1 July 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 provide a written submission to the Select Committee on Foreign Interference through Social Media regarding Twitter's response to the recent global pandemic and our efforts, both globally and in Australia, to respond to COVID-19.

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. As the global community faces the COVID-19 pandemic together, Twitter is helping people find reliable information, connect with others, and follow what's happening in real time.

We trust this written submission will provide a useful input to the Government's work in this space. 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 global problem sets together.

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 and that conversation is never more important than during crisis events, including the global COVID-19 pandemic.

Our service gives people the ability to share what is happening and provides people with insights into a diversity of perspectives on critical issues, all in real time.

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.

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.

Enumerated in our policies, 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.[1] 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.[2]

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.[3] When we identify such activity, we require an individual using the service to confirm human control of the account or their identity.[4]

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

[1] Twitter (September 2019). Platform manipulation and spam policy. https://help.twitter.com/en/rules-and-policies/platform-manipulation.
[2] *Ibid*.
[3] *Ibid*.
[4] Twitter (June 2020). Our range of enforcement options. https://help.twitter.com/en/rules-and-policies/enforcement-options.

engaging in some form of platform manipulation.[5] We have also implemented mandatory email or phone verification for all new accounts.

Rules Prohibiting Attributed Activity

We know that certain groups and individuals engage in persistent, organised 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.[6] For instance, if we are able to identify activity associated with the Russian Internet Research Agency (IRA), 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.

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.[7] 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 sections of the media distribute the contents of a hack through their own reporting. These actions potentially achieve

[5] Twitter (June 2019). Transparency Report. https://transparency.twitter.com/en/platform-manipulation.html.

[6] Twitter Safety (12 June 2020). Disclosing networks of state-linked information operations we've removed. https://blog.twitter.com/en_us/topics/company/2020/information-operations-june-2020.html.

[7] Twitter (June 2020). Distribution of hacked materials policy. https://help.twitter.com/en/rules-and-policies/hacked-materials.

the aim of the hostile actor to amplify a desired message to large audiences in spite of Twitter's efforts to remove offending accounts.

Political Advertising Policy and State-Controlled Media Advertising Policy

Twitter continues to evolve and adapt its advertising policies to protect the integrity of our service and ensure easier access to credible information on Twitter. We believe that there is a difference between engaging in conversations with accounts you choose to follow and the content you see from advertisers in your Twitter experience, which may be from accounts you're not currently following. We have policies for both, but we have higher standards for our advertisers.

There are two such recent changes also pertinent to discouraging and mitigating foreign interference. First, Twitter now globally prohibits the promotion of political content on the platform. The company made the decision in October 2019 that such political messaging reach should be earned, not bought. Political entities and individuals are allowed to have a presence on Twitter. But we do not allow ads of any type by political or electoral candidates, political parties or elected or appointed officials. We define political content as content that references a candidate, political party, elected or Government official, election, referendum, ballot measure, legislation, regulation, directive or a judicial outcome.[8]

Second, also in 2019 Twitter determined that news media entities controlled by state authorities may not purchase advertisements.[9] Any affected accounts are free to continue to use Twitter to engage in public conversation, just not our advertising products. This policy also extends to individuals reporting on behalf of, or who are directly affiliated with, such entities as well. When we define state-controlled media in this context, we look at the following criteria:

- Control by state authorities entails lack of financial or editorial independence. Independence means autonomous control over editorial content by publishers, broadcasters, editors, and journalists without undue interference from state authorities through financial resources, direct or indirect political pressures, and/or control over production and distribution.
- State authorities mean government bodies and institutions.
- News media refers to all relevant sources of news and commentary.
- This policy does not apply to taxpayer-funded entities, including independent public broadcasters.

[8] Twitter (June 2020). Twitter for Business Political content FAQs. https://business.twitter.com/en/help/ads-policies/proh bited-content-policies/political-content/faqs.html.

[9] Twitter (19 August 2019). Updating our advertising policies on state media. https://blog.twitter.com/en_us/topics/company/2019/advertising_policies_on_state_media.html.

The determinations of those entities and their affiliates that may fall under this policy are made on the basis of critical issues of media freedom and independence, such as control of editorial content, financial ownership influence, or interference over broadcasters, editors, and journalists, direct and indirect exertion of political pressure and/or control over the production and distribution process. The policy is also informed by established academic and civil society leaders in this space globally.

Products that Safeguard the Conversation

Approach to Misinformation

At Twitter, we prioritise 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 input has guided our work.[10] 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.[11]

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, which is open, real-time, and conversational. Our overarching goal is to provide context, not fact-check.[12]

We prioritise based on the highest potential for harm, focusing exclusively on high-profile cases of manipulated media, civic integrity, and COVID-19 at this time. Likelihood, severity, and type of potential harm — along with reach and scale — factor into these decisions. Due to

[10] Harvey, D. (11 November 2019). Help us shape our approach to synthetic and manipulated media. https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html.

[11] Achuthan, A. and Roth, Y. (4 February 2020). Building rules in public: Our approach to synthetic & manipulated media. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html.

[12] Pickles, N. and Roth, Y. (11 May 2020). Updating our Approach to Misleading Information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.

the large potential reach and persuasive impact of media content, we started with a policy on manipulated media.

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.

Synthetic and Manipulated Media

We have closely tracked the challenges associated with so-called 'deep fake' 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 11 November 2019, we released a draft of our rules governing synthetic and manipulated media that purposely attempts to mislead or confuse people.[13] 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 4 February 2020, we announced Twitter's policy on synthetic and manipulated media.[14] Under our Twitter 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.[15]

When applying this policy, we review a number of criteria to evaluate 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.[16]

---

[13]Harvey, D. (11 November 2019). Help us shape our approach to synthetic and manipulated media. https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html.

[14] Twitter (June 2020). Synthetic and manipulated media policy. https://help.twitter.com/en/rules-and-policies/manipulated-media.

[15] Achuthan, A. and Roth, Y. (4 February 2020). Building rules in public: Our approach to synthetic & manipulated media. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html.

[16] *Ibid*.

Second, we evaluate whether the media are shared in a deceptive manner. Under this review, we also consider: (1) whether the context in which media is shared could result in confusion or misunderstanding; or (2) 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; (1) the metadata associated with the media; (2) the information on the profile of the person sharing the media; and (3) 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.

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 our public archive — the only one in the industry to date.

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.[17] To date, it is the only public archive of its kind, now spanning operations across 15 countries, and

[17] Gadde, V. and Roth, Y. (17 October 2018). Enabling further research of information operations on Twitter. https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html.

includes more than nine terabytes of media and 200 million Tweets.[18] 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 and work we are committed to continuing into the future.

Partnerships with Stakeholders and Governments

Information sharing and collaboration are critical to Twitter's success in preventing hostile foreign actors from disrupting meaningful political conversations on the service. We believe these data sets are valuable for research, transparency, and accountability. Going forward, we expect to be communicating this information more widely to meet our goal of serving the public conversation, remove bad faith actors, and to advance public understanding of these critical topics. In the future we're going to offer more clarity in the public archive around impression counts and attempt to further measure the tangible impact of information operations on the public conversation, as well as continue to formalise our academic partnerships to ensure they're globally diverse and advancing public understanding of these issues.[19]

As an example of our efforts, we partnered with the Australian Strategic Policy Institute (ASPI) and the Stanford Internet Observatory (SIO) to provide them with advance access to the data and enable independent research from subject matter experts to provide analysis and insights to accompany the data disclosure as part of our recent disclosure of state-linked information operations.[20]

We have well-established relationships with relevant Australian law enforcement agencies, including the Australian Department of Home Affairs, the Australian Security Intelligence Organisation, the Australian Federal Police, and the Australian Electoral Commission. We look forward to continued cooperation with federal, state, and local government agencies on foreign interference and election integrity because in certain circumstances, only they have access to information critical to our joint efforts to stop bad faith actors.

Additionally, we have significantly deepened our partnership with industry peers, establishing formal processes for information sharing and a regular cadence of discussion about shared

[18] Twitter (June 2019). Transparency Report. https://transparency.twitter.com/en/information-operations.html.
[19] Twitter Safety (12 June 2020). Disclosing networks of state-linked information operations we've removed. https://blog.twitter.com/en_us/topics/company/2020/information-operations-june-2020.html.
[20] *Ibid.*

threats.[21] This collaboration is a critical part of our efforts to mitigate malicious activity that is not restricted to a single platform or service.

Overview of Twitter's Response to COVID-19

Based on our policies, rules, and principles outlined above, we have employed concerted internal and external efforts to help people find authoritative health information, build partnerships, raise relief funds, contribute pro bono advertising support, and ultimately protect the public conversation to ensure people are getting the right message from the right source throughout the progression of the global COVID-19 pandemic.[22]

With a critical mass of expert organisations, official government accounts, health professionals, and epidemiologists on our service, our goal is to elevate and amplify authoritative health information as far as possible.

Providing Access to Credible Information

Global expansion of the COVID-19 search prompt

Launched six days before the official designation of the virus in January 2020, we continue to expand our dedicated search prompt feature to ensure that when people come to Twitter for information about COVID-19, they are met with credible, authoritative content at the top of their search experience. We have been consistently monitoring the conversation on the service to make sure keywords, including common misspellings, also generate the search prompt. Additionally, we have halted any auto-suggest results that are likely to direct individuals to non-credible content on Twitter.[23]

In each country where we have launched the initiative, we have partnered with the national public health agency or the World Health Organization (WHO) to assist in the search prompt activation.[24] The proactive search prompt is in place with official local partnerships - 90 of them in 79 countries and in 30 local languages across all continents. In Australia, we partnered with the Department of Health to direct people to their pandemic webpage, and then updated the search prompt with the Australian Government's dedicated COVID-19 website as the government's response progressed. We also provide related resources from the WHO for Australian users.

[21] Technology Coalition (June 2020). https://www.technologycoalition.org/.

[22] Twitter (June 2020). Coronavirus: staying safe and informed on Twitter. https://blog.twitter.com/en_us/topics/company/2020/covid-19.html.

[23] Chu, J. and McDonald, J. (29 January 2020). Helping the world find credible information about novel #coronavirus. https://blog.twitter.com/en_us/topics/company/2020/authoritative-information-about-novel-coronavirus.html.

[24] @WHO, https://twitter.com/WHO.

Figure 1. Screenshot of COVID-19 Twitter search prompt directing people to the Australian Government COVID-19 website and WHO for resources.

Working with the Federal Minister for Communications and the Australian Department of Infrastructure, Transport, Regional Development and Communications, we also launched a search prompt in Australia that responds to queries related to 5G and COVID-19. The prompt directs people searching for information about 5G and COVID-19 to authoritative sources of information from the Department of Health, and informs them that the Australian Government has not found evidence of a link between 5G and coronavirus.

Figure 2. Screenshot of COVID-19 and 5G Twitter search prompt directing people to the Australian Department of Health and Australian Government COVID-19 Twitter account for resources.

Expanding Topics during COVID-19

Twitter Topics is a setting that allows users to select and follow trends they're interested in from a list of Topics. Once a user follows a Topic, more news and Tweets related to it will start showing up in their regular Twitter feed.[25]

Many people come to Twitter to stay informed on their interests, and we want to give them a way to easily see all the things they care about when browsing their Home timeline. People have an additional need when it comes to interests which more directly impact their health, families, communities, and lives in that they want to be sure that they can trust the information

[25] Twitter (June 2020). Follow topics. https://help.twitter.com/en/using-twitter/follow-and-unfollow-topics.

they are seeing. With this in mind, we created a COVID-19 followable Topic so that people can stay informed on coronavirus right from their Home timeline.

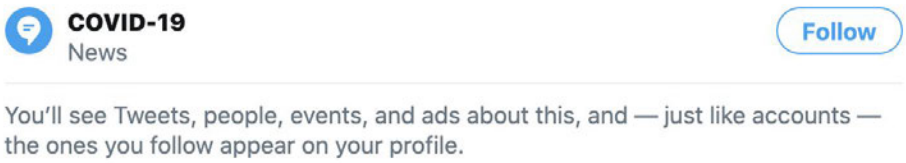

Figure 3. Screenshot of COVID-19 followable Topic where people can quickly follow credible Twitter accounts for resources and information related to the pandemic.

Verifying Authoritative Accounts

While our public verification program has been paused,[26] Twitter continues to verify official government accounts, media, expert organisations, health professionals, and epidemiologists in local markets with a blue badge.[27] We have worked with local healthcare organisations, universities, and the online community to verify health experts and doctors in Australia who provide local advice and information.

Curated Event Feature for COVID-19

We also launched a global Events feature[28] and a localised Australian Event feature[29] containing credible information and the latest facts about COVID-19. This feature assembles the most recent Tweets of credible sources, including the Australian and State Departments of Health, relevant state and federal health ministers, international agencies, news and media outlets, doctors, health professionals, academics, and researchers, etc., to provide up to date information.

It is available at the top of the home timeline for everyone in Australia in the Explore tab. While the Events feature is continually updated, we will also share specific platform and policy updates from our Twitter owned and operated handle @TwitterSafety too.[30]

[26] Twitter (June 2020). About verified accounts. https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.
[27] @TwitterSupport (21 March 2020). https://twitter.com/TwitterSupport/status/1241155701822476288
[28] @Twitter (June 2020). https://twitter.com/i/events/1219057585707315201
[29] @Twitter (June 2020). https://twitter.com/i/events/1242311410245222400
[30] @TwitterSafety, https://twitter.com/TwitterSafety

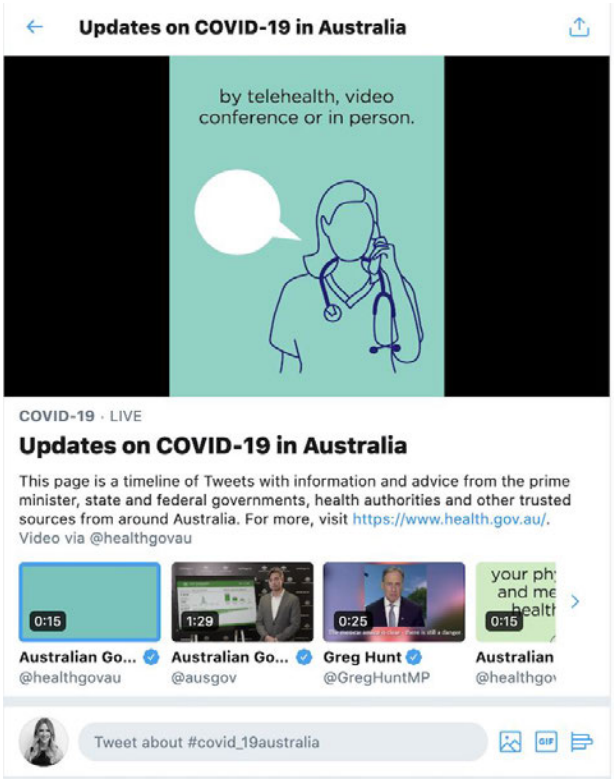

Figure 4. Screenshot of COVID-19 Australian Event feature containing updates from credible government organisations, officials, media outlets, and more.

Furthering our Partnerships

Our global Public Policy team has open lines of communication with relevant multinational stakeholders, including the WHO, numerous global government and public health organisations, and officials around the world, to ensure they can troubleshoot account issues, get their experts verified, and seek strategic counsel as they use the power of Twitter to mitigate harm.

We're also in close contact with our industry peers and attend relevant cross-functional forums and meetings.[31] As people across the country and around the world #StayHome to slow the spread of COVID-19, it's imperative we think about the significant impact these measures have on those experiencing substance use disorders and those in recovery. For example, in partnership with the Center for Safe Internet Pharmacies,[32] Twitter is building on its collaboration with Google, Facebook, and Microsoft to showcase a variety of resources on TechTogether.co to support the #RecoveryMovement.[33] The site is a collection of resources to

[31] @TwitterComms (17 March 2020). https://twitter.com/TwitterComms/status/1239712022096297985
[32] @Safemedsonline, https://twitter.com/safemedsonline
[33] TechTogether (June 2020). http://techtogether.co/

help those experiencing a substance use disorder and any associated stigmas and helping provide resources for those battling substance abuse disorders in recovery.

#BuildforCOVID19 Global Online Hackathon and #DataforGood

Twitter also joined our industry peers organising a Global Online Hackathon around COVID-19. This initiative is an opportunity for developers to build software solutions that drive social impact, with the aim of tackling some of the challenges related to the coronavirus pandemic.[34]

As a uniquely open service, Twitter data is being accessed via our public API every day, and our research hub is publicly available. Through our ongoing #DataforGood partnership with the UN Global Pulse Lab, agencies such as UNICEF and WHO are also able to access Twitter's Enterprise APIs to conduct further public health research.[35] We are also exploring further #DataForGood partnerships to assess how our data products can enhance academic and NGO understandings of public health emergencies now and into the future.

Enabling Research of the Public Conversation in a Time of Crisis

As a uniquely open service, our data is being used in research every day and our researchers hub is publicly available.[36] To further support Twitter's ongoing efforts to protect the public conversation and help people find authoritative health information around COVID-19, we released a new endpoint in Twitter Developer Labs to enable approved developers and researchers to study the public conversation about COVID-19 in real time.[37]

This is a unique dataset that covers many tens of millions of Tweets daily and offers insight into the evolving global public conversation surrounding an unprecedented crisis. The conversations happening on Twitter are highly insightful and have the potential to help the world better understand the COVID-19 pandemic and ultimately contribute to making Twitter, and the world, a better place through use cases like:

- Researching the spread of the disease
- Understanding the spread of misinformation
- Crisis management, emergency response, and communication within communities
- Developing machine learning and data tools that can help the scientific community answer key questions about COVID-19
- And many others that we'll look to our community to help us understand

[34] COVID-19 Global Hackathon (26 March 2020). https://covid-global-hackathon.devpost.com/
[35] Crowell, C. (2016). Twitter and UN Global Pulse Lab announce data partnership. https://blog.twitter.com/en_us/a/2016/twitter-and-un-global-pulse-announce-data-partnership.html
[36] Twitter Developer (June 2020). https://developer.twitter.com/en/use-cases/academic-researchers
[37] Tornes, A. (29 April 2020). Enabling study of the public conversation in a time of crisis. https://blog.twitter.com/developer/en_us/topics/tools/2020/covid19_public_conversation_data.html

Making this access available for free is one of the most unique and valuable things Twitter can do as the world comes together to protect our communities and seek answers to pressing challenges. We hope this data can produce unprecedented research on how we collectively respond to the challenges of a global pandemic and to provide learnings that allow humanity to better prepare in the future.

Countering Misinformation

Protecting the Conversation During COVID-19

The power of a uniquely open service during a public health emergency is clear. The speed and borderless nature of Twitter presents an extraordinary opportunity to communicate in real time and ensure people have access to the latest information from expert sources around the world. Journalists, experts, and engaged citizens Tweet side-by-side correcting and challenging public discourse second by second. These vital interactions happen on Twitter every day, and we are working to ensure that we surface the highest quality and most relevant content and context first.

Twitter has a zero-tolerance approach to the artificial amplification of public health misinformation content on our platform, as well as any attempt to abuse or manipulate our service.[38] We continue to invest in detection tools and technology to combat malicious automation and manipulation of our service. This investment has yielded positive results – we have seen a 50 percent drop in challenges to accounts for suspected breaches of our platform manipulation policy.[39]

We're not seeing significant coordinated platform manipulation efforts around these issues in Australia. We also understand there's been a lot of discussion of "bots" online recently, which has become a loaded and often misunderstood term.

People often refer to bots when describing everything from automated account activity to individuals who would prefer to be anonymous for personal or safety reasons. The term is used to mischaracterise accounts with numerical usernames that are auto-generated when you create an account, and more worryingly, as a tool by those in positions of political power to tarnish the views of people who may disagree with them or counter online public opinion that's not favorable.

There are also many commercial services that purport to offer insights on bots and their activity online, and frequently their focus is entirely on Twitter due to the free data we provide through

[38] Twitter (June 2020). Platform manipulation policy. https://help.twitter.com/en/rules-and-policies/platform-manipulation
[39] Twitter (June 2019). Transparency report. http://transparency.twitter.com

our public APIs. Unfortunately, this research is rarely peer-reviewed and often does not hold up to scrutiny further confusing the public's understanding of what's really happening.[40]

In order to counter misinformation, especially in the context of COVID-19, we adopted the following measures to protect our platform:[41]

- **Increasing our use of machine learning and automation** to take a wide range of actions on potentially abusive and manipulative content. We want to be clear: while we work to ensure our systems are consistent, they can sometimes lack the context that our teams bring, and this may result in us making mistakes. As a result, we will not permanently suspend any accounts based solely on our automated enforcement systems. Instead, we will continue to look for opportunities to build in human review checks where they will be most impactful.

- **Building systems that enable our team to continue to enforce our rules remotely around the world.** We're increasing our employee assistance and wellness support for everyone involved in this critical work, and ensuring people's privacy and security stay a top priority.

- **Instituting a global content severity triage system** so we are prioritising the potential rule violations that present the biggest risk of harm and reducing the burden on people to report them.

- **Executing daily quality assurance checks** on our content enforcement processes to ensure we're agile in responding to this rapidly evolving, global issue.

- **Engaging our partners around the world** to ensure escalation paths remain open and urgent cases can be brought to our attention.

- **Reviewing the Twitter Rules in the context of COVID-19** and considering ways in which they may need to evolve to account for new account behavior.[42]

- **Broadening our definition of harm** to address content that goes directly against guidance from authoritative sources of global and local public health information. Rather than reports, we are enforcing this in close coordination with trusted partners, including public health authorities and governments, and continue to use and consult with information from those sources when reviewing content.

[40] Pickles, N. and Roth, Y. (18 May 2020). Bot or not? The facts about platform manipulation on Twitter. https://blog.twitter.com/en_us/topics/company/2020/bot-or-not.html
[41] Twitter (31 October 2019). 15th Transparency Report: Increase in proactive enforcement on accounts. https://blog.twitter.com/en_us/topics/company/2019/twitter-transparency-report-2019.html
[42] Twitter (June 2020). http://twitter.com/rules

- We'll continue to prioritize removing content when it has a clear call to action that could directly pose a risk to people's health or well-being, but we want to make it clear that we will not be able to take enforcement action on every Tweet that contains incomplete or disputed information about COVID-19. This is not meant to limit good faith discussion or expressing hope about ongoing studies related to potential medical interventions that show promise.
- We may also apply the public interest notice[43] in cases where world leaders[44] violate the COVID-19 guidelines.

In addition to our proactive work, we also have the option for people who see suspicious activity to report it to us.[45] We add these signals to the hundreds of others we use to inform our technical approach and make our service better.

We remain vigilant and will continue to substantially invest in our abilities to ensure that trends, search, and other common areas of the service are protected from malicious behaviors. As ever, we also welcome constructive and open information sharing from governments and academics to further our work in these areas.

Automated Technology during COVID-19

We are using automated technology to help us review reports more efficiently by surfacing content that's most likely to cause harm and should be reviewed first. Additionally, automation helps us proactively identify rule-breaking content before it's reported. Our systems learn from past decisions by our review teams, so over time, the technology is able to help us rank content or challenge accounts automatically.

For content that requires additional context, such as misleading information around COVID-19, our teams will continue to review those reports manually. What you can expect if you file a report during this time: (1) if you've reported an account or Tweet to us, it will take longer than normal for us to get back to you, and we appreciate patience as we continue to make adjustments; and (2) because these automated systems don't have all of the context and insight our team has, we'll make mistakes. If you think we've made a mistake, we have made it possible for people to appeal using our online form.[46]

Ads Policies related to COVID-19

Based on our Inappropriate Content Policy, Twitter prohibits advertising content that is inflammatory or provocative and is likely to provoke strong negative reactions.[47] We will halt

[43] Twitter (15 October 2019). World Leaders on Twitter. https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html
[44] Twitter Safety (27 June 2019). Defining public interest on Twitter. https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html
[45] Twitter (June 2020). Spam and fake accounts policy. https://help.twitter.com/en/safety-and-security#spam-and-fake-accounts
[46] Twitter (June 2020). https://help.twitter.com/forms/general
[47] Twitter (June 2020). Twitter for Business, Political content policy.
https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/inappropriate-content.html

any attempt by advertisers to opportunistically use the COVID-19 outbreak to target inappropriate ads. We also prohibit ads that contain inappropriate content globally.[48] This includes ads relating to sensitive topics or events (e.g. deaths, natural or industrial disasters, violent attacks, etc.)

In response to the shifting advertising landscape, and in order to support helpful causes during this time, we made adjustments to allow managed clients and partners to advertise content containing implicit or explicit reference to COVID-19 in the following use cases, with restrictions:

- Adjustments to business practices and/or models in response to COVID-19
- Support for customers and employees related to COVID-19

The following restrictions apply to these use cases:

- Distasteful references to COVID-19 (or variations) are prohibited
- Content may not be sensational or likely to incite panic
- Prices of products related to COVID-19 may not be inflated
- The promotion of certain products related to COVID-19 may be prohibited. We currently prohibit the advertising of facemasks and alcohol hand sanitizers. Please note that other products may be added to this list and enforcement can be retroactive.
- The mention of vaccines, treatments and test kits is permitted, only in the form of information, from news publishers which have been exempted under the Political Ads Content policy.[49]

Government entities that want to disseminate public health information are permitted to promote ads on COVID-19, including the Australian and State Departments of Health and relevant official government COVID-19 accounts.

In the case of COVID-19, we have put additional safeguards into place in order to facilitate the sharing of trusted public health information and to reduce potential harm to users. We are currently prohibiting the promotion of all medical masks and alcohol hand sanitisers due to strong correlation to COVID-19 and instances of inflated prices globally, and we will continue to monitor and evaluate the need for further restrictions going forward.

Proactive Enforcement and Spam Detection

Since introducing our updated policies on March 18, we have removed thousands of Tweets around the globe for containing misleading and potentially harmful content from Twitter. Additionally, our automated systems have challenged approximately 4.3 million accounts

[48] *Ibid.*
[49] *Ibid.*

which were targeting discussions around COVID-19 with spammy or manipulative behaviors. We will continue to use both technology and our teams to help us identify and stop spammy behavior and accounts.

Philanthropic initiatives and awareness campaigns

- Emoji #SafeHands
  In partnership with the WHO, Twitter launched a hashtag emoji to support the #SafeHands campaign. The emoji was activated when people use one of the following hashtags: #SafeHands #WashYourHands #HandWashChallenge #HandWashingChallenge #HandWashing #SafeHandsChallenge
- World Health Day
  To celebrate World Health Day (7 April), and all the health workers keeping us safe during this crisis, Twitter launched an emoji that was activated through the following hashtags: #WorldHealthDay #StayHome #StayAtHome #HealthyAtHome #StayAtHomeSaveLives.

On June 30 for #WorldSocialMediaDay, in line with our work to promote authoritative information on public health and civic integrity issues, we provided pro bono Ads for Good[50] advertising grants to support the UN's Pause campaign to encourage people to #TakeCareBeforeYouShare to help stop the spread of misinformation about COVID-19.[51]

On April 7 led by our official @Twitter account, we also asked people to share their stories or raise their hands if they have friends and family in healthcare so the world could thank them. At the same time, since there have been more than six million questions Tweeted about COVID-19, for #WorldHealthDay, @Twitter facilitated a Q&A with the @WHO on our platform to provide answers to some of the most frequently asked questions.[52]

Figure 5. Screenshot of @Twitter hosted Q&A with @WHO.

[50] Twitter Marketing (14 December 2011). Twitter Ads for Good. https://blog.twitter.com/en_us/a/2011/twitter-ads-for-good.html.
[51] @Policy (30 June 2020). https://twitter.com/Policy/status/1278095924330364935?s=20.
[52] @Twitter (8 April 2020). https://twitter.com/Twitter/status/1247542368514887690?s=20.

Protecting and Supporting Journalists

Right now, every journalist is a COVID-19 journalist. Since journalism is core to our service, and we have a deep and enduring responsibility to protect that work, Twitter contributed to two critical global organisations that are working tirelessly to uphold the fundamental values of a free press during this pandemic.

We donated one million dollars evenly distributed between the Committee to Protect Journalists and the International Women's Media Foundation. These funds will be used to ensure these organisations can continue their work in the face of new economic strains and to directly support journalists. Their shared efforts to advocate for the rights of vulnerable reporters and to guarantee an equal share of voice for women in the industry has never been more relevant or important.

In Australia, we also partnered locally with the Walkley Foundation to provide a twenty-five thousand dollar cash grant and a twenty-five thousand dollar pro bono Ads for Good grant to support the production of public interest journalism in remote and regional areas.[53]

Helping Brands Communicate During the Crisis

We have also been working with brands and partners to help them respond to COVID-19. The pandemic is not a marketing opportunity to capitalise on, and we do not recommend brands opportunistically linking themselves to a health scare. However, we want to recognise that this is a new reality and requires thoughtful navigation from all of us moving forward.

We also know that Twitter is a platform that plays a significant role in crisis communications, and can be a powerful tool for governments and companies alike to communicate with their citizens, customers, employees, and the broader ecosystem.[54]

We have worked closely with our partners to anticipate changes in behaviors and patterns, as well as ensure they are communicating accurate and reliable information. In times of crises, people want credible information. We've seen that verified people on Twitter are about 2.4x more likely to participate in COVID-19 conversation than non-verified people, and 75 percent of COVID-19 related Tweets are actually Retweets.[55] In other words, the primary method of sharing information during a time of crisis is through Retweeting.

[53] Wa kley Foundation (June 2020). Walkley Grants for freeland regional journalism. https://www.walkleys.com/grants/walkley-grants-for-freelance-regional-journalism/
[54]Twitter Marketing (June 2020). Twitter ad performance and research for COVID-19. https://marketing.twitter.com/na/en/insights/twitter-ad-performance-research-covid19
[55]Twitter (June 2020). Coronavirus: staying safe and informed on Twitter. https://blog.twitter.com/en_us/topics/company/2020/covid-19.html.

As part of an industry working group convened by the Minister for Industry, Science and Technology and the Department of Industry, Science, Energy and Resources, we have contributed resources for COVID-19 response efforts. Additionally, our marketing and brand strategy teams have provided trainings for brands, which contain guidance for businesses that possess useful and reliable information that might help people navigate the uncertainty. For example, retail or e-commerce brands can keep the public updated on stock to help mitigate panic buying. Additionally, when it comes to setting policies and supporting their employees through this uncertain period, there is truly an opportunity to lead by example and get others to follow in their steps.

Figure 6. Screenshot of @Channel4 Tweet.

In times like this, when the newscycle can be overwhelming, a bit of light distraction and levity can also go a long way. As humans, we're programmed to seek out connection. While social distancing measures can offer protection and security, they come with significant impact on individuals, communities, and the world at large.

We have also partnered with mental health partners in Australia, like Beyond Blue, Reachout Australia, and Lifeline Australia to provide advertising grants and amplification through our Twitter owned and operated handles. We've seen non-COVID-19 related positive stories capture people's attention as good things continue to happen, despite the context right now. We encourage organisations and companies to continue to connect with and celebrate these moments, as and when appropriate.

As the global community continues to face the COVID-19 pandemic together, Twitter will continue to help people find reliable information, connect with others, and follow what's happening in real time.

Conclusion

As detailed previously, the threat of foreign and domestic interference is one that has evolved since 2016 with new tactics, tools, and vulnerabilities. We have prepared accordingly.

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.

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.

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 around the world engage in discussions 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, and a wide range of stakeholders need to play their part too.

Now is the time to unleash the best of democratic values and encourage courageous communication in every sense. A well-informed public at home is still the strongest response to foreign or domestic threats. We have never been more focused on taking enforcement action against hostile actors and fostering an environment conducive to healthy, meaningful dialogue, and we look forward to working with the Committee on these vital issues.

Further information

For more information regarding internal measures we have taken to protect our employees, contractors, clients, and partners during the pandemic, please visit our blog posts:

- Our contingency strategy to protect the conversation
- Our working guidance to our employees and partners to keep them safe
- Our partnerships and public engagement strategies
- Protecting and supporting journalists during COVID-19
- Helping the world find credible information about novel #coronavirus