated accounts automatically by tracking account behaviour. This lets us tackle attempts to manipulate conversations on Twitter at scale, across languages, and different time zones.

As patterns of malicious activity evolve, we're adapting to meet them head-on. Our investments in this space are having a positive impact:

- **In May 2018, our systems identified and challenged more than 9.9 million potentially spammy or automated accounts per week.** That's up from 6.4 million in December 2017 and 3.2 million in September 2017.
- Due to technology and process improvements during the past year, we are now removing 214 percent more accounts for violating our spam policies on a year-on-year basis.
- At the same time, **the average number of spam reports we received through our reporting flow continued to drop — from an average of approximately 25,000 per day in March, to approximately 17,000 per day in May 2018.** We've also seen a ten percent drop in spam reports from search as a result of our recent changes. These decreases in reports received means people are encountering less spam in their timeline, search, and across the Twitter product.
- We're also moving rapidly to curb spam and abuse originating via Twitter APIs.[8] **In Q1 2018, we suspended more than 142,000 applications in violation of our rules — collectively responsible for more than 130 million low-quality, spammy Tweets.** We've maintained this pace of proactive action removing an average of more than 49,000 malicious applications per month in April and May this year. We are increasingly using automated and proactive detection methods to find misuses of our platform before they impact anyone's experience. **More than half of the applications we suspended in Q1 2018 were suspended within one week of registration, many within hours.**

[7] Internet Society (30 October 2015). Policy Brief: The Challenge of Spam. Retrieved from https://www.internetsociety.org/policybriefs/spam

[8] "Twitter Developer Policy." Twitter. https://developer.twitter.com/en/docs/tweets/search/overview. Web (30 August 2018).

These numbers tell us that our tools are working. We're preventing or catching more of this activity ourselves before it is seen on Twitter.

These are just some of our tools:

- **The most effective way to fight suspicious bots is stopping them before they start.** To do this, we've built systems to identify suspicious attempts to log in to Twitter, including signs that a login may be automated or scripted. Importantly, much of this defensive work is done through machine learning and automated processes on our back end, and we have been able to significantly improve our automatic spam and bot-detection tools.
- **We're investing in systems to stop bad content at its source** if its point of origin corresponds with a known bad actor. However, the use of proxy servers, virtual private networks (VPNs), and other forms of authentication, may obscure the true origin of traffic on Twitter. We are working on better identifying the true origins of traffic and blocking activity from suspicious sources.
- **We're also improving how we detect and cluster accounts that were created by a single entity or a single suspicious source.**
- **Detecting non-human activity patterns.** Using signals like the frequency and timing of Tweets and engagements, we've built models that can detect whether an activity on Twitter is likely automated. We're expanding how we use these signals to restrict how users see suspicious accounts.
- **Compromised account detection.** To stop bad actors from exploiting otherwise healthy accounts to spread malicious content, we're investing in new ways to identify potentially compromised accounts. For example, we're building systems to detect when login activity is inconsistent with a user's typical behavior and to help get compromised accounts back under the control of their owners.
- **Checking suspicious content.** Accounts and content detected by our systems are subject to a number of enforcement actions and limitations including: being placed in a "read only" mode pending authentication, having the reach of Tweets limited based on suspicious origin or low-quality content, the removal of associated content, and account suspension.
- **Third-party apps.** We're also continuing to invest in proactively identifying and acting against applications that violate our developer policies, including bots and automated apps. While some bots can provide a vital public utility in times of crisis and natural disaster, we're committed to combating the minority of apps that create spam and abuse via our application programming interface (API).
- **Preventing false positives.** Any automated system for detecting spam or bots has a chance of false positives, and it's our goal to bring this as low as possible. That's why we typically give users caught by our spam detections an opportunity to verify that they're legitimate before we suspend them from the platform. During this window, accounts may still appear on Twitter and via our public API, even though they are not

able to create new Tweets and engagements. We also limit the visibility of these accounts and their content in both Search and Trends.

- **Improving phone verification.** When we detect suspicious activity from an account, we may require that user to verify their phone number to regain access to Twitter. However, as spammers have adapted their techniques, we've found that not all phone numbers are equally trustworthy. We've improved our phone reputation system to identify suspicious carriers and numbers and prevent their repeated use to pass verification challenges.

It's also important to note that third-party research of the impact of bots and automation on Twitter systematically under-represents our enforcement actions. This is because these defensive actions are not visible via our API and because they take place shortly after content is created and delivered via our streaming API. Furthermore, researchers using an API often overlook the substantial in-product features that prioritize the most relevant content. Based on user interests and choices, we limit the visibility of low-quality content using tools such as Quality Filter and Safe Search, both of which are on by default for all of Twitter's users and active for more than 97 percent of users.

## Gaming Trending Topics

Over the years we've invested heavily in thwarting spam and other automated attempts to manipulate Trends. We take active measures to protect against trend gaming, such as excluding automated Tweets and users from our calculations of a Trend.

Importantly as spammers change their tactics, we actively modify our technological tools to address such situations. Since June 2017, we've been able to detect an average of 130,000 accounts per day who are attempting to manipulate Trends and have taken steps to prevent that impact. This is an area where saying more about those steps would only help bad actors, but we will keep looking for new ways to illustrate our efforts with examples on the important progress we've made on this front.

## Electoral Outreach

We engage with national and state-based electoral commissions regularly and consistently bolster our security and agent review coverage during key moments of election cycles around the world. We will continue to do this and expand our outreach so that we've got strong, clear escalation processes in place for all potentialities during major elections and political events.

## Supporting Media Literacy and Accurate Emergency Information

In our increasingly polarized public sphere, we believe developing critical media skills is more important than ever. We recognise that Twitter is an important part of a larger ecosystem of how news and information spreads online, and we have a responsibility to support external programs that empower our users, connecting them with resources to give them control over their online experience.

Our partners Common Sense Media, the National Association for Media Literacy, the Family Online Safety Institute and Connect Safely, amongst others, have helped us to craft materials and conduct workshops to help our users learn how to process online information and understand which sources of news have integrity. We focus on elements like verification of sources, critical thinking, active citizenship online, and the breaking down of digital divides.

## Next Steps

Over the coming months, Twitter will be rolling out several changes to the actions we take when we detect spammy or suspicious activity, including introducing escalations for enforcement of suspicious logins, Tweets, and engagements, as well as shortening the amount of time suspicious accounts remain visible on Twitter while pending confirmation.

These are not meant to be definitive solutions. We've been fighting against these issues for years, and as long as there are people trying to manipulate Twitter, we will be working hard to stop them.

Twitter is where the world goes to see what's happening, and we are a platform founded on a commitment to transparency. We take that legacy and responsibility seriously.

We are committed to ensuring that Twitter is safe and secure for all users and serves to advance healthy civic discussion and engagement, and we will continue in our efforts to protect Twitter against bad actors and networks of malicious automation and manipulation.