EMC Submission No 126
Received 26 March 2021

26 March 2021

Committee Secretary
Victorian Parliamentary Committee on Electoral Matters
Parliament of Victoria
Spring Street
East Melbourne VIC 3002

By email: [redacted]

Dear Committee Secretary,

Thank you for the opportunity to provide a written submission to the Victorian Parliamentary Electoral Matters Committee inquiry into the impact of social media on elections and electoral administration.

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

This submission supplements [redacted] It also draws from public contributions Twitter made to the Australian Joint Standing Committee on Electoral Matters' inquiry into the 2019 Australian Federal Election, as well as the Australian Senate Select Committee on Foreign Interference through Social Media.

Twitter is committed to working with the Australian Government, the Victorian 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.

As we head into future elections, Twitter is committed to building on the knowledge and expertise gained from the recent elections and creating a place where people from around the world come together to safely engage in an open and free exchange of ideas.

Thank you again for the opportunity to input to this important process.

Kind regards,

[redacted]

Kara Hinesley
Director of Public Policy
Australia and New Zealand

[redacted]

Kathleen Reen
Senior Director of Public Policy
Asia Pacific

Introduction

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

Deploying a full range of policies, products, and partnerships, we are proud of the role Twitter has played in past elections. We thank both the Australian Electoral Commission and the Victorian Electoral Commission, along with our local and global partners, who enabled us to contribute to such vital processes with success. While our work to improve the health of the conversation has evolved, our work is never done. Our submission here will focus on:

- The lessons learned from global elections and how this translates in our work in Australia and Victorian Elections;
- Our approach to disclosing state-affiliated information operations on our service and how it makes us the only company in the industry to do this; and
- Our efforts to safeguard the conversation on Twitter, including updates to our rules governing election information and political advertising.

I. LESSONS LEARNED FROM GLOBAL ELECTIONS

The public conversation on Twitter is never more important than during elections, and access to the free flow of information is underpinned by an Open Internet. Twitter plays a unique role as a digital square for public conversation as our service shows the world what is happening. By democratising access to information, people are able to be better informed, learn from each other, and glean 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.

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. We possess a deeper understanding of both the scope and tactics used by malicious actors to manipulate our service and sow division across Twitter more broadly. Our efforts enable Twitter to fight this threat while maintaining the integrity of peoples' experiences and supporting the health of the public conversation on our service.

A. *How this applies in Australia and Victorian Elections*

The democratic process, which is essential to the Australian way of life, relies on the exchange of ideas, information, and healthy debate. We are committed to creating a service where Australians can access that information and engage in those conversations about the policies and issues that affect them.

Detailed below is a summary, and then further information about our approaches, our teams, and our latest efforts at the forefront of addressing challenges to healthy, safe, and credible election conversations.

Drawing on our insights from previous elections, both globally and in Australia, we have strengthened our products, policies, and operational aspects of Twitter's service. With focus on protecting and supporting the conversation during the course of elections, we have undertaken intensive efforts to identify and combat any attempt to undermine or abuse our service, including:

- First, updating and enforcing our updated Civic Integrity Policy, as well as our policies on account ban evasion and hacked materials;[1]
- Second, utilising the tools within our Partner Support Portal that allow trusted partners, like the Victorian Electoral Commission, to directly and swiftly report deliberately misleading election-related content to us.[2]
- And third, challenging suspicious accounts; between January and June 2020, our internal, proactive tools challenged over 135.7 million accounts globally for engaging in suspected spammy behaviour, including those engaging in suspected platform manipulation.[3]

From the outset of any election, we establish a dedicated internal cross-functional team to lead our election integrity work. This team undertakes intense preparation, including assessments of potential malicious or automated activity by bad actors, either foreign or domestic – work that is ongoing even outside of election cycles.

Transparency and openness are deep-seated values at the heart of Twitter which define and guide our methodology around state-backed information operations and disclosures. When we identify evidence of state-backed information operations on Twitter, we publish archives of Tweets and media associated with those operations.[4] Twitter is the only company to make its

[1] https://help.twitter.com/en/rules-and-policies/hacked-materials
[2] https://help.twitter.com/en/rules-and-policies/election-integrity-policy
[3] Platform manipulation refers to the use of Twitter to mislead others and/or disrupt their experience by engaging in bulk, aggressive, or deceptive activity. This activity includes, but is not limited to, spam, malicious automation (malicious use of bots), and inauthentic account abuse (fake accounts). https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jan-jun
[4] https://blog.twitter.com/en_us/topics/company/2019/information-ops-on-twitter.html

archive – now the largest of its kind in the industry – of state-backed information fully available to the research community; we encourage researchers and experts to tap into these data sets to conduct their own investigations and share their independent analysis with the world. The purpose of this effort is not just transparency, but also to deter such behaviours, and encourage new findings that will inform collective efforts – our own, and for the wider industry.

We also verify political candidates on Twitter and on-board government organisations and political parties onto Twitter's Partner Support Portal, which allow these organisations to rapidly report suspected violations of the Twitter Rules. No reports were made via the Partner Support Portal throughout the 2018 Victorian Election period.

Recognising that our Australian audience is active during an election campaign and often turns to Twitter to talk about politics, we also create dedicated Event pages in the Twitter Explore Tab summarising daily election news from credible sources and showcasing the very best of what's happening on Twitter.[5] In this way, Twitter's goal is to find and highlight high-quality and media rich Tweets that provide insight and context to the election conversations happening on our service. We are committed to upholding high standards of accuracy, impartiality, and fairness in our curation efforts, which are designed to feature compelling, original, and diverse content.

Twitter is constantly investing in technology, developing new policies, and building meaningful partnerships to further our understanding of the political and social context within which Twitter operates. In the sections below, we explain the work undertaken for each of these pillars in more detail.

B. *Ongoing Efforts to Safeguard Elections*

Twitter continues to demonstrate a strong commitment to transparency regarding our election integrity efforts. We publish biannual Transparency Reports, routine recaps of global elections on our blog, and issue reports of our findings for relevant electoral authorities.[6] We are proud to publicly document our efforts to increase voter turnout, combat voter suppression, and provide greater clarity on the limited state-backed foreign information operations we proactively remove from the service.[7]

Our work on this issue is not done, nor will it ever be. It is clear that information operations and platform manipulation 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. As such, 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

[5] https://twitter.com/TwitterAU/status/1116160976645017602?s=20
[6] https://transparency.twitter.com/
[7] https://blog.twitter.com/content/dam/blog-twitter/official/en_us/company/2019/2018-retrospectiv e-review.pdf

efforts are more powerful together in combating these threats.

The process of investigating suspected foreign influence and information campaigns is an ongoing one. We remain vigilant about identifying and eliminating abuse on the service perpetrated by hostile foreign actors, and we will continue to invest in resources and leverage our technological capabilities to do so.

II. STATE-BACKED INFORMATION OPERATIONS

To date, our teams have not observed or found any foreign manipulation or foreign malicious activity related to the suppression or interference with an election in Australia.

In other markets outside of Australia, where we have observed foreign manipulation or foreign malicious activity, as referred above, we release full, comprehensive archives of Tweets and media associated with potential information operations that we have found on our service in line with our strong principles of transparency and with the goal of improving understanding of foreign influence and information campaigns.

These datasets live in our public archive of state-backed information operations – the largest of its kind in the industry. First launched in October 2018, the archive has been accessed by thousands of researchers from around the world, who, in turn, have conducted independent, third-party investigations of their own.[8]

Prior to the release of these datasets, Twitter shared examples of alleged foreign interference in political conversations on Twitter by the Internet Research Agency (IRA) and provided the public with a direct notice if they interacted with these accounts. 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.

We recognise that, as a private company, there are threats we cannot understand and address alone. Thus, we will continue to work together with elected officials, government entities, industry peers, outside experts, and other stakeholders so that the global community can understand the fuller context in which these threats arise.

III. SAFEGUARDING THE CONVERSATION

We strongly believe that any attempt to undermine the integrity of our service undermines freedom of expression. To address and mitigate those challenges, we have made numerous updates to the Twitter Rules that govern our policies relating to elections and political

[8]https://blog.twitter.com/en_us/topics/company/2019/new-disclosures-to-our-archive-of-state-backed-information-operations.html

advertising.

A. *Twitter Rules relating to safety and behaviour-first approach*

The Twitter Rules are a living document, and our policies and enforcement options evolve continuously to address emerging behaviours online.[9] Given the scale we're working at and erring on the side of protecting the voice of people who use Twitter, context is crucial when we review content, and we take a variety of signals into account to help us make the most informed decisions possible. Every day we observe new ways people behave on the service, and our team works hard to ensure we're enforcing our rules fairly and consistently, especially in light of these changing online norms.

Twitter has adopted a 'behaviour-first' approach, which utilises machine learning to identify behaviours that distort and detract from the public conversation on Twitter. This system is informed by thousands and thousands of account behaviours. We also look at how accounts are connected to those that violate our Rules and how they interact with each other.

These signals inform how we organise and present content in communal areas like conversation and search. The result is that people contributing to the healthy conversation will be more visible in conversations and in Twitter Search, while unhealthy behaviour that distorts and distracts from the public conversation is less visible.

Twitter has also 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
- Behaviour that might indicate a coordinated attack.

Previously, our actions were predicated on people reporting accounts to us directly or content that violated the Twitter Rules before we could take action. By using new tools to address this conduct from a behavioural 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.

[9] https://blog.twitter.com/en_us/topics/company/2018/TheTwitterRulesALivingDocument.html

This approach, which continues to evolve, 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 to reduce abuse within conversations.

Further, over the past year, we made additional 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.[10]
- 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.[11]
- 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.[12]
- In December 2019, we expanded and diversified our Trust and Safety Council, which brings together experts and organisations from around the world, including Australia and specifically in Victoria too, to help advise us as we develop our products, programs, and the Twitter Rules.[13]
- In July 2019, we expanded our rules against Hateful Conduct to include language that dehumanises others on the basis of religion or caste. In March 2020, we further added to this policy to prohibit language that dehumanises people on the basis of age, disability, or disease. Then December 2020, we further expanded our hateful conduct policy substantially for a third time to also prohibit language that dehumanises people on the basis of race, ethnicity, or national origin;[14]These steps also reflect the important process of systems to support application and enforcement of these critical changes.
- In August 2020, we began including the sender's public profile information in Direct Messages, and indicate how the sender is connected to the receiver,which can help people quickly identify potentially abusive content;[15]
- In July 2020, we introduced a new URL Policy to limit or prevent the spread of malicious URL links to content outside Twitter;[16]
- In 2020, we also introduced *conversational controls,* a feature that allows people to choose who will be able to reply to your Tweet. You'll see a default setting of 'Everyone can reply' next to a globe icon in the compose Tweet box. Clicking or tapping this prior to posting your Tweet allows you to choose if you want 'Everyone,' 'People you follow,'

[10]https://blog.twitter.com/en_us/topics/events/2020/safer-internet-day-2020-creating-a-better-internet-for-all.html
[11] *Ibid*.
[12] https://help.twitter.com/en/using-twitter/mentions-and-replies
[13] https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-trust-and-safety-council.html
[14] https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html
[15]https://twitter.com/TwitterSupport/status/1296560915886870529?utm_campaign=OT_NL_GBL_EN_OneTeamWeekly_082420_200824&utm_medium=email&utm_source=Eloqua
[16] https://help.twitter.com/en/safety-and-security/phishing-spam-and-malware-links

or 'Only people you mention' can reply to you.[17] These controls are more than about choices; they are about enabling the people who use our service to manage who they are in conversations with too.

Keeping people safe on Twitter remains a top priority. We won't stop working to build a healthier Twitter, so people feel safe and are able to find high-quality information on our service.

Some of the changes and updates we're actively working on toward that goal include further improving our technology to help us review content that breaks Twitter's Rules faster and before it's reported, specifically Tweets with private information, threats, and other types of abuse. We are also making it easier for people who use Twitter to share more specifics when reporting so we can take action faster, especially when it comes to protecting people's physical safety. We are also adding more notices to people who use our service, to provide deeper context for our enforcement decisions (e.g. if a Tweet breaks our rules but remains on the service because the content is in the public interest).

Going forward, we will continue to work closely with governments, civil society, and industry to address and expand 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.

B. *Twitter Rules relating to elections*

As summarised above, we have made a number of recent updates to the rules governing the use of our service to better protect the conversation. In addition to new prohibitions on inauthentic activity, ban evasion, and hacked materials, we have also codified our policy regarding civic integrity governing multiple categories of manipulative behavior and content related to elections.[18]

First, an individual cannot share false or misleading information about how to participate in an election or other civic event. This includes but is not limited to misleading information about how to vote, register to vote, requirements for voting, and the official announced date or time of an election.

Second, an individual cannot share false or misleading information intended to intimidate or dissuade voters from participating in an election. This includes, but is not limited to, misleading claims that polling places are closed, that polling has ended, or other misleading information relating to votes not being counted.

---

[17] https://help.twitter.com/en/using-twitter/twitter-conversations
[18] https://help.twitter.com/en/rules-and-policies/election-integrity-policy

Third, we do not allow misleading claims about police or law enforcement activity related to polling places or elections, long lines, equipment problems, voting procedures, or techniques which could dissuade voters from participating in an election, and threats regarding voting locations.

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

At the same time, we have prioritised our approach to tackle misinformation based on the highest potential for harm, which is why our primary focus is also on three areas overall; synthetic and manipulated media, civic integrity, and laterally, COVID-19. We understand that not all misleading information is the same and hence adopt a flexible approach to protect and serve the public conversation taking place during elections including our approach to tackle misinformation.

For instance on 5 March 2020, we put into effect our policy to address synthetic and manipulated media on Twitter.[19] We believe that we need to consider how synthetic media is shared on Twitter in potentially damaging contexts. We also want to listen and consider a variety of perspectives in our policy development process, and we want to be transparent about our approach and values.

Through an initial public feedback period for the development of our synthetic and manipulated media policy, we gathered more than 6,500 responses from people around the world on the policy and issues. We also consulted with a diverse, global group of civil society and academic experts on our draft approach. In summary, some of the highlights of what we learned included that people felt:

- Twitter should give more information: Globally, more than 70 percent of people who use Twitter said “taking no action” on misleading altered media would be unacceptable. Respondents were nearly unanimous in their support for Twitter providing additional information or context on Tweets that have this type of media.
- That this type of content should be labeled: Nearly 9 out of 10 individuals said placing warning labels next to significantly altered content would be acceptable. That is about as many who said it would be acceptable to alert people before they Tweet misleading altered media.
- And, if it is likely to cause harm, it should be removed: More than 90 percent of people who shared feedback support Twitter removing this content when it’s clear that it is intended to cause certain types of harm.

[19] https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

Based on this approach and feedback, we introduced the rule in 2020 that states people may not deceptively share synthetic or manipulated media that are likely to cause harm. In addition, we may also label Tweets containing synthetic and manipulated media to help people understand the media's authenticity and to provide additional context.

As we continually learn from elections and civic processes happening around the world, we will continue to consider a wider spectrum of remediation options to address content that does not meet the strict standards for outright removal, and review flexible frameworks that will address marginal content while still providing transparency, and upholding our core commitment to free expression.

C. *Twitter Rules relating to political advertising*

On 30 October 2019, Twitter's chief executive officer Jack Dorsey announced that we made the decision to stop all political advertising on Twitter globally. We remain the only platform to date to implement a ban on political advertising.

We believe political message reach should be earned, not bought. This means bringing ads from political candidates and political parties to an end. We define political content as content that refers to a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome.[20]

Ads that contain references to political content, including appeals for votes, solicitations of financial support, and advocacy for or against any of the above-listed types of political content, are prohibited under this policy. We also do not allow ads of any type by candidates, political parties, or elected or appointed government officials.

A political message earns reach when people decide to follow an account or retweet. Paying for reach removes that decision, forcing highly optimised and targeted political messages on people. We believe this decision should not be compromised by money. 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.

We believe our approach to political advertising does not compromise free expression because candidates and political parties can continue to share their content organically. This is about paying for reach, and paying to increase the reach of this political speech has significant ramifications for democratic infrastructures.

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

Conclusion

As we head into future elections, Twitter is deeply committed to building on the knowledge and expertise gained from the elections we have overseen to date, and creating a place where people in Australia and from around the world come together to engage in an open and free exchange of ideas.

All people who use Twitter must have confidence in the integrity of the information found on the service. We continue to invest in our efforts to address threats posed by hostile actors, and to work consistently to foster an environment conducive to healthy, meaningful conversations on our service. Thank you again for the opportunity for Twitter to contribute and participate in this process, and we look forward to working with the Committee on these important issues.