



9 February 2022

Committee Secretary
Australian Select Committee on Social Media and Online Safety
PO Box 6021
Parliament House
Canberra ACT 2600

By email: smos.reps@aph.gov.au

Dear Chair,

Thank you for the opportunity to provide responses to questions taken on notice following Twitter's appearance as part of the Australian Select Committee on Social Media and Online Safety inquiry into social media and online safety.

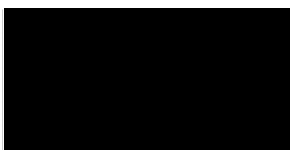
We have endeavoured to provide answers to the best of our ability within the time allotted for the questions outlined by the Committee below.

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

Please don't hesitate to let us know if there is any additional information we can provide to assist the Committee during the course of this inquiry.

Thank you again for the opportunity to provide input as part of this important process.

Kind regards,



Kara Hinesley
Director of Public Policy
Australia and New Zealand



Kathleen Reen
Senior Director of Public Policy
Asia Pacific



Australian Select Committee on Social Media and Online Safety Inquiry into social media and online safety

Answers to Questions on Notice Twitter Inc.

1. [Last year], a New York Times article detailed how the Chinese Communist Party (CCP) monitors Twitter to identify and then intimidate and harass critics offshore, particularly those with relatives still in China, pressuring them to remove criticism on platforms like Twitter. This story cites a Chinese police manual that 'specifically called for monitoring activity on websites' and reports they're then intimidated by authorities from China, often using their families. The article states that: *Actions against people for speaking out on Twitter and Facebook have increased in China since 2019, according to an online database aggregating them. Are you aware of this practice, and is there anything you're doing in response to it?*

...there was one specific story in this New York Times article that did concern me and that I am keen to get your response to. The article described one account of a 23-year-old Chinese student in Australia who received a video call from police using her father's phone. These police then pushed her to remove her Twitter account. This was dealing with a parody account, not in her own name, that she'd created.

The article described the situation like this:

This time, calling her via video chat, they told her to report to the station when she returned to China and asked how much longer her Australian visa was valid. Fearful, she denied owning the Twitter account but filmed the call and kept the account up. A few months later, Twitter suspended it.

After an inquiry from The Times, Twitter restored the account. A Twitter spokeswoman said it had been taken down in error. Can you explain what happened here?

—Deputy Chair Tim Watts MP

The Twitter account referenced in the New York Times article first published on 31 December 2021 and updated on 1 January 2022 entitled “A Digital Manhunt: How Chinese Police Track Critics on Twitter and Facebook,” was suspended in error by an incorrect enforcement action and has since been restored.¹

We take action on accounts when they violate the Twitter Rules or Terms of Service, and our view is that those apply equally to everyone. We do not succumb to pressure to take down accounts that are not in violation of our publicly stated policies.

The Chinese Government has blocked access to Twitter in China. Our most recent Twitter Transparency Report shows that in the latest reporting period, we have not disclosed any account information to Chinese authorities, nor removed content on the basis of a legal demand from Chinese authorities.²

State-backed information operations

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

Our data disclosures on such operations are part of Twitter’s efforts to contribute to such independent analysis. We also believe that independent analysis of this activity by researchers is a key step toward promoting shared understanding of these threats.

In line with this commitment, as of January 2022, we have proactively shared 48 datasets from 22 countries, which contain over 200 million Tweets and 9 terabytes of media.³

¹ <https://www.nytimes.com/2021/12/31/technology/china-internet-police-twitter.html>

² <https://transparency.twitter.com/en/reports/countries/cn.html>

³ <https://transparency.twitter.com/en/reports/information-operations.html>



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

In June 2020, we shared relevant data with research partner, the Australian Strategic Policy Institute (ASPI), who published the report *Retweeting through the Great Firewall* utilising these data sets.⁴ Their analysis looked at a dataset of over 23,000 Twitter accounts and almost 350,000 Tweets that occurred from January 2018 to April 2020, which we attributed to Chinese state-linked actors and took the accounts offline.

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

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

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

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

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

Similar studies support ASPI report's findings. Graphika has undertaken two studies of a persistent campaign targeting the Hong Kong protests, Guo Wengui (Miles Kwok) and other critics of the Chinese Government.⁵ Researchers at Bellingcat have also previously reported on networks targeting Guo Wengui and the Hong Kong protest movement.⁶ Additional research from Miburo's Nick Monaco was recently released in December 2021 that examined this campaign regarding "Spamouflage," a well-known disinformation actor aligned with the CCP, identified previously in a 2019 Graphika report. The report explicitly states that: "Spamouflage, though a well-known actor with long-established tactics, techniques, and procedures (TTPs), continues to proliferate on Facebook and YouTube. Accounts are typically actioned much more quickly on Twitter."⁷

As part of our continued commitment to address these issues, in December 2021 we also shared relevant data disclosures with ASPI related to the removal of a network of accounts that amplified Chinese Communist Party narratives related to the treatment of the Uyghur population in Xinjiang.⁸ As part of this data disclosure, we released a representative sample of 2,048 accounts and also removed a network of 112 accounts connected to "Changyu Culture," a private company backed by the Xinjiang regional government.

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

- 2. Are you concerned that there's potential for, let's call it, foreign interference, manipulating your moderation and trust and safety policies? It does seem a disturbing coincidence that this account [see question above] was attracting that kind of attention from these authorities and was then taken down in this way.**

— Deputy Chair Tim Watts MP

⁴ <https://www.aspi.org.au/report/retweeting-through-great-firewall>

⁵ https://public-assets.graphika.com/reports/Graphika_Report_Spamouflage_Returns.pdf

⁶ <https://www.bellingcat.com/news/2020/05/05/uncovering-a-pro-chinese-government-information-operation-on-twitter-and-facebook-analysis-of-the-milesquo-bot-network/>

⁷ <https://miburo.substack.com/p/spamouflage-survives>

⁸ https://blog.twitter.com/en_us/topics/company/2021/disclosing-state-linked-information-operations-we-ve-removed



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

Platform manipulation

People are not permitted to use Twitter in a manner intended to artificially amplify, suppress information, or engage in behaviour that manipulates or disrupts other people's experience on the service.

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

Some of the factors that we take into account when determining whether an account is fake include the use of stock or stolen avatar photos; the use of stolen or copied profile bios; and the use of intentionally misleading profile information, including profile location.

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

We prioritise identifying suspicious account activity, such as exceptionally high-volume Tweeting with the same hashtag or mentioning the same @handle without a reply from the account being addressed. When we identify such activity, we require an individual using the service to confirm human control of the account or their identity.

We have increased our use of challenges intended to catch automated accounts, such as reCAPTCHAs (that require individuals to identify portions of an image or type words displayed on screen), and password reset requests that protect potentially compromised accounts.

In our most recent Transparency Report, we challenged over 130 million accounts for engaging in suspected spammy behaviour, including those engaged in suspected platform manipulation.⁹ We have also implemented mandatory email or phone verification for all new accounts.

Since 2018, we also introduced a registration process for developers requesting access to our application programming interfaces (APIs) to prevent the registration of spammy and low quality apps, and we are continuing to roll out improvements to our proactive enforcements against common policy violations.¹⁰

Automation and automated accounts

People often refer to bots when describing everything from automated account activity to individuals who would prefer to be pseudonymous for personal or safety reasons, or avoid a photo because they've got strong privacy concerns.

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

It's important to note, however, that not all forms of automation are violations of the Twitter Rules. We've seen innovative and creative uses of automation to enrich the Twitter experience. For example, accounts that track air quality, earthquakes, or general reminders to drink your water like @tinycarebot.¹¹

Automation can also be a powerful tool, like a conversational bot that can help find information about orders or voting information, like the Twitter Direct Message Chatbot we set up for the 2019 Australian Election.¹² This kind of innovative tooling has proved safe and efficient for a myriad of civic and corporate functions, especially at a time of social distancing.

The presence of a bot account on Twitter is not an indication that content from that user is shown or distributed in the same way as organic content — and our actions to limit the spread of spammy or automated content are not

⁹ <https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jul-dec>

¹⁰ https://blog.twitter.com/developer/en_us/topics/tips/2018/automation-and-the-use-of-multiple-accounts

¹¹ <https://twitter.com/tinycarebot>

¹² https://blog.twitter.com/en_au/topics/company/2019/get--ausvotes2019-election-information-through-twitter-



available to developers or researchers using the public APIs. The end user experience of someone using the Twitter app or website is not replicated by looking at an unfiltered stream of content obtained via our public API.

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

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

Commitment to privacy

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

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

We believe that individuals should control the personal data that is shared with companies and provide them with the tools to help them control their data. Through the account settings on Twitter, we give people the ability to make a variety of choices about their data privacy, including limiting the data we collect, determining whether they see interest-based advertising, and controlling how we personalise their experience.¹³ In addition, we provide them with the ability to access information about advertisers that have included them in tailored audiences to serve them ads, demographic and interest data about their account from ad partners, and information Twitter has inferred about them.

- 3. I would just like to ask you, specifically given that it goes to gender, whether comments by a Twitter user calling a woman a 'cavorting whore' would meet the threshold for your hateful conduct policy? I also ask you, in relation to 'I hate this f-ing b-i-t-c-h, pinko slut b-i-t-c-h', whether that is considered to be hateful conduct policy or not. Does that breach your standards? What action would be taken in relation to a comment such as that against a person posting this? I will ask a series of others as well, but those two are specific.**

– Chair Lucy Wicks MP

When it comes to reviewing content or accounts for violating the Twitter Rules, context matters, and there are a range of different factors our team considers. For example: is the behaviour directed at an individual, or group of people? Has the report been filed by the target of the potential abuse or a bystander? Does the person have a history of violating our policies? What is the severity of the violation?

With regards to abusive behaviour, our rules state that users may not engage in the targeted harassment of someone, or incite other people to do so. We consider abusive behaviour an attempt to harass, intimidate, or silence someone else's voice.¹⁴

We believe in freedom of expression and open dialogue, but that means little as an underlying philosophy if voices are silenced because people are afraid to speak up. In order to facilitate healthy dialogue on the platform, and empower individuals to express diverse opinions and beliefs, we prohibit behaviour that harasses or intimidates, or is otherwise intended to shame or degrade others. In addition to posing risks to people's safety, we understand that abusive behaviour may also lead to physical and emotional hardship for those affected.

Some Tweets may seem to be abusive when viewed in isolation, but may not be when viewed in the context of a larger conversation. When we review this type of content, it may not be clear whether it is intended to harass an individual, or if it is part of a consensual conversation. To help our teams understand the context of a

¹³ <https://help.twitter.com/en/privacy>

¹⁴ <https://help.twitter.com/en/rules-and-policies/abusive-behavior>



conversation, we may need to hear directly from the person being targeted, to ensure that we have the information needed prior to taking any enforcement action. However, we review both first-person and bystander reports of such content.

We will review and take action against reports of accounts targeting an individual or group of people with any of the following behaviour within Tweets or Direct Messages.

Violent threats

We prohibit content that makes violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., “I will kill you.” We have a zero tolerance policy against violent threats, and those deemed to be sharing violent threats will face immediate and permanent suspension of their account.

Wishing, hoping, or calling for serious harm on a person or group of people

We do not tolerate content that wishes, hopes, promotes, incites, or expresses a desire for death, serious bodily harm or serious disease against an individual or group of people. This includes, but is not limited to:

- Hoping that someone dies as a result of a serious disease e.g., “I hope you get cancer and die.”
- Wishing for someone to fall victim to a serious accident e.g., “I wish that you would get run over by a car next time you run your mouth.”
- Saying that a group of individuals deserves serious physical injury e.g., “If this group of protesters don’t shut up, they deserve to be shot.”

About wishes of harm exceptions on Twitter

We recognise that conversations regarding certain individuals credibly accused of severe violence may prompt outrage and associated wishes of harm. In these limited cases, we will request the user to delete the Tweet without any risk of account penalty, strike, or suspension. Examples are, but not limited to:

- “I wish all rapists to die.”
- “Child abusers should be hanged.”

Unwanted sexual advances

While some consensual nudity and adult content is permitted on Twitter, we prohibit unwanted sexual advances and content that sexually objectifies an individual without their consent. This includes, but is not limited to:

- sending someone unsolicited and/or unwanted adult media, including images, videos, and GIFs;
- unwanted sexual discussion of someone’s body;
- solicitation of sexual acts; and
- any other content that otherwise sexualises an individual without their consent.

Using insults, profanity, or slurs with the purpose of harassing or intimidating others

We take action against the use of insults, profanity, or slurs to target others. In some cases, such as (but not limited to) severe, repetitive usage of insults or slurs where the primary intent is to harass or intimidate others, we may require Tweet removal. In other cases, such as (but not limited to) moderate, isolated usage of insults and profanity where the primary intent is to harass or intimidate others, we may limit Tweet visibility as further described below. Please also note that while some individuals may find certain terms to be offensive, we will not take action against every instance where insulting terms are used.

Encouraging or calling for others to harass an individual or group of people

We prohibit behaviour that encourages others to harass or target specific individuals or groups with abusive behaviour. This includes, but is not limited to; calls to target people with abuse or harassment online and behaviour that urges offline action such as physical harassment.

Denying mass casualty events took place

We prohibit content that denies that mass murder or other mass casualty events took place, where we can verify that the event occurred, and when the content is shared with abusive intent. This may include references to such



an event as a “hoax” or claims that victims or survivors are fake or “actors.” It includes, but is not limited to, events like the Holocaust, school shootings, terrorist attacks, and natural disasters.

Consequences

When determining the penalty for violating this policy, we consider a number of factors including, but not limited to, the severity of the violation and an individual’s previous record of rule violations. The following is a list of potential enforcement options for content that violates this policy:

- Downranking Tweets in replies, except when the user follows the Tweet author.
- Making Tweets ineligible for amplification in Top search results and/or on timelines for users who don’t follow the Tweet author.
- Excluding Tweets and/or accounts in email or in-product recommendations.
- Requiring Tweet removal.

For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent suspension.

Suspending accounts whose primary use we’ve determined is to engage in abusive behaviour as defined in this policy, or who have shared violent threats.

Guiding principles and standards

Twitter is committed to respecting the human rights of our users, in line with the expectations articulated in the UN Guiding Principles on Business and Human Rights.¹⁵ We have looked to internationally recognised human rights standards to guide our approach to content policy and enforcement, including those related to the protection of freedom of expression, privacy, security, non-discrimination, and ensuring due process. We recognise the value of international standards in guiding how we navigate instances where human rights may be in tension. Each of our Twitter Rules is designed to address specific harms on the platform, and we try to ensure that content moderation actions we take are both necessary and proportionate to addressing such harms.

Statistics related to rule enforcement

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

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

- Impressions for rule-violating Tweets: Our impressions metric captures the number of views a violative Tweet received prior to removal. From 1 January 2021, through 30 June 2021, Twitter required account holders to remove 4.7M Tweets that violated the Twitter Rules.¹⁶
- Of the Tweets removed, 68% received fewer than 100 impressions prior to removal, with an additional 24% receiving between 100 and 1,000 impressions.¹⁷
- In total, impressions on these violative Tweets accounted for less than 0.1% of all impressions for all Tweets during that time period.¹⁸

We also stepped up the level of proactive enforcement across the service and invested in technological solutions to respond to ever-evolving malicious online activity. Today, by using technology, 65% of the abusive content we action is surfaced proactively for human review, instead of relying on reports from people using Twitter.¹⁹

4. **For the record I would like the number of users in Australia; the ages and genders of the users; the full-time-equivalent numbers employed in trust and safety teams compared with the number of employees in other aspects of Twitter operations; and how many of them are based in Australia, with a similar sort of comparison—employees in Australia and the proportion of those who are in**

¹⁵ https://www.ohchr.org/documents/publications/guidingprinciplesbusinessshr_en.pdf

¹⁶ https://blog.twitter.com/en_us/topics/company/2021/transparency-19

¹⁷ *Ibid.*

¹⁸ *Ibid.*

¹⁹ *Ibid.*



the trust and safety teams. I would also like some data on the number of complaints. There are no privacy issues there, and I think that you might have already publicly released some of that information. I'm only concerned about Australia at the moment. I would like the number of complaints, the turnaround times and how many actually need to be escalated. I imagine your operational team would probably sort out a lot of them, but you would keep data on how many need to be escalated. So I will put all of that on notice.

– Ms Celia Hammond MP

We share the Australian Government's desire to promote a healthy digital information ecosystem, and Twitter remains focused on protecting and empowering users to participate in the public conversation every day. Twitter is an open service that's home to a world of diverse people, perspectives, ideas and information. We're committed to protecting the health of the public conversation, and we take that commitment seriously.

Twitter is an inherently open, public, real time service. In addition to our suite of advertising products, Twitter's service is primarily composed of user-generated content. In line with our commitments to preserving freedom of expression and the Open Internet, we believe it is vital to strike the right balance between taking proactive steps to protect from harm with human rights and other vital interests, including freedom of expression, privacy, and ensuring we do not act as the sole arbiter of truth.

The Twitter Rules make clear that users are responsible for the content they post and that advertisers must adhere to specific quality guidelines on our service. There are a wide-ranging set of rules that Twitter enforces when content is posted that infringes those rules.²⁰ In addition to content that violates the Twitter Rules or Terms of Service, there are Tweets that we do remove, such as illegal Tweets. Many countries, including Australia, have laws that may apply to Tweets and/or Twitter account content. In our continuing effort to make our services available to people everywhere, we duly review valid and properly scoped requests from authorised entities or legal court orders and take action where necessary.

Globally, Twitter tracks and reports on the reports of violative content and actions taken in the Twitter Transparency Report found in at the Twitter Transparency Centre.²¹ It should be noted, that while Twitter does not report numbers for enforcement of Twitter's Spam or Platform Manipulation policies as enforcement of these policies are a key cornerstone of Twitter's efforts to combat the spread of misinformation and disinformation. However on average, Twitter challenges approximately 5 million accounts per week to prevent these violations.²²

Below we have included Australia-specific approximate data²³ for July to December 2020, as reported in Twitter's Initial Report on the Australian Code of Practice on Disinformation and Misinformation in May 2021.²⁴

- **All Twitter Rules** (see the Rules Enforcement page for a list of rules)²⁵
 - a. 37,000 Australian accounts were actioned for violations of the Twitter Rules.
 - b. 7,200 Australian accounts were suspended for violations of the Twitter Rules.
 - c. 47,000 pieces of content authored by Australian accounts were removed for violations of the Twitter Rules.
- **COVID-19 misleading information**
 - More than 50 Australian accounts were actioned for violations of the COVID-19 misleading information policy.²⁶
 - Less than 10 Australian accounts were suspended for violations of the COVID-19 misleading information policy.²⁷
 - More than 50 pieces of content authored by Australian accounts were removed for violations of the COVID-19 misleading information policy.²⁸
- **Civic integrity**

²⁰ <https://help.twitter.com/en/rules-and-policies#general-policies>

²¹ <https://transparency.twitter.com/>

²² <https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jan-jun>

²³ Australian accounts are identified using the account's country designation. The country designation is assigned automatically at sign up, but also can be manually modified by the user.

²⁴ <https://digi.org.au/wp-content/uploads/2021/05/20210504-APPENDIX-2-AUSTRALIAN-CODE-OF-PRACTICE-ON-DISINFORMATION-AND-MISINFORMATION-Twitter-Initial-Report.pdf>

²⁵ <https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jan-jun>

²⁶ <https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy>

²⁷ *Ibid.*

²⁸ *Ibid.*



- d. More than 40 Australian accounts were actioned for violations of the civic integrity policy.²⁹
- e. Zero) Australian accounts were suspended for violations of the civic integrity policy.³⁰
- f. About 70 pieces of content authored by Australian accounts were removed for violations of the civic integrity policy.³¹

Labelling for misinformation was a new enforcement remedy added in 2020 and 2021. Between 28 January 2021 – 28 April 2021, approximately 25,000 accounts had at least one Tweet labelled for misinformation. This metric reflects accounts across all countries.

Employees and users

In our most recent quarterly report for Q3 2021, we have shared that Twitter now employs approximately 7,100 employees worldwide, up 33% year over year.³² With regards to moderation and enforcement, we have a specialised, global team that enforces the Twitter Rules with impartiality and provides 24/7 global coverage in all time zones, including Australia, and in multiple different languages. We are continuing to build more capacity to address increasingly complex issues.

These teams are not composed of public-facing representatives for the company. They are dedicated internally-facing personnel who go through rigorous training on the Twitter Rules, our enforcement philosophy, and all of our tooling. They review content based on our policies only and never on the basis of political orientation. This is by design and is a core principle of how we do our work. Review teams will always be a critical part of our approach but the future in this area is technological and that's where we are placing most of our attention going forward. Content moderation policies and how we enforce Twitter Rules are uniform across the globe, and we have always used a combination of machine learning and human review. By using technology, 65% of the abusive content we now action is surfaced proactively for human review instead of relying on reports from people using Twitter.³³

In our last quarterly reporting period, our global average monetizable daily active users (mDAU) reached 211 million, up 13% year over year, driven by ongoing product improvements and global conversation around current events.³⁴ Twitter defines monetizable daily active usage or users (mDAU) as people, organisations, or other accounts who logged in or were otherwise authenticated and accessed Twitter on any given day through twitter.com or the Twitter app that are able to show ads. Average mDAU for a period represents the number of mDAU on each day of such period divided by the number of days for such period. Changes in mDAU are a measure of changes in the size of our daily logged in or otherwise authenticated active total accounts. We do not publicly provide mDAU breakdowns by country or region. Our geographic location data collected for purposes of reporting is based on the IP address or phone number associated with the account when an account is initially registered on Twitter. The IP address or phone number may not always accurately reflect a person's actual location at the time they engaged with our platform. For example, someone accessing Twitter from the location of the proxy server that the person connects to rather than from the person's actual location.

5. **Dr Brian Tyson, a doctor in the USA, has recently written a book, *Overcoming the COVID-19 Darkness: How Two Doctors Successfully Treated 7000 Patients*. He documents how, through his clinic, he treated 7,000 patients using hydroxychloroquine and ivermectin with zero deaths and only a handful of hospitalisations. How can a doctor like that, a qualified medical practitioner with such experience, be deplatformed and banned from Twitter?**

– Mr Craig Kelly MP

Twitter suspended the account referenced for violations of the Twitter Rules on COVID-19 misleading information.

We share the belief it is vital for Australians to identify trusted health information online. At Twitter, our primary goal in addressing COVID-19 misleading information has been focused on removing demonstrably false or potentially misleading content that has the highest risk of causing harm, as well as surfacing credible content from

²⁹ <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>

³⁰ *Ibid.*

³¹ *Ibid.*

³² https://s22.q4cdn.com/826641620/files/doc_financials/2021/q3/Final-Q3'21-Shareholder-letter.pdf

³³ *Ibid.*

³⁴ <https://investor.twitterinc.com/home/default.aspx>



authoritative sources through Twitter's Explore tab,³⁵ dedicated search prompts for COVID-19,³⁶ and vaccination information from trusted partners, like the Australian Department of Health.³⁷

As scientific understanding of the COVID-19 pandemic continues to develop, we've observed the emergence of persistent conspiracy theories, alarmist rhetoric unfounded in research or credible reporting, and a wide range of unsubstantiated rumours, when left uncontextualised, can prevent the public from making informed decisions regarding their health, and puts individuals, families, and communities at risk.

To make sure we are addressing such COVID-related misinformation head on, we created dedicated policies to address it starting in March 2020. We've updated those as conditions and issues have evolved. Since introducing our COVID-19 guidance in 2020, Twitter has challenged 11.7 million accounts, suspended 4,110 accounts, and removed over 72,062 pieces of content worldwide.³⁸

Last year, Twitter updated our policies to cover new enforcement actions in relation to misleading information around the COVID-19 vaccines. Our policies state that people may not use Twitter's services to share false or misleading information about COVID-19 which may lead to harm, such as increased exposure to the virus, or adverse effects on public health systems. This includes sharing content that may mislead people about the nature of the COVID-19 virus; the efficacy and/or safety of preventative measures, treatments, or other precautions to mitigate or treat the disease; official regulations, restrictions, or exemptions pertaining to health advisories; or the prevalence of the virus or risk of infection or death associated with COVID-19.

We also seek to protect robust, public debate about the response to COVID-19 recognising that the state of scientific knowledge about certain aspects of the pandemic and public response to it (including the development of vaccines) is still relatively nascent. In the absence of other policy violations, the following are generally not in violation of this policy:

- Strong commentary, opinions, and/or satire, provided these do not contain false or misleading assertions of fact.
- Counterspeech. We allow for direct responses to misleading information which seek to undermine its impact by correcting the record, amplifying credible information, and educating the wider community about the prevalence and dynamics of misleading information.
- Personal anecdotes or first-person accounts.
- Public debate about the advancement of COVID-19 science and research, including debate about research related to COVID-19, such as the effectiveness of treatments and mitigation measures, so long as the claims don't misrepresent research findings.

A critical aspect of getting this balance right is our quickly evolving efforts to address impact. The consequences for violating our COVID-19 misleading information policy depends on the severity and type of the violation and the account's history of previous violations. In instances where accounts repeatedly violate this policy, we will use a strike system to determine if further enforcement actions should be applied. We believe this system further helps to reduce the spread of potentially harmful and misleading information on Twitter, particularly for high-severity violations of our rules. The actions we take may include the following:

- **Tweet deletion:** for high-severity violations of this policy, including (1) misleading information related to the nature or treatment of the COVID-19 virus and (2) pandemic or COVID-19 vaccines that invoke a deliberate conspiracy by malicious and/or powerful forces, we will require you to remove this content. We will also temporarily lock you out of your account before you can Tweet again. Tweet deletions accrue 2 strikes.
- **Labelling:** in circumstances where we do not remove content which violates this policy, we may provide additional context on Tweets sharing the content where they appear on Twitter. This means we may:
 - Apply a label and/or warning message to the Tweet;
 - Show a warning to people before they share or like the Tweet;
 - Reduce the visibility of the Tweet on Twitter and/or prevent it from being recommended;
 - Turn off likes, replies, and Retweets; and/or

³⁵ <https://help.twitter.com/en/twitter-guide/topics/how-to-get-started-with-twitter/how-to-use-the-explore-tab-twitter-help>

³⁶ https://blog.twitter.com/en_us/topics/company/2020/stepping-up-our-work-to-protect-the-public-conversation-around-covid-19.html

³⁷ https://blog.twitter.com/en_au/topics/company/2020/helping-you-find-reliable-public-health-information-on-twitter.html

³⁸ <https://transparency.twitter.com/en/reports/covid19.html#2021-jan-jun>



- Provide a link to additional explanations or clarifications, such as in a curated landing page or relevant Twitter policies.

In most cases, we will take all of the above actions on Tweets we label. We prioritise producing Twitter Moments in cases where misleading content on Twitter is gaining significant attention and has caused public confusion on our service. Tweets that are labelled and determined to be harmful will accrue 1 strike.

- **Account locks and permanent suspension:** if we determine that an account is dedicated to Tweeting or promoting a particular misleading narrative (or set of narratives) about COVID-19, this would also be grounds for suspension.

We are also currently running an experiment in Australia (as well as in South Korea and the USA, which recently expanded to Brazil, Spain, and the Philippines³⁹) to allow users to report potentially misleading content.⁴⁰ From mid-2021 to January 2022, we received over 3 million reports via this experiment. Since August last year, Australian Twitter users have been able to flag tweets that seem misleading via a new option on the platform's content reporting mechanism. As part of our commitment to exploring and testing new ways to address potentially misleading information on Twitter, people are also able to report potential misleading information that our policies do not currently cover.

We will continue to maintain and build strong collaborative relationships with relevant multinational and local stakeholders to help inform our approach and updates to our relevant policies. We will continue to work with our trusted partners to fight COVID-19 misinformation, promote credible information, and share important public health messages on Twitter.

6. Additional information regarding Twitter's policy development process and public consultations.

– Deputy Chair Tim Watts MP

It's our responsibility to create rules on Twitter that are fair and set clear expectations for everyone on our service. That's why we began processes to seek public input from around the globe on how we will address key policies.

In 2020, we announced our public consultation on synthetic and manipulated media.⁴¹ We undertook a survey on our initial draft of this rule, as well as Tweets that included the hashtag #TwitterPolicyFeedback, 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. We then shared what we learned and how it shaped the update to The Twitter Rules, how we'll treat this content when we identify it, as well as how we would label or remove the content on Twitter as part of this change.⁴²

Given the strong response we received through this initial interaction of public consultation, we built on this process for how we define the public interest on Twitter, as well as our principles and approach to world leaders on our service,⁴³ and engaged in targeted consultations for updates to our hateful conduct policy.⁴⁴

We work hard to ensure that a global perspective is reflected and that we maintain a consistent approach across the multiple public surveys. These consultations have helped us understand the evolving community expectations, as well as what type of enforcement action people generally believe is appropriate in given situations. Additionally with each change or new policy we create, we also consult with a range of human rights experts, civil society organisations, and academics worldwide whose feedback is reflected in forthcoming revisions to Twitter's policy framework.

We want to serve the public conversation, and our aim is to have rules and policies that appropriately balances fundamental human rights and considers the global context in which we operate.

Twitter Trust & Safety Council

We also work with some of the leading online safety experts who reside on our global Twitter Trust & Safety Council for guidance and feedback on our policies and products.⁴⁵

³⁹ <https://twitter.com/TwitterSafety/status/1483076718730649607?s=20&t=xsLeJg4RAFYrKGA-xR661w>

⁴⁰ <https://www.theguardian.com/australia-news/2021/aug/18/twitter-to-allow-australian-users-to-flag-potential-misinformation-during-month-long-trial>

⁴¹ <https://twitter.com/TwitterSafety/status/1186403736995807232>

⁴² https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media

⁴³ https://blog.twitter.com/en_us/topics/company/2021/calling-for-public-input-on-our-approach-to-world-leaders

⁴⁴ https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate

⁴⁵ <https://about.twitter.com/en/our-priorities/healthy-conversations/trust-and-safety-council>



The Trust & Safety Council is a group of independent expert organisations from around the world. Together, they advocate for safety and advise us as we develop our products, programs, and rules. At the end of 2019, we expanded the Council to include even more global experts and diverse perspectives.

The Council is made up of several advisory groups, each dedicated to issues critical to the health of the public conversation. Areas of focus include Online Safety and Harassment, Human and Digital Rights, Suicide Prevention and Mental Health, Child Sexual Exploitation, and Dehumanisation. We are proud to work with a number of Australian nonprofits who are longstanding members of the Trust & Safety Council, including the Alannah & Madeline Foundation, Beyond Blue, Black Rainbow, Lifeline Australia, Kids Helpline, Project Rockit, Reachout Australia, and the Young & Resilient Research Centre at the University of Western Sydney.

Public Application Programming Interfaces (APIs)

In line with our principles of transparency and to improve understanding of the public conversation, Twitter makes available real-time access to the global conversation through our free, open application programming interfaces (APIs).⁴⁶

At a high level, APIs are the way computer programs “talk” to each other so that they can request and deliver information. This is done by allowing a software application to call what's known as an endpoint: an address that corresponds with a specific type of information we provide (endpoints are generally unique like phone numbers).⁴⁷ Twitter allows access to parts of our service via APIs to allow people to build software that integrates with Twitter, like a solution that helps a company respond to customer feedback on Twitter.

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

Academic partnerships

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

With issues that are new, complex and rapidly evolving, a lot of work is also being done to better define the scope and extent of online behaviours to be studied, so as to make research methodologies more robust. Due to our public nature, Twitter is frequently scrutinised; however, we encourage robust study and academic partnerships through recent expansion of APIs coupled with our own public disclosures, like our state-backed information operations database.⁴⁸

- 7. When asked at recent public hearings whether certain statements would breach a platform's terms of service and allow for removal, a number of social media companies outlined that the answer depended on the context of the comments. Below are some examples of abusive and derogatory comments posted online. Can you outline to the committee the context in which these statements would not breach your policies and remain on your site?**

Would the context differ for a private individual and a public figure?

- **Language suggesting a woman should put a bag over her head to suffocate herself**
- **“Take half a brick and hide behind a bush”**
- **“I bet she rages so hard a natural disaster occurs every time she has her period”**
- **“That woman is the reason nature designed the human hand to grasp a penis in a pleasing manner”**
- **“Cavorting whore”**

⁴⁶ <https://developer.twitter.com/en>

⁴⁷ <https://help.twitter.com/en/rules-and-policies/twitter-api>

⁴⁸ <https://transparency.twitter.com/en/reports/information-operations.html>



- **“Every homophobic whore deserves to have their c*nt sliced open with a chainsaw. Including the filthy sluts in the Australian Parliament... I’m not advocating violence, but if any of those bigots was set on fire, I’d toast marshmallows in the flames.”**

Twitter is reflective of real conversations happening in the world and that sometimes includes perspectives that may be offensive, controversial, and/or bigoted to others. While we welcome everyone to express themselves on our service, we will not tolerate behaviour that harasses, threatens, or uses fear to silence the voices of others.

We have the Twitter Rules in place to help ensure everyone feels safe expressing their beliefs and we strive to enforce them with uniform consistency.⁴⁹

Our policy development process

Creating a new policy or making a policy change requires in-depth research around trends in online behaviour, developing clear external language that sets expectations around what’s allowed, and creating enforcement guidance for reviewers that can be scaled across millions of Tweets.

While drafting policy language, we gather feedback from a variety of internal teams as well as our Trust & Safety Council. This is vital to ensure we are considering global perspectives around the changing nature of online speech, including how our rules are applied and interpreted in different cultural and social contexts. Finally, we train our global review teams, update the Twitter Rules, and start enforcing the new policy.

Our enforcement philosophy

We empower people to understand different sides of an issue and encourage dissenting opinions and viewpoints to be discussed openly. This approach allows many forms of speech to exist on our platform and, in particular, promotes counterspeech: speech that presents facts to correct misstatements or misperceptions, points out hypocrisy or contradictions, warns of offline or online consequences, denounces hateful or dangerous speech, or helps change minds and disarm.

Thus, context matters. When determining whether to take enforcement action, we may consider a number of factors, including (but not limited to) whether:

- the behaviour is directed at an individual, group, or protected category of people;
- the report has been filed by the target of the abuse or a bystander;
- the user has a history of violating our policies;
- the severity of the violation;
- the content may be a topic of legitimate public interest.

To elaborate further, we analyse these factors in detail.

- (1) Is the behaviour directed at an individual or group of people?

To strike a balance between allowing different opinions to be expressed on the platform, and protecting our users, we enforce policies when someone reports abusive behaviour that targets a specific person or group of people. This targeting can happen in a number of ways (for example, @mentions, tagging a photo, mentioning them by name, and more).

- (2) Has the report been filed by the target of the potential abuse or a bystander?

Some Tweets may seem to be abusive when viewed in isolation, but may not be when viewed in the context of a larger conversation or historical relationship between people on the platform. For example, friendly banter between friends could appear offensive to bystanders, and certain remarks that are acceptable in one culture or country may not be acceptable in another. To help prevent our teams from making a mistake and removing consensual interactions, in certain scenarios we require a report from the actual target (or their authorised representative) prior to taking any enforcement action.

- (3) Does the user have a history of violating our policies?

We start from a position of assuming that people do not intend to violate our Rules. Unless a violation is so egregious that we must immediately suspend an account, we first try to educate people about our Rules and give

⁴⁹ <https://help.twitter.com/en/rules-and-policies/twitter-rules.html>



them a chance to correct their behaviour. We show the violator the offending Tweet(s), explain which Rule was broken, and require them to remove the content before they can Tweet again. If someone repeatedly violates our Rules then our enforcement actions become stronger. This includes requiring violators to remove the Tweet(s) and taking additional actions like verifying account ownership and/or temporarily limiting their ability to Tweet for a set period of time. If someone continues to violate Rules beyond that point then their account may be permanently suspended.

(4) What is the severity of the violation?

Certain types of behaviour may pose serious safety and security risks and/or result in physical, emotional, and financial hardship for the people involved. These egregious violations of the Twitter Rules — such as posting violent threats, non-consensual intimate media, or content that sexually exploits children — result in the immediate and permanent suspension of an account. Other violations could lead to a range of different steps, like requiring someone to remove the offending Tweet(s) and/or temporarily limiting their ability to post new Tweet(s).

(5) Is the behaviour newsworthy and in the legitimate public interest?

Twitter moves at the speed of public consciousness and people come to the service to stay informed about what matters. Exposure to different viewpoints can help people learn from one another, become more tolerant, and make decisions about the type of society we want to live in.

To help ensure people have an opportunity to see every side of an issue, there may be the rare occasion when we allow controversial content or behaviour which may otherwise violate our Rules to remain on our service because we believe there is a legitimate public interest in its availability. Each situation is evaluated on a case by case basis and ultimately decided upon by a cross-functional team.

Some of the factors that help inform our decision-making about content are the impact it may have on the public, the source of the content, and the availability of alternative coverage of an event.

- *Public impact of the content:* A topic of legitimate public interest is different from a topic in which the public may be curious. We will consider what the impact is to citizens if they do not know about this content. If the Tweet does have the potential to impact the lives of large numbers of people, the running of a country, and/or it speaks to an important societal issue then we may allow the content to remain on the service. Likewise, if the impact on the public is minimal we will most likely remove content in violation of our policies.
- *Source of the content:* Some people, groups, organisations and the content they post on Twitter may be considered a topic of legitimate public interest by virtue of their being in the public consciousness. This does not mean that their Tweets will always remain on the service. Rather, we will consider if there is a legitimate public interest for a particular Tweet to remain up so it can be openly discussed.
- *Availability of coverage:* Everyday people play a crucial role in providing firsthand accounts of what's happening in the world, counterpoints to establishment views, and, in some cases, exposing the abuse of power by someone in a position of authority. As a situation unfolds, removing access to certain information could inadvertently hide context and/or prevent people from seeing every side of the issue. Thus, before actioning a potentially violating Tweet, we will take into account the role it plays in showing the larger story and whether that content can be found elsewhere.

Public figures

Twitter generally actions Tweets that violate our rules. However, we recognise that sometimes it may be in the public interest to allow people to view Tweets that would otherwise be taken down. We consider content to be in the public interest if it directly contributes to understanding or discussion of a matter of public concern.⁵⁰

At present, we limit exceptions to one critical type of public-interest content—Tweets from elected and government officials—given the significant public interest in knowing and being able to discuss their actions and statements. These are known as our public-interest exceptions.⁵¹

As a result, in rare instances, we may choose to leave up a Tweet from an elected or government official that would otherwise be taken down. Instead we will place it behind a notice providing context about the rule violation that allows people to click through to see the Tweet. Placing a Tweet behind this notice also limits the ability to

⁵⁰ <https://help.twitter.com/en/rules-and-policies/public-interest>

⁵¹ *Ibid.*



engage with the Tweet through likes, Retweets, or sharing on Twitter, and makes sure the Tweet isn't algorithmically recommended by Twitter. These actions are meant to limit the Tweet's reach while maintaining the public's ability to view and discuss it.

We are currently in the process of updating this policy.⁵² As politicians and government officials evolve how they use our service, we want our policies to remain relevant to the ever-changing nature of political discourse on Twitter and protect the health of the public conversation. Last year, we opened up a public survey to receive feedback that helped inform the development of our policy framework. Nearly 49,000 people from around the globe took time to share their feedback on how content from world leaders should be handled on our service. During this process, we've also engaged experts, including NGOs, governments, academics, and civil society, to ensure we're hearing as many diverse and thoughtful perspectives as possible. As our teams review and distil the data, we've been looking for key themes, new ideas, and creative thinking so we can begin to develop an update to our approach and consider next steps. We are crafting a framework that captures these issues, which we are working to have in place soon.

Content moderation principles

We support the Santa Clara Principles on Transparency and Accountability in Content Moderation to consider how best to obtain meaningful transparency and accountability around demands for increasingly aggressive moderation of user-generated content on Twitter.⁵³ We also share detailed information about how we enforce the Twitter Rules in our bi-annual Transparency Reports.⁵⁴

At Twitter, we have identified our responsibilities and limits. There are Tweets that we do remove, such as illegal Tweets. Many countries, including Australia, have laws that may apply to Tweets and/or Twitter account content. In our continuing effort to make our services available to people everywhere, we duly review valid and properly scoped requests from authorised entities or legal court orders and take action where necessary.

Transparency is vital to protecting freedom of expression; thus, in the case of legal demands concerning content removal, we also have a counter-notice process where we will promptly notify affected users unless we are prohibited from doing so (e.g. if we receive a court order under seal). When content has been withheld, we also clearly indicate within the product and publish requests to withhold content on Lumen—unless, similar to our practice of notifying users, we are prohibited from doing so.

Our position on freedom of expression carries with it a mandate to protect our users' right to speak freely and preserve their ability to contest having their private information revealed. While we may need to release information as required by law, we try to notify Twitter users before handing over their information whenever we can so they have a fair chance to counter the request if they so choose.

⁵² https://blog.twitter.com/en_us/topics/company/2021/calling-for-public-input-on-our-approach-to-world-leaders

⁵³ <https://santaclaraprinciples.org/>

⁵⁴ <https://transparency.twitter.com/>