Australian Code of Practice on Disinformation and Misinformation

Twitter Annual Transparency Report
Reporting period: 2021

Summary

Twitter's mission is to serve the public conversation. Transparency is fundamental to our work in achieving that mission. This annual report outlines our commitments and progress under the Australian Code of Practice on Disinformation and Misinformation (the Code) to demonstrate Twitter's efforts to protect the public conversation and uphold the integrity of our service.

We are committed to providing meaningful transparency reporting to the public. We do this both in partnership with government, academics, and civil society under schemes like the Code and through existing, proactive self reporting initiatives like the biannual Twitter Transparency Report.

Under the Code, Twitter has made meaningful commitments and progress on all mandatory objectives and applicable opt-ins. Encouragingly, many measures outlined under the Code were already underway through Twitter's proactive policy enforcement and reporting measures, including the Twitter Transparency Reports, Twitter Transparency Centre, and Information Operations disclosures.

Key insights from the report

- Overview of ongoing work to protect against platform manipulation and enhance authenticity
- Efforts to disrupt global state-backed information operations to preserve integrity in public conversation
- Trends concerning the enforcement of the Twitter Rules
- Explanation of the range of measures and enforcement options available

As online behaviours evolve, we continue to experiment, iterate, and strengthen our approach to protecting the public conversation on Twitter and our place in the online information ecosystem. We are moving with urgency, purpose, and commitment as we develop and enforce a range of policy, procedural, and product changes to the Twitter platform.

Commitments under the Code

Twitter's mission is to serve the public conversation. As a global platform for public conversation, it provides a network that connects users to people, information, ideas, opinions, and news in real-time. The company's services include live commentary, connections and conversations. Through the mobile Twitter app or desktop version, the Twitter platform provides social networking and microblogging services through 280-character Tweets that can also feature images, video, audio, and GIFs. The company can also be used as a marketing tool for businesses through its promoted products including Promoted Tweets, Promoted Accounts, and Promoted Trends.

Twitter is a singular platform and service. Therefore, as a signatory to the Australian Code of Practice on Disinformation and Misinformation and consistent with our initial report, Twitter indicates in the following table where the company has opted-in to the following Objectives and Outcomes under the Code, and provides details on these measures in our following report.

Objective 1: Provide Safeguards against Harms that may arise from Disinformation and Misinformation.	Outcome 1a: Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures.	**Opt-in**
	Outcome 1b: Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.	**Opt-in**
	Outcome 1c: Users can report content and behaviours to Signatories that violates their policies under 5.10 through publicly available and accessible reporting tools	**Opt-in**
	Outcome 1d: Users will be able to access general information about Signatories actions in response to reports made under 5.11.	**Opt-in**
Objective 2: Disrupt advertising and monetisation incentives for Disinformation	Outcome 2: Advertising and/or monetisation incentives for Disinformation are reduced.	**Opt-in**
Objective 3: Work to ensure the security and integrity of services and products delivered by digital platforms.	Outcome 3: The risk that Inauthentic User Behaviours undermine the integrity and security of Services and Products is reduced.	**Opt-in**
Objective 4: Empower consumers to make better informed choices of digital content.	Outcome 4: Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.	**Opt-in**
Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.	Outcome 5: Users are better informed about the source of Political Advertising.	**Not applicable** *Twitter globally prohibits the promotion of political content.*
Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.	Outcome 6: Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.	**Opt-in** *Requirements that researchers must adhere to, in order to be eligible, are outlined in the initial report.*
Objective 7: Signatories publicise the measures they take to combat Disinformation and Misinformation.	Outcome 7: The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.	**Opt-in**

Objective 1: Safeguards against Disinformation and Misinformation

Outcome 1a: Reducing harm by adopting scalable measures

Twitter addresses misinformation and disinformation through a range of policies, enforcement actions, and product solutions. Under our policies, Twitter defines misleading content ('misinformation') as claims that have been *confirmed to be false by external, subject-matter experts or include information that is shared in a deceptive or confusing manner*.

This content is identified through a combination of human review and technology, and through partnerships with global third-party experts.

We manage the risk of public harm in many ways. The combination of actions we take are meant to be proportionate to the level of potential harm from that situation. People who repeatedly violate our policies may be subject to temporary or permanent suspensions.

Depending on potential for offline harm, we limit amplification of misleading content or remove it from Twitter if offline consequences could be immediate and severe.

In other situations, we aim to inform and contextualise by sharing timely information or credible content from third-party sources. This is done by:

- **Labelling content**: For claims that do not meet our threshold for removal, outlined in the policies above, we may label the Tweet to give readers a notice and/or share additional context with them. Labelled Tweets are subject to reduced visibility. Labels are visible in all Twitter-supported languages.
- **Prompting a user when they engage with a misleading Tweet**: When you try to share a Tweet that was labelled for violating one of our policies, you will see a prompt to help you find additional context and consider whether or not to amplify the Tweet to your followers.
- **Creating Twitter Moments**: Learn from other people on Twitter and trusted sources about what's happening in the world and what that might mean for Twitter's users. Twitter Moments are available in multiple global regions. More information about Moments is available here.
- **Launching prebunks**: During important events (e.g. COVID-19 pandemic, elections), we may proactively feature informative messages or updates to counter misleading narratives that emerge. In the past, we've launched prebunks about the COVID-19 vaccine, mail-in voting ballots, election results, and more. Users will see prebunks directly in their Twitter Timeline.
- **Search prompts:** When someone searches for certain hashtags or keywords on Twitter – such as those associated with a civic event or a natural disaster – a notification will be shown at the top of the search results directing people to the latest, authoritative information from credible sources.

Additionally, we're testing opportunities for people to share feedback directly with Twitter and the community. While the actions Twitter takes against a misleading Tweet are driven by the Twitter Rules, the public conversation is better served with diverse participation.

CASE STUDY 1: Birdwatch
Twitter is currently testing a feature called Birdwatch in the US. Birdwatch aims to create a better informed world by empowering people on Twitter to collaboratively add notes to potentially misleading Tweets. Pillot contributors can write notes on any Tweet and if enough other contributors rate that note as helpful, highly ranked Birdwatch notes may be publicly shown on a Tweet.

During this phase, Birdwatch notes will be publicly visible on Tweets to only a small group of US customers. However, all people in the US can see these notes on a separate Birdwatch site. Based on initial random sample surveys of people in the US found:

- The majority of people found Birdwatch notes helpful.
- People were 20-40% less likely to agree with a potentially misleading Tweet's content after reading a Birdwatch note about it, compared to those who did not

Birdwatch notes do not represent Twitter's viewpoint and cannot be edited or modified by our teams. A Tweet with a Birdwatch note will not be labelled, removed, or addressed by Twitter unless it is found to be violating Twitter Rules, Terms of Service, or our Privacy Policy. Failure to abide by the rules can result in one's removal from the Birdwatch pilot, and/or other remediations. Anyone can report notes they believe aren't in accordance with those rules by clicking or tapping the menu on a note, and then selecting "Report."

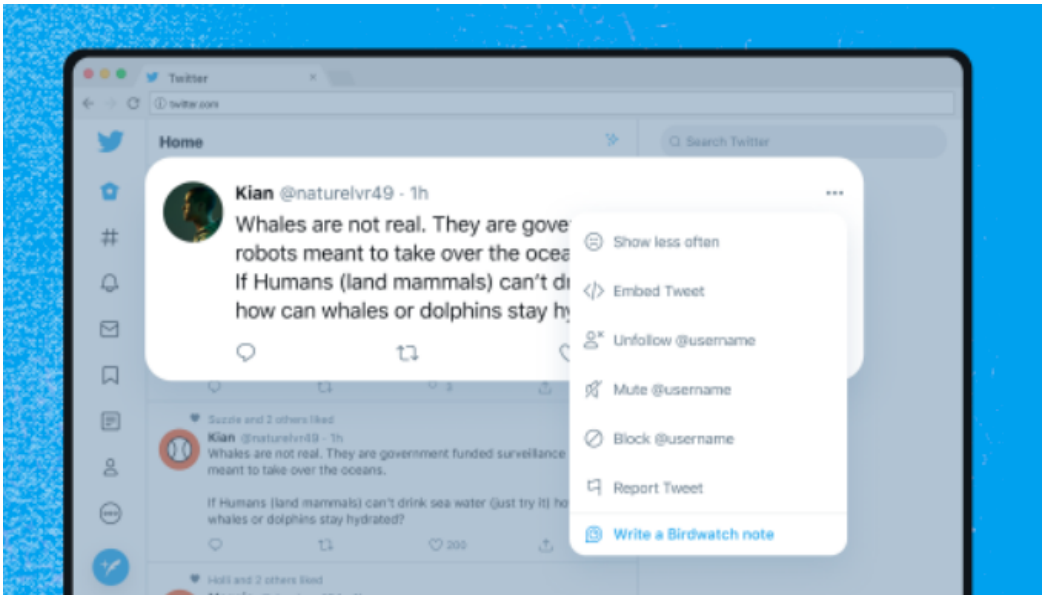

Image 1: Depicts the drop-down menu available for Contributors to write a Birdwatch Note.

More information about Birdwatch and the ongoing pilot is available on Github here.

CASE STUDY 2: Twitter Conversation Settings
Over the past two years, Twitter has built and expanded features called Conversation Settings to give people the ability to control who can reply to Tweets they post. Everyone that uses Twitter is able to use these settings so unwanted replies don't get in the way of meaningful conversations.

Before a user Tweets, they can choose who can reply to their Tweet with three options: (1) everyone (standard Twitter, and the default setting), (2) only people they follow, or (3) only people they @mention. Tweets with the latter two settings will be labelled and the Reply icon will be grayed out for people who can't reply. People who can't reply will still be able to view, Retweet, Retweet with Comment, and Like these Tweets.

The feedback on Conversation Settings has helped people feel safer, allowed people to still see various points of view, and fostered more meaningful conversations on Twitter.

Once we deployed this feature, we saw:

- People who have submitted abuse reports are 3x more likely to use these settings.
- 60% of people found it helped block out noise and instead didn't use Mute or Block.
- People more frequently look for additional commentary when replies are limited, and the new Retweets with Comments timeline is visited 4x more often on Tweets using these Conversation Control settings.

More recently, Twitter is testing the feature to allow people to remove @mentions of themselves in Tweets. This beta test launched as a global experiment on 7 April 2022, and gives people the ability to unmention themselves from a single Tweet, thread, or conversation.

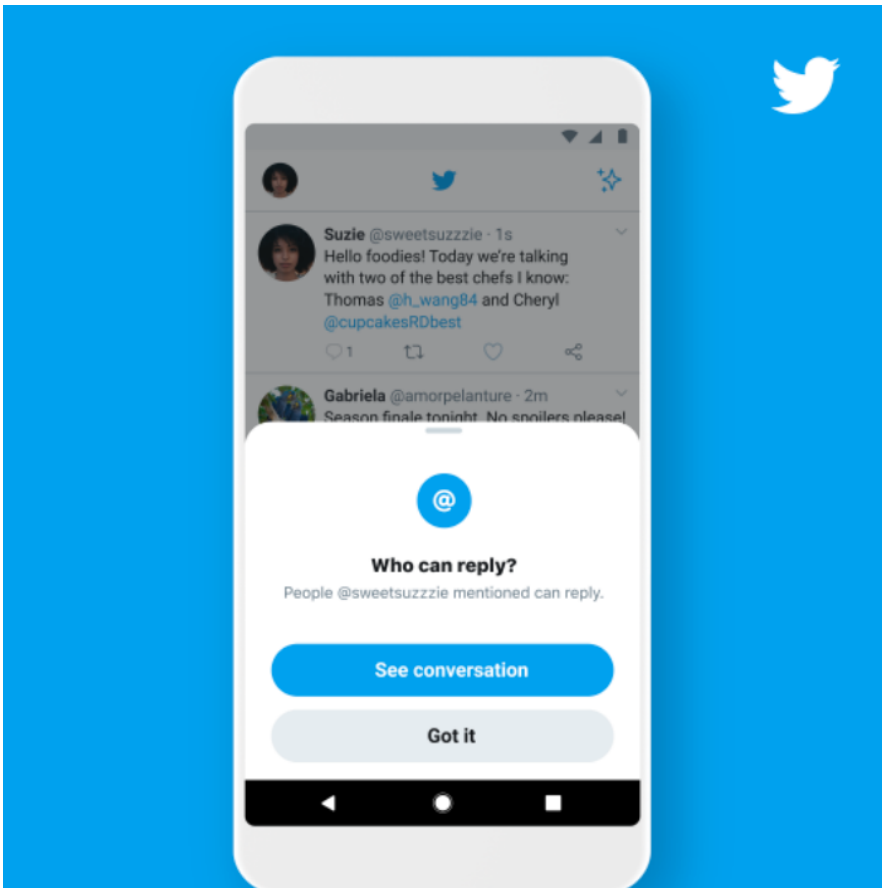

Image 2: Depicts how a person can choose their Conversation settings.

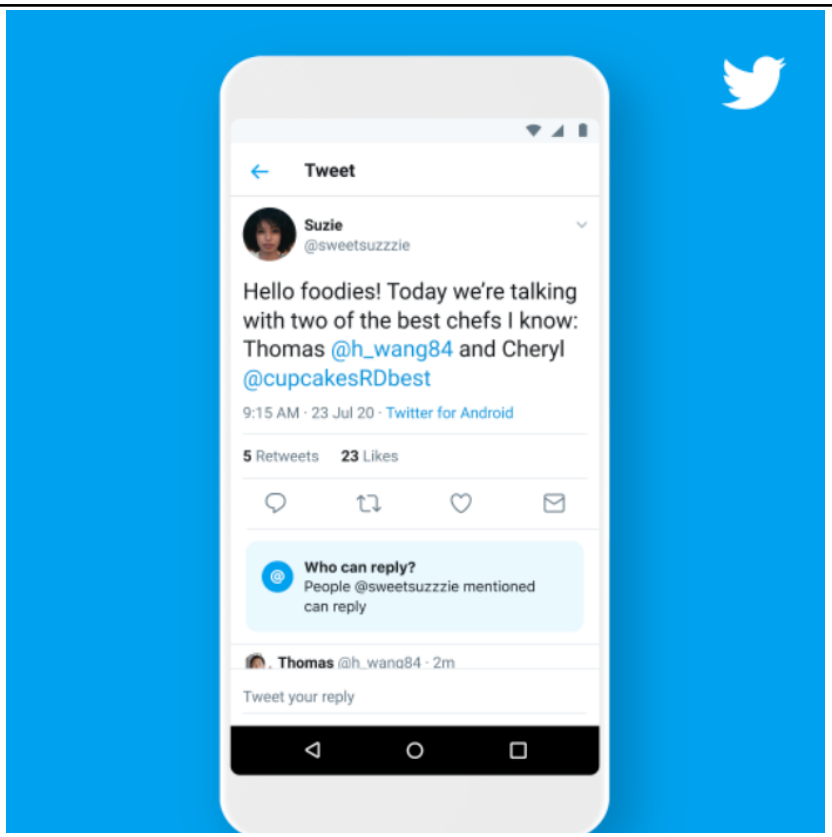

Image 3: Shows the label applied when a person uses Conversation settings.

Twitter's approach to rules and policies

Twitter is committed to sharing our policies in accessible, plain language. We have updated our rules and recognise the importance of having policies that can be easily understood by the general public. This is vital for complex topic areas like defining and identifying platform manipulation, misinformation, and disinformation, which are being continually evaluated and further refined by subject matter experts. Our policies, and corresponding transparency reports, outline a variety of different behaviours that might fall under these categories.

In addition to our policies, which enable us to remove or provide further context on potentially misleading information, we have adopted a number of strategies to signal authenticity, accuracy and authoritative information on the service. These include Explore tab, Moments, Search Prompts, and Verification, as well as labelling government and state-sponsored media accounts so that users can evaluate the source of information. We provide this context both through Twitter's Curation team and through trusted partnerships with government, experts, and civil society groups. Our Curators don't act as reporters or creators of original work. They organise and present content that already exists on Twitter in Moments, explanatory content on Trends, in Lists and more. They are guided by publicly available principles, including impartiality and accuracy, which is available on our Help Centre.

For example, in addressing COVID-19 misleading information, our work has focused on removing demonstrably false or potentially misleading content that has the highest risk of causing harm, as well as surfacing credible content from authoritative sources. The latter has been driven through profile verification, our curated COVID-19 tab in Explore, dedicated COVID-19 pages in the Explore tab, and Search Prompts for COVID-19 and vaccinations, in partnership with agencies such as the Australian Federal Department of Health or World Health Organization (WHO).

The launch of our evergreen COVID-19 and immunisation search prompts is built on our existing work to guard against the artificial amplification of non-credible content about the safety and effectiveness of vaccines. When people in Australia search certain keywords related to COVID-19 or vaccination on Twitter, a prompt appears at the top of the search bar, alerting them to credible sources where they

can find the most up to date information from the Federal Department of Health. In April 2021, a timeline prompt also connected people in Twitter in Australia with the COVID-19 vaccine landing page from the Federal Department of Health and a curated Moment, made by our Curation team, that covered a wide range of topics such as: vaccine safety, effectiveness, vaccines available, distribution plans, how to stay safe before/after vaccinations, and more.

Finally, we believe Twitter has a responsibility to protect the integrity of the public conversation — including working with research, academic, and civil society partners on the timely disclosure of information about attempts to manipulate Twitter to influence elections and other civic conversations by foreign or domestic state linked entities. More about these efforts are covered below.

Outcome 1b: Inform users about what content is targeted

New initiatives in communicating to users what constitutes mis/disinformation

Twitter misleading information prompts

- Our work to limit the spread of misleading information goes beyond elections. Twitter began testing a feature that before a user Retweets or Quote Tweets any labelled Tweet that breaks Twitter's misleading information rules, they will see a warning prompt, which alerts users when they go to share a Tweet that has been flagged under our rules against misinformation. This mechanism works to slow the spread of misinformation and provide more context on why the Tweet breaks our Rules.
- Twitter has already seen results from these prompts. The platform recently reported that users are opening articles 40% more often when shown its new 'read before retweeting' message.
- This latest update, as Twitter notes, will apply to all Tweets tagged as containing misleading information, and that small bit of pushback could definitely make more people think twice about re-distributing such messages.

CASE STUDY 3: Prompt to open an article before sharing on Twitter
Starting in September 2020, Twitter added a new prompt when people Retweet an article that they haven't opened on Twitter, we may ask if they would like to open it first. Insights into the effectiveness of the prompt, and how it's changed user behavior when they're shown the alert demonstrated:

- People open articles 40% more often after seeing the prompt
- People opening articles before Retweeting increased by 33%
- Some people didn't end up Retweeting after opening the article

These numbers underline the value of simple prompts like this in getting users to think twice about what it is they're distributing on Twitter.

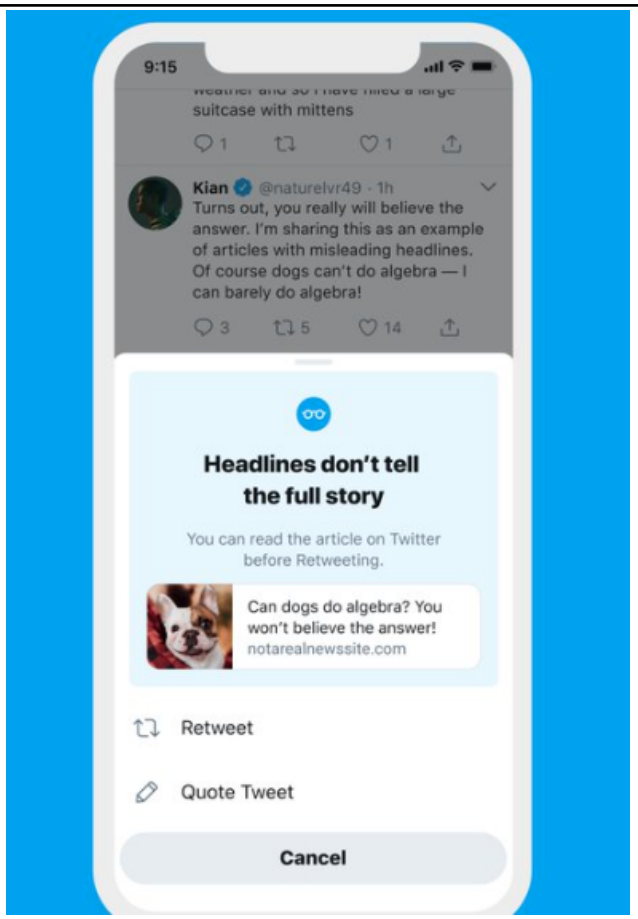

Image 5: The prompt that appears when a user Retweets an article they haven't opened on Twitter.

Policies that combat misleading information and disinformation on Twitter

Twitter has a range of existing, publicly available, policies that outline our definitions, enforcement options and reporting guidelines for inauthentic behaviour, platform manipulation, and misinformation. Our approach to these complex issues is never static. We continually evolve our policies to address new challenges and online behaviours, engaging experts and the public in consultation along the way. The key policies relevant to the operation of the Code are outlined and linked below.

Platform manipulation and spam policy
Users may not use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behavior that manipulates or disrupts people's experience on Twitter.

In line with Twitter's mission to serve the public conversation, we have engaged in long term, proactive work to help people find reliable information, and express themselves freely and safely on our service. Under our platform manipulation and spam policy, we do not allow spam or other types of platform manipulation. We define platform manipulation as using Twitter to engage in bulk, aggressive, or deceptive activity that misleads others and/or disrupts their experience.

Platform manipulation can take many forms and our rules are intended to address a wide range of prohibited behavior, including:

- inauthentic engagements, that attempt to make accounts or content appear more popular or active than they are
- coordinated activity, that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting
- coordinated harmful activity that encourages or promotes behavior which violates the Twitter Rules

It is important to note that while this policy prohibits inauthentic accounts, it does not apply to those using Twitter pseudonymously or as a parody, commentary, or fan account. The ability to speak anonymously, or to use a pseudonym, has been a core tenet of our service since its inception and we believe that the right to remain anonymous online is essential to preserving free expression.

The consequences for violating this policy depend on the severity of the violation as well as any previous history of violations. Our action is also informed by the type of inauthentic activity that we have identified. The actions we take may include the following:

- **Anti-spam challenges**
    - When we detect suspicious levels of activity, accounts may be locked and prompted to provide additional information (e.g. a phone number) or to solve a reCAPTCHA.
- **Denylisting URLs**
    - We denylist or provide warnings about URLs we believe to be unsafe. Read more about unsafe links, including how to appeal if we've falsely identified your URL as unsafe.
- **Tweet deletion and temporary account locks**
    - If the platform manipulation or spam offence is an isolated incident or first offence, we may take a number of actions ranging from requiring deletion of one or more Tweets to temporarily locking account(s). Any subsequent platform manipulation offences will result in permanent suspension.
    - In the case of a violation centering around the use of multiple accounts, users may be asked to choose one account to keep. The remaining accounts will be permanently suspended.
    - If we believe a user may be in violation of our fake accounts policy, we may require that they provide government-issued identification (such as a driver's licence or passport) in order to reinstate their account.
- **Permanent suspension**
    - For severe violations, accounts will be permanently suspended at first detection. Examples of severe violations include:
        - operating accounts where the majority of behavior is in violation of the policies described above;
        - using any of the tactics described on this page to undermine the integrity of elections;
        - buying/selling accounts;
        - creating accounts to replace or mimic a suspended account; and
        - operating accounts that Twitter is able to reliably attribute to entities known to violate the Twitter Rules.

People on Twitter who believe their account was locked or suspended in error, can submit an appeal.

COVID-19 misleading information policy

Users may not use Twitter's services to share false or misleading information about COVID-19 which may lead to harm.

Even 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 rumors, which left uncontextualised can prevent the public from making informed decisions regarding their health, and puts individuals, families, and communities at risk.

In this context, content that is demonstrably false or misleading and may lead to significant risk of harm (such as increased exposure to the virus, or adverse effects on public health systems) may not be shared on Twitter. 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. In

addition, we may label Tweets which share misleading information about COVID-19 to reduce their spread and provide additional context.

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.

- **Content removal**
  - 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
    - 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.
  - 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.
- **Permanent suspension**
  - For severe or repeated violations of this policy, accounts will be permanently suspended.
  - Repeated violations of this policy are enforced against on the basis of the number of strikes an account has accrued for violations of this policy:
    - 1 strike: No account-level action
    - 2 strikes: 12-hour account lock
    - 3 strikes: 12-hour account lock
    - 4 strikes: 7-day account lock
    - 5 or more strikes: Permanent suspension

People on Twitter who believe their account was locked or suspended in error, can submit an appeal.

Civic integrity policy

Users may not use Twitter's services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process. In addition, we may label and reduce the visibility of Tweets containing false or misleading information about civic processes in order to provide additional context.

We also prohibit attempts to use our services to manipulate or disrupt civic processes, including through the distribution of false or misleading information about the procedures or circumstances around participation in a civic process. In instances where misleading information does not seek to directly manipulate or disrupt civic processes, but leads to confusion on our service, we may label the Tweets to give additional context.

The consequences for violating our civic integrity policy depends on the severity and type of the violation and the accounts' 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:

- **Content removal**
  - For high-severity violations of this policy, including (1) misleading information about how to participate, and (2) suppression and intimidation, 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.
- **Profile modifications**
  - If you violate this policy within your profile information (e.g. your bio), we will require you to remove this content. We will also temporarily lock you out of your account before you can Tweet again. If you violate this policy again after your first warning, your account will be permanently suspended.
- **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 content where it appears in the Twitter product;
    - Show a warning to people before they share or like the content;
    - Turn off people's ability to reply, Retweet, or like the Tweet;
    - Reduce the visibility of the content on Twitter and/or prevent it from being recommended.

Synthetic and manipulated media policy

Users may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm ("misleading media"). In addition, we may label Tweets containing misleading media to help people understand their authenticity and to provide additional context. The consequences for violating our synthetic and manipulated media policy depends on the severity of the violation.

- **Tweet Deletion**
  - For high-severity violations of the policy, including misleading media that have a serious risk of harm to individuals or communities, we will require you to remove this content.
- **Labelling**
  - In circumstances where we do not remove content which violates this policy, we may provide additional context on Tweets sharing the misleading media 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
    - Provide a link to additional explanations or clarifications, such as in a curated landing page (Twitter Moments) or relevant Twitter policies.
  - In most cases, we will take a combination 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.
- **Account locks**
  - If we determine that an account has advanced or continuously shares harmful misleading narratives that violate the synthetic and manipulated media policy, we may temporarily reduce the visibility of the account or lock or suspend the account.

Outcome 1c: Users can easily report offending content

Twitter has a number of avenues for people to report potential violations of the Twitter Rules and Terms of Service for review. People can report directly from an individual Tweet, List, or Profile for certain violations, including: including: spam, abusive or harmful content, inappropriate ads, self-harm and impersonation. Other reporting options available in the Twitter Help Centre include spam and system abuse, abusive or harmful content, hateful conduct, child sexual exploitation, violent threats, self harm, inappropriate ads, private information, unauthorised use of a trademark or copyrighted materials, sale or promotion of counterfeit goods, and impersonation. People can report specific content in a Direct Message, Moment, Twitter Space, or person in a Space. Additionally, users can report violations on behalf of another person.

Platform manipulation and spam policy

Under our Platform Manipulation and Spam policy, anyone on Twitter can report accounts or Tweets that violate the criterion defined under the policy or that display inauthentic behaviours, using Twitter's public reporting flow. This is available in-app, on desktop, and via our reporting forms. Respecting that the terms 'disinformation' and 'misinformation' can be unfamiliar to and misunderstood by those without a technical background, our policies clearly outline what inauthentic behaviours look like on Twitter so it's easy to understand the variety of violative content that can be reported and that we can take action on. These reports are then used in aggregate to help refine our enforcement systems and identify new and emerging trends and patterns of behavior.

People using Twitter can also make reports related to Twitter Ads that might potentially violate our policies. These will be assessed against the Twitter Ads Policy, the Twitter Rules, and Terms of Service and any enforcement action will be taken in line with these policies.

In addition to public reporting for platform manipulation, we also have partner reporting mechanisms for other priority areas, such as COVID-19 and Civic Integrity. We provide trusted government partners, public health and electoral authorities access to a Partner Support Portal, a dedicated reporting flow that allows expert and trusted reporters to escalate potentially violative content and access expedited support and human review by our Support teams. By giving experts, who can readily identify the accuracy of certain information, i.e medical misinformation, we are able to address reporting-quality concerns. It is important we can take preventative measures against the potential misuse of public reporting mechanisms as a high-volume of erroneous reports could compromise the efficacy of harm reduction strategies.

Misleading information reporting beta

- In the face of misleading information, we aim to create a better informed world so people can engage in healthy public conversation. We work to mitigate detected threats and also empower customers with credible context on important issues. To help enable free expression and conversations, we only intervene if content breaks our rules. Otherwise, we lean on providing people that use Twitter with additional context.
- This reporting flow is in the beta testing phase, and is currently available in limited testing to some people in Australia, Brazil, Philippines, South Korea, Spain, and the US. These reports are reviewed and acted on independently from other Tweet reporting flows (e.g. for abuse), as this test flow is used to inform our misinformation-related strategy and operations.
- In markets where the feature is available, including Australia, users can report misinformation by clicking the three-dot menu in the upper-right of a tweet, then choosing the "report tweet" option. From there, they'll be able to click the option "it's misleading."

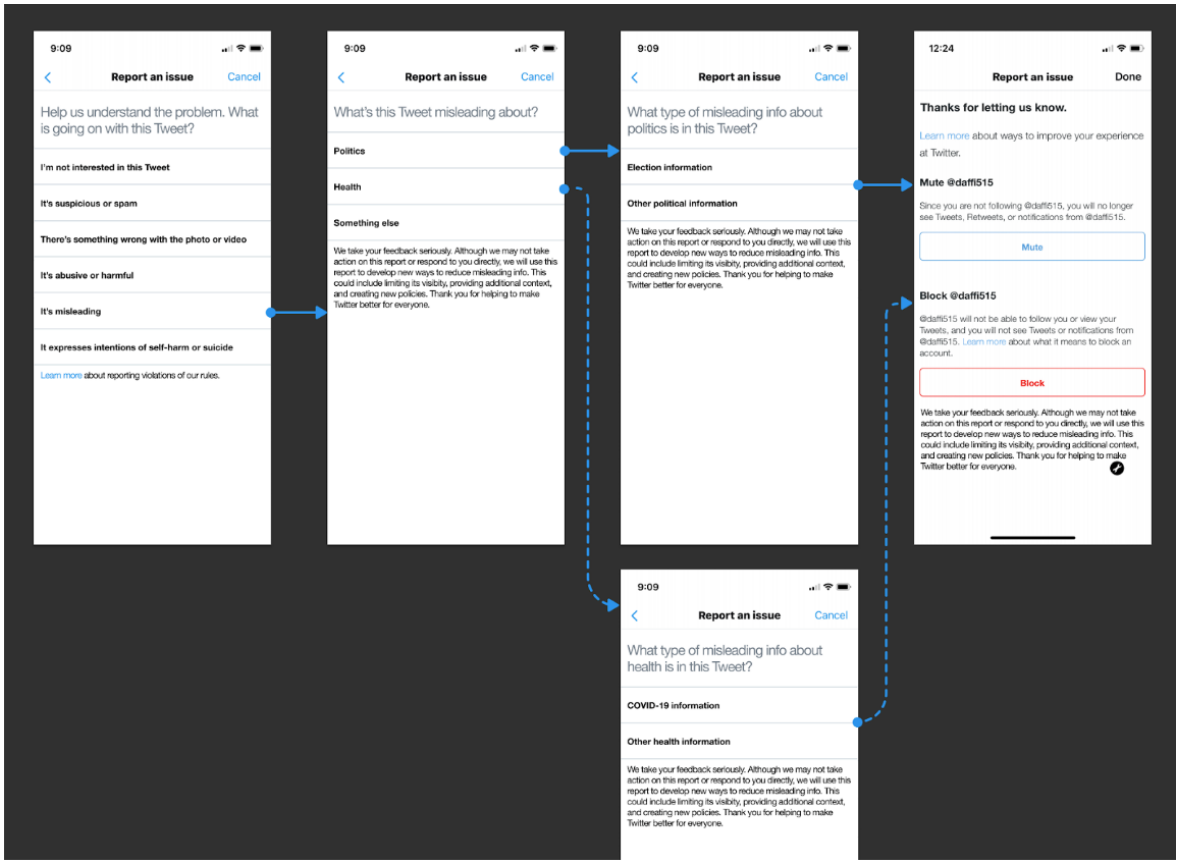

Image 6: Misleading Information Reporting Beta user reporting flow (subject to change).

Outcome 1d: Information about reported content available

Twitter Transparency Reports and relevant data

Twitter has a comprehensive set of policies addressing a wide range of behaviors that are intended to manipulate the public conversation. These behavioral rules captured in our Platform Manipulation and Spam Policy, apply across content types, and serve as the basis for our enforcements against all forms of disinformation and coordinated manipulation. Additionally, Twitter's misinformation policies are focused on identifying content that is demonstrably false or misleading and may lead to significant risk of harm. In line with these priorities, our misinformation policies remained focused on COVID-19, synthetic and manipulated media (SAMM), and civic integrity. Twitter leverages proprietary tools to enforce these policies.

Twitter's approach to enforcement of misinformation and disinformation has evolved over the course of 2020 and 2021. In March 2020, Twitter launched its COVID-19 misleading information policy. Twitter initially enforced this policy through removal of violative content. In May 2020, we expanded its enforcement options to add labels as an option for COVID misinformation. In December 2020, Twitter launched removals of content with, specifically vaccine misinformation, and in March 2021, added labels as an enforcement option for COVID vaccine misinformation.

Twitter has also expanded its enforcement options for violations of the Civic Integrity Policy. In September 2020, Twitter added labels for Civic Integrity violations, in addition to content removal. Additionally in January 2021, Twitter adopted a strike policy for Civic Integrity to address repeat offenders.

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. It should be noted, that while Twitter does not report numbers for enforcement of Twitter's Platform Manipulation or Spam policies, enforcement of these policies are a key cornerstone of Twitter's efforts to combat the spread of

misinformation and disinformation. Twitter challenged approximately 130.3 million accounts to prevent these violations in our reporting period from January to June 2021.

Actions taken to counter spam tend to fluctuate for a variety of reasons, such as the volume of attempted Twitter signups, as well as the volume of spam campaigns targeting our service at any point in time. These anti-spam challenges decreased by approximately 9% compared to the previous reporting period. We believe this can be attributed to ongoing efforts to reduce the impact of anti-spam challenges on legitimate users during this reporting period.

During the latest reporting period from January to June 2021, Twitter globally received 12.9 million accounts reported for review. This represented a decrease of 6% reported accounts compared in the prior transparency reporting period. Reported content is reviewed to determine whether it violates any aspects of the Twitter Rules, independent of its initial report category. For example, content reported under Twitter's private information policy may be found to violate – and be actioned under – our hateful conduct policies. Twitter may also determine that reported content does not violate the Rules at all. The policy categories do not map cleanly to the ones in the Accounts Actioned section below. This is because people typically report content for possible Twitter Rules violations through the Help Centre or in-app reporting.

Twitter is committed to providing due process and to better ensure that the enforcement of the Twitter Rules is fair, unbiased, proportional and respectful of human rights, influenced by the spirit of the Santa Clara Principles on Transparency and Accountability in Content Moderation and other multi stakeholder processes. We will continue to invest in expanding the information available about how we do so in future reports.

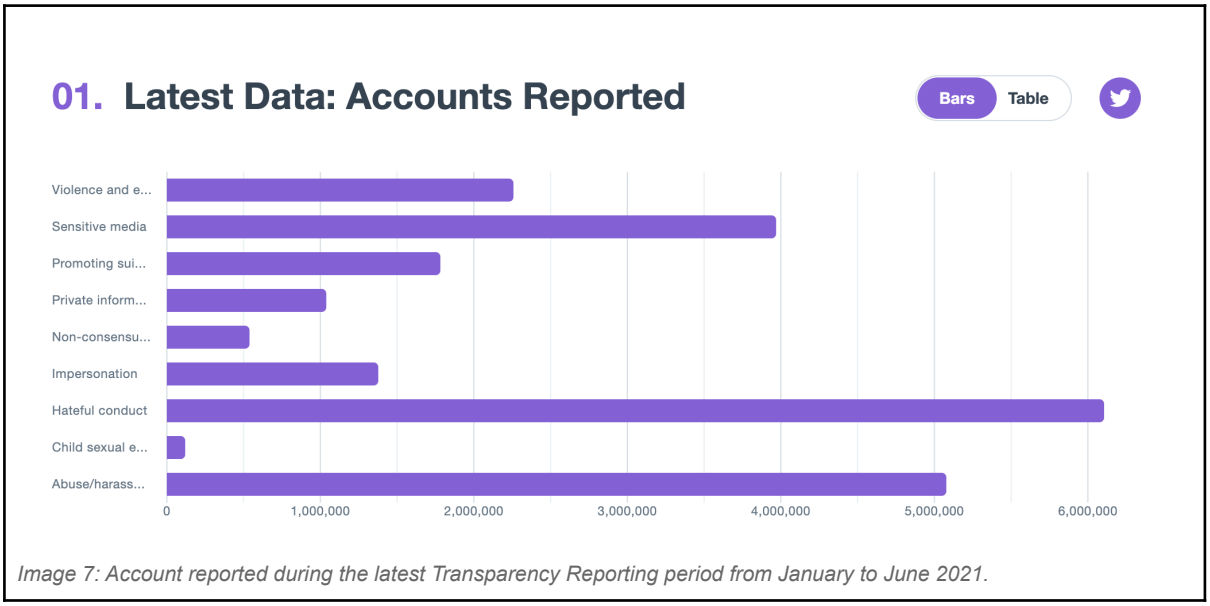

Image 7: Account reported during the latest Transparency Reporting period from January to June 2021.

At the time of this reporting, the following are the approximate global numbers available for the aforementioned time period from January to June 2021:

- **All Twitter Rules** (see the Rules Enforcement page for a list of rules)
  - 4.8 million accounts were actioned for violations of the Twitter Rules.[1]
  - 1.2 million accounts were suspended for violations of the Twitter Rules.[2]

[1] "Accounts actioned" reflects the number of unique accounts that were suspended or had some content removed for violating the Twitter Rules. This does not include labels applied.
[2] "Accounts suspended" reflects the number of unique accounts that were suspended.

    - 5.9 million pieces of content were removed for violations of the Twitter Rules.[3]
    - This represents a 36% increase in accounts actioned, a 23% increase in accounts suspended, and a 32% increase in content removed as compared to the last Twitter Transparency reporting period from January to June 2021.
- **COVID-19 misleading information** (see the COVID-19 Misinformation page)
    - In April 2022:
        - 2,135 accounts were challenged for violations of the policy.
        - 1,329 accounts were suspended for violations of the policy.
        - 5,320 pieces of content were removed for violations of the policy.
    - Since January 2020 when the pandemic began, Twitter has challenged 11.72M accounts, suspended 8,126 accounts, and removed 83,999 pieces of content that violated our COVID-19 misleading information policy.
- **Civic integrity policy** (see the Civic integrity policy page)
    - 581 accounts were actioned for violations of the civic integrity policy.
    - 23 accounts were suspended for violations of the civic integrity policy.
    - 593 pieces of content were removed for violations of the civic integrity policy.

For purposes of this report, we have also compiled the Australia-specific approximate data for January to June 2021.[4]

- **All Twitter Rules** (see the Rules Enforcement page for a list of rules)
    - 39,607 Australian accounts were actioned for violations of the Twitter Rules.
    - 7,851 Australian accounts were suspended for violations of the Twitter Rules.
    - 51,394 pieces of content authored by Australian accounts were removed for violations of the Twitter Rules.
- **COVID-19 misleading information**
    - 817 Australian accounts were actioned for violations of the policy.
    - 35 Australian accounts were suspended for violations of the policy.
    - 1,028 pieces of content authored by Australian accounts were removed for violations of the policy.
- **Civic integrity policy**
    - 6 Australian accounts were actioned for violations of the civic integrity policy.
    - We do not have any recorded data of Australian accounts being suspended for violations of the civic integrity policy during the specified reporting period.
    - 6 pieces of content authored by Australian accounts were removed for violations of the civic integrity policy.

The next Twitter Transparency Report is scheduled to be published in the second half of 2022, which will cover the period from July to December 2021.

Twitter reporting options

We believe that transparency is a key principle in our mission to protect the Open Internet, and advancing the Internet as a global force for good. As outlined above, the Twitter Transparency Centre provides data about reports received and actions taken on content that violates Twitter policies, including sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, and information operations. This is an ongoing reporting scheme for Twitter and we will continue to share updates on the trends in violative content on our service and our enforcement.

People who report potentially violative content on Twitter will also receive a response directly from our Support teams about the results of our investigation and any relevant enforcement action taken.

Enforcement of Twitter Rules

[3] "Content removed" reflects the number of unique pieces of content (such as Tweets or an account's profile image, banner, or bio) that Twitter required account owners to remove for violating the Twitter Rules. The country value presented here relates to the country associated with the account that created the violative content.

[4] Countries are assigned by inferring an account's country based on the following data from a user; sign-up location, user selected location and IP addresses. While every effort is made to present accurate data, we cannot guarantee that the assigned country reflects the true location of a user at a specific instance when an enforcement action occurred.

In addition to our work to combat misleading information and disinformation, Twitter promotes the health of the public conversation through scaling our efforts to remove content that violates the Twitter Rules before it can be widely distributed. For example, in our latest Transparency Report from 1 January 2021 to 30 June 2021, Twitter reported that it globally took action on 4.8M Tweets that violated the Twitter Rules, requiring account holders to remove violative content. Of the Tweets removed, 68% received fewer than 100 impressions prior to removal, with an additional 24% receiving between 100 and 1,000 impressions. In total, impressions on these violative Tweets accounted for less than 0.1% of all impressions for all Tweets during that time period.

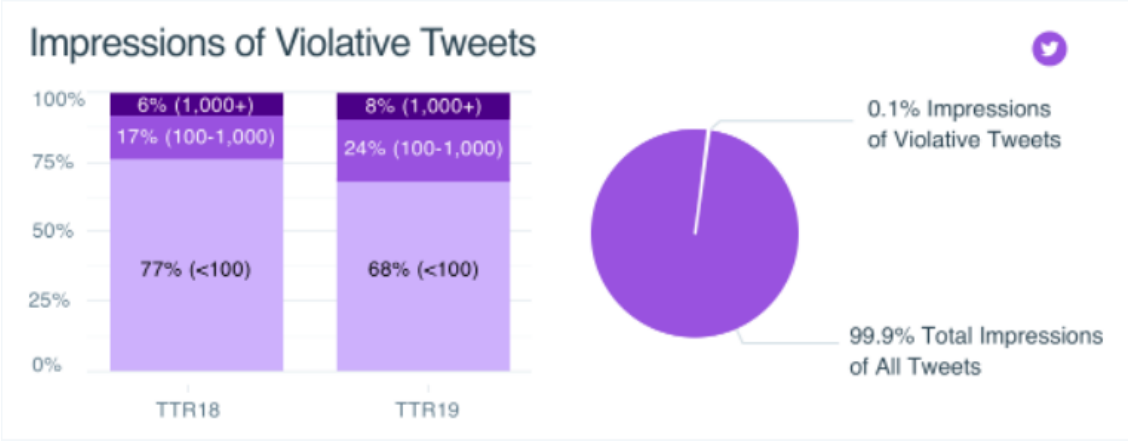

Image 8: Impressions of violative Tweets from the Twitter Transparency Report from 1 January 2021 to 30 June 2021.

Additional information regarding insights into information requests and removal requests originating from Australia can be found in our Transparency Centre here.

Objective 2: Disrupt advertising and monetisation incentives for disinformation

Promoted content on Twitter must also adhere to our existing Twitter Rules. In addition, we publish specific policies for advertisers that share standards for that are outlined below.

Political content advertising policy

Twitter globally prohibits the promotion of political content under our political content advertising policy. We have made this decision based on our belief that political message reach should be earned, not bought.

Inappropriate content advertising policy

Our policy on inappropriate content advertising prohibits advertising deemed to be dangerous or exploitative, misrepresentative, along with misleading synthetic or manipulated content and content engaged in coordinated harmful activity.

Quality advertising policy

Our quality advertising policy outlines standards for advertisers including that ads should represent the brand or product being promoted and cannot mislead users into opening content by including exaggerated or sensationalised language or misleading calls to action.

Demonetisation of misleading information

Twitter automatically demonetises publisher content monetised through the Amplify Pre-Roll program that receives a misleading information label. Tweets receiving this label also cannot be promoted as ads under our Inappropriate Content policy.

People using Twitter can also make reports related to Twitter Ads that might potentially violate our policies. These will be assessed against the Twitter Ads Policy, the Twitter Rules and Terms of Service and any enforcement action will be taken in line with these policies.

State-affiliated media

State-affiliated media may not purchase advertisements on Twitter per our ads content policies (and within Twitter Ads Policy overall). This policy extends to individuals reporting on behalf of, or who are directly affiliated with such entities. State-affiliated media (as above) is defined as outlets where the state exercises control over editorial content through financial resources, direct or indirect political pressures, and/or control over production and distribution. Unlike independent media, state-affiliated media frequently use their news coverage as a means to advance a political agenda.

Climate-forward approach to ads

People around the world use Twitter to connect with others passionate about protecting our planet. Last year, we introduced a dedicated Topic to help people find personalised conversations about climate change, and to support conversation around #COP26, we rolled out pre-bunks — hubs of credible, authoritative information across a range of key themes, like the science backing climate change, made available in the Explore tab, Search, and Trends.

To better serve these conversations, Twitter recently announced that misleading advertisements on Twitter that contradict the scientific consensus on climate change are prohibited, in line with our inappropriate content policy. We believe that climate denialism shouldn't be monetised on Twitter, and that misrepresentative ads shouldn't detract from important conversations about the climate crisis.

This approach is informed by authoritative sources, like the Intergovernmental Panel on Climate Change Assessment Reports. We recognise that misleading information about climate change can undermine efforts to protect the planet. In the coming months, we'll have more to share on our work to add reliable, authoritative context to the climate conversations happening on Twitter.

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.

Expanding access beyond information operations

In October 2018, we published the first comprehensive, public archive of data related to state-backed information operations. Since then, we've shared 37 datasets of attributed platform manipulation campaigns originating from 17 countries, spanning more than 200 million Tweets and nine terabytes of media. More than 26,000 researchers have accessed these datasets, empowering an unprecedented level of empirical research into state-backed attacks on the integrity of the conversation on Twitter.

We strive to provide timely updates, alongside comprehensive data, whenever our teams identify and remove these campaigns, however, this year, due to technical issues and significant risks to the physical safety of our employees posed by certain disclosures, we have only provided one update. During this time, we've been working to identify a sustainable path forward, without compromising on our goals of providing meaningful transparency.

As we advance data-driven transparency in 2022 and beyond, we shared a number of important lessons learned in our endeavours:

- **Meaningful transparency begins with access to data**. The data we publish about information operations allows researchers to understand not just that a platform manipulation

campaign took place and that Twitter removed it — but precisely which narratives that campaign aimed to advance, and how widely they spread on Twitter. Access to raw content, rather than limited samples and aggregate information, is important.
- **Raw data isn't accessible to everyone**. Many of the datasets we've released include hundreds of thousands of Tweets and gigabytes of media. Processing this information often requires advanced tooling and capabilities. Academics, independent researchers, NGOs and data journalists play a key part in translating raw data into meaningful insights, as well as providing critical context in understanding how bad actors operate. Partnerships with the Stanford Internet Observatory and Australia Strategic Policy Institute have helped put these datasets in analytic and narrative context, along with a conference dedicated to studying this data we held in conjunction with the Carnegie Institute.
- **Confident attribution isn't always possible**. Our transparency approach has focused on activity we can confidently attribute to a state actor. Emergent behaviors, including the use of disinformation-for-hire vendors and increasing operational security, sometimes make confident attribution impossible based solely on Twitter's own data. This doesn't make the activity in question less important to analyze, but our policies presently prevent dataset disclosure in these cases. Moreover, access to this data, without attribution, may allow experts to piece together operations across multiple platforms and services that is not possible by just one company.
- **Information operations are just one area of public concern**. We've provided an unprecedented level of transparency about state-backed information operations, given their severe impact on public discourse around the world. As Camille François and evelyn douek have pointed out, other content moderation domains of equal public concern don't receive the same treatment.

Building off these learnings, we're changing our approach in an effort to continue to provide expanded transparency about our content moderation actions. Here's what you'll see in the coming months:

- In line with our long term commitment to principles of transparency, and to improve public understanding of inauthentic influence and manipulation campaigns, in 2022 we will launch the Twitter Moderation Research Consortium (TMRC) that will create a new global expert group of academics, members of civil societies, NGOs, and journalists to study platform governance issues.
- Membership in the consortium will be granted to groups or individuals with:
    - A proven track record of research on content moderation and integrity topics (or affiliation with a group that does such research, such as a university, research lab, or newspaper).
    - Appropriate plans and systems for safeguarding the privacy and security of the data provided by the consortium.

We will be fully public about the standards used to determine membership in the consortium, and will bias towards inclusion and access, particularly for emerging researchers and researchers from historically under-represented communities and parts of the world.

Twitter will not exercise any control or judgment over the findings or focus areas of the research produced using this data by members of the consortium.

The more than 200 researchers around the world with existing access to our unhashed information operations datasets will be invited to join the consortium through an expedited process. Other qualifying individuals and institutions are welcome to apply. We will share additional details about this process in early 2022 in advance of any disclosures to the consortium.

We will provide comprehensive data about attributed platform manipulation campaigns to members of the consortium, who may independently choose to publish their findings on the basis of the data we share and their own research. Under this model, we will also begin to share data about platform manipulation campaigns for which we have not been able to arrive at confident attribution to a state actor, and campaigns where we are unable to provide broad access due to employee safety concerns.

Later in 2022, we will for the first time share similarly comprehensive data about other policy areas, including misinformation, coordinated harmful activity, and safety.

As part of this change, we will discontinue our fully public dataset releases, prioritising release to the consortium. Existing datasets will continue to be available for download indefinitely — and our public data offerings, including free access to our APIs (including the full archive of Tweets) remain available.

Approach to monitoring performance

Meaningful transparency between companies, regulators, civil society, and the general public is fundamental to the work we do at Twitter. This transparency is a key tenet of our efforts to preserve and protect the Open Internet. In line with this philosophy, for the past ten years our biannual Twitter Transparency Report has highlighted trends in requests made to Twitter from around the globe.

We believe the open exchange of information can have a positive global impact and through our efforts to provide meaningful transparency, we endeavour to earn public trust, and enable accountability. Recognising that the public as well as policymakers and regulators want to be better informed of our enforcement processes, we launched the new Twitter Transparency Centre in 2020 to make our data easier to understand and more comprehensive to cover a broader array of our transparency efforts. Over time, we have significantly expanded the information we disclose. We now include sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, and state-backed information operations.

Our goal with these changes is to provide more transparency about more issues, while grappling with the considerable safety, security, and integrity challenges in this space. We'll continue to learn and iterate on our approach over time and share those findings publicly along the way.

Objective 4: Empower consumers to make better informed choices of digital content.

Twitter recognises the importance of helping users identify trusted information and make informed choices in today's information ecosystem. As outlined above, we have adopted a number of strategies to signal authenticity, accuracy and authoritative information on the Twitter service. These include Explore tab, Moments, Search Prompts, and Verification, as well as labelling government and state-sponsored media accounts so that users can evaluate the source of information.

We provide this context both through Twitter's Curation team and through trusted partnerships with government, experts, and civil society groups. Our Curators don't act as reporters or creators of original work. They organise and present content that already exists on Twitter in Moments, explanatory content on Trends, in Lists and more. They are guided by publicly available principles, including impartiality and accuracy, which is available on our Help Centre.

About government and state-affiliated media account labels on Twitter

Twitter is where people come to see what's happening and to hear from their governments and government officials. We believe that safety and free expression go hand-in-hand, especially when interacting with these leaders and associated institutions, and adding context to what people see on Twitter helps them have a more informed experience on Twitter.

Since our last report under the Australian Code of Practice on Disinformation and Misinformation, we have made additional progress on our state media and government account labelling efforts.

As reported last year, in 2020, we began applying account labels to two categories of state-affiliated accounts: (1) the accounts of key government officials; and (2) accounts belonging to state-affiliated media entities. Included in this initial action were accounts from countries represented in the five permanent members of the United Nations Security Council.

Then in 2021, we expanded these labels to accounts from Group of Seven (G7) countries, and to a majority of countries to which Twitter has attributed state-linked information operations. We also apply labels to the personal accounts of heads of state for these countries. We regularly report such state-linked information operations with a dedicated section, Information Operations, within the Twitter Transparency Report.

What are government and state-affiliated media account labels?

- Labels on state-affiliated accounts provide additional context about accounts that are controlled by certain official representatives of governments, state-affiliated media entities and individuals closely associated with those entities.
- The label appears on the profile page of the relevant Twitter account and on the Tweets sent by and shared from these accounts. Labels contain information about the country the account is affiliated with and whether it is operated by a government representative or state-affiliated media entity.
- Additionally, these labels include a small icon of a flag to signal the account's status as a government account and of a podium for state-affiliated media.

Image 9: Example of state-affiliated media labels on Twitter.

How government accounts are defined

Our focus is on senior officials and entities that are the official voice of the nation state abroad, specifically accounts of key government officials, including foreign ministers, institutional entities,

ambassadors, official spokespeople, and key diplomatic leaders. Where accounts are used solely for personal use and do not play a role as a geopolitical or official Government communication channel, we will not label the account.

How state-affiliated media accounts are defined

State-affiliated media is defined as outlets where the state exercises control over editorial content through financial resources, direct or indirect political pressures, and/or control over production and distribution. Accounts belonging to state-affiliated media entities, their editors-in-chief, and/or their senior staff may be labelled.

State-financed media organisations with editorial independence, like the BBC in the UK or NPR in the US for example, are not defined as state-affiliated media for the purposes of this policy.

Which accounts currently have a label?

Currently, labels appear on relevant Twitter accounts from China, France, Russia, the United Kingdom, the United States, Belarus, Canada, Germany, Italy, Japan, Cuba, Ecuador, Egypt, Honduras, Indonesia, Iran, Saudi Arabia, Serbia, Spain, Thailand, Turkey, Ukraine, and the United Arab Emirates that are:

- Government accounts heavily engaged in geopolitics and diplomacy
- State-affiliated media entities;
- Individuals, such as editors or high-profile journalists, associated with state-affiliated media entities;
- This policy will be expanded to include additional countries in the future.

Additionally, labels will distinguish between individual government accounts and institutional government accounts. In terms of functionality with state-affiliated media entities, Twitter will not recommend or amplify accounts or their Tweets with these labels to people.

Currently, labels are not applied to Australian government accounts; however, our teams are continuing to expand this feature across the service in the future.

Our approach with labels in Ukraine

In February 2022, we expanded our approach to state accounts by adding labels to Tweets that share links to Russian state-affiliated media websites. Much of the content from state-affiliated media came from these link shares, and not just from the accounts we'd been labelling as state-affiliated media.

Since 28 February 2022, we've labelled more than 260,000 unique Tweets in this expanded category. As is standard with our labels, these Tweets are not eligible for amplification, meaning they aren't recommended in a user's Home Timeline, Notifications, and other places on Twitter. Our interventions have made a difference, contributing to a 30% reduction of the reach of this content.

We've now begun adding additional labels to multiple state-affiliated media accounts from Belarus. We have also labelled one state-affiliated media account in Ukraine. As the conflict – and online conversation – evolves, we want to equip people on Twitter with context and enable informed experiences on the service. We believe that people have the right to know when a media account is affiliated with a state actor.

During periods of conflict, enabling access to factual, reliable information, and making clear which accounts are controlled by participating states, is critical. To serve this mission, we'll apply government and state-affiliated media labels to states engaged in inter-state conflict, in addition to those countries already covered by our existing work. The policy will help ensure we add context to conversations about global conflicts equitably. We intend to expand on our approach, labelling more countries over time.

The recently announced European Union (EU) sanctions legally require Twitter to withhold certain state-affiliated media content in EU member states, and we are complying with this requirement. Our global approach outside of the EU, including Australia, continues to focus on de-amplifying this type of state-affiliated media content across our service and providing important context through our labels.

Twitter's relevant policies regarding misleading information

Additionally, Twitter's relevant policies and strategies for this section are the same as our policies outlined under Objective 1b. For reference, an abridged version of these policies is included below.

Platform manipulation and spam policy

- Users may not use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behavior that manipulates or disrupts people's experience on Twitter. While this policy prohibits fake accounts, it does not apply to those using Twitter pseudonymously or as a parody, commentary, or fan account. The ability to speak anonymously, or to use a pseudonym, has been a core tenet of our service since its inception and we believe that the right to remain anonymous online is essential to preserving free expression.

COVID-19 misleading information policy

- Users may not use Twitter's services to share false or misleading information about COVID-19 which may lead to harm. In this context, content that is demonstrably false or misleading and may lead to significant risk of harm (such as increased exposure to the virus, or adverse effects on public health systems) may not be shared on Twitter. 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. In addition, we may label Tweets which share misleading information about COVID-19 to reduce their spread and provide additional context.

Civic integrity policy

- Users may not use Twitter's services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process. In addition, we may label and reduce the visibility of Tweets containing false or misleading information about civic processes in order to provide additional context.

- We also prohibit attempts to use our services to manipulate or disrupt civic processes, including through the distribution of false or misleading information about the procedures or circumstances around participation in a civic process. In instances where misleading information does not seek to directly manipulate or disrupt civic processes, but leads to confusion on our service, we may label the Tweets to give additional context.

Synthetic and manipulated media policy

- Users may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm ("misleading media"). In addition, we may label Tweets containing misleading media to help people understand their authenticity and to provide additional context.

Objective 5: Improve public awareness of the source of political advertising carried on digital platforms

Objective 5 is not applicable to Twitter.

Political content policy

Since 2019, Twitter globally prohibits the promotion of political content under our Political Content advertising policy. We have made this decision based on our belief that political message reach should be earned, not bought.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research

In line with our commitments to transparency, Twitter is the only major service to make public conversation data proactively available via an application programming interface (API) for the purposes of research. By harnessing the power of the Twitter API, partners are able to tap into the public conversation and study collective issues facing global communities to bring about new insights to universal issues, devise fresh approaches to problems, and foster social good. Research conducted with the Twitter API must adhere to the Twitter Developer Policy, which is linked in our publicly available information about our approach to providing academic access to data.

Transparency is core to Twitter's approach. Through initiatives such as our open developer platform, our information operations archive, and our disclosures in the Twitter Transparency Center and Lumen, we continue to support third-party research of what's happening on Twitter. We'll continue to build on these efforts and inform the public as we improve Twitter in the open. The following are highlights from the past year:

- Twitter API for Academic Research: In early 2021, we launched a dedicated Academic Research product track on the new Twitter API giving qualified researchers access to the entire history of public conversation and elevated access to real-time data for free. This track provides qualified academics the opportunity to access new endpoints, including the full history of public conversation data, a higher volume of Tweets, and more precise filtering capabilities.
- Algorithmic bias bounty challenge: When we introduced our commitment to responsible machine learning, we also said, "the journey to responsible, responsive, and community-driven machine learning systems is a collaborative one." That's why we introduced the industry's first algorithmic bias bounty competition to draw on the global ethical AI community's knowledge of the unintended harms of saliency algorithms to expand our own understanding and to reward the people doing work in this field.
- Twitter Moderation Research Consortium (TMRC): As noted in Objective 3, Twitter announced the creation of a new global expert group of academics, members of civil societies and NGOs, and journalists to study platform governance issues.
- Launch of an API curriculum: "Getting started with the Twitter API for Academic Research" is now being used at universities, enabling students and teachers to learn how to use Twitter data for academic research. It is currently starred by over 200 academics on Github.
- Creation of a Developer Platform Academic Research advisory board: This group of 12 scholars began work with our team this year to better understand how we can enhance the use of the Twitter API for academic research, while increasing meaningful dialogue between the Twitter Academic program and the academic community.
- Developer research highlights: We published and continued to spotlight key research areas Twitter teams are working on today in an effort to inspire even more researchers to pursue these topics.
- We have also partnered with non-government organisations on global awareness campaigns and initiatives, such as UNESCO for the evergreen, custom emoji activated by the #ThinkBeforeSharing hashtag. #ThinkBeforeSharing aimed to increase comprehension and

media literacy and help people learn how to identify, debunk, react to and report on conspiracy theories to prevent their spread.

In line with our principles of transparency and to improve public understanding of inauthentic influence campaigns, as mentioned above, Twitter has also published public archives of Tweets and media that we believe resulted from state-backed information operations. We have collaborated with research and civil society partners to increase access, transparency and meaningful interpretation of this information, including 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. Using this data, ASPI produced a number of reports – including Tweeting through the Great Firewall and Retweeting Through the Great Firewall – and an interactive website that takes people through their analysis of the publicly accessible data from Twitter's Information Operations Archive.

Objective 7: Signatories will publicise the measures they take to combat Disinformation

We have put extensive work into updating, developing, and educating our users on Twitter's rules and enforcement actions. This ultimately supports and improves public understanding of the wide variety of inauthentic behaviours that can be addressed to protect the integrity of our service.

In addition to committing to annual reporting under the Code, and updating our policies as our rationales, approaches, or enforcement options evolve, we will also continue to disclose data under the Twitter Transparency Centre. This data is available and open to analysis for government, Twitter users, and the general public. We adopted transparency and open data principles with the establishment of these initiatives and aim to continually improve their accessibility and usefulness to the public by publicising our disclosures on the Twitter blog and through media activity in each reporting period and continually evaluating how easily the language and structure of our reporting can be understood by external audiences.

As we look toward the future, Twitter is working to increase transparency and understanding on our approach to content moderation. As we continue to invite trusted partners and the public to share feedback on ways to make Twitter safe, it's important to be transparent about how we develop and enforce the Twitter Rules.

Our newly formed Content Governance Initiative (CGI) aims to do this by developing a governance framework that provides a consistent and principled approach to the development, enforcement, and assessment of our global rules and policies. To build our governance framework, we're engaging external stakeholders and have created an additional advisory group on our Trust and Safety Council. We'll continue collaborating with this group and cross-functional teams across Twitter to establish standardised guidelines on policy development, enforcement, and appeals that help drive a common understanding of Twitter's approach to content moderation. The framework's principles and guidelines will aim to fulfill the following objectives:

- Build legitimacy and trust through transparency and accountability.
- Deepen our commitment to good governance and human rights.
- Provide additional clarity on Twitter's content moderation processes.
- Affirm our commitment to serving a diverse and inclusive global community.

We recognise that achieving these objectives will not be easy. Content moderation at scale is a highly complex and challenging process. This initiative reflects our ongoing commitment to working systematically — in partnership with external stakeholders around the world — to improve the transparency and consistency of our content moderation processes.

Concluding remarks

Twitter's mission is to serve the public conversation. Trust in the information we consume every day is critical to how we engage online. Trust in information found online can be established and enhanced when companies, civil society and regulators are transparent about the data and processes they rely on. This, alongside thoughtful policies developed by governments around the world, can help elevate many of the core principles of the Open Internet. For our part, since 2012 we have consistently published bi-annual updates, detailing our enforcement actions housed in the Twitter Transparency Center.

Coupled with this, we have been explicit in our advocacy for an Open Internet that is global, available to all, built on open standards and rooted in the protection of human rights. As we've said, protecting the Open Internet — which continues to be under threat — requires meaningful transparency, which is essential to holding companies and governments to account. This work is core to who we are as a company – connecting back to the very first Twitter Transparency Report in 2012, one of the first such reports in the industry.

Our approach, as outlined in this report, remains closely aligned with our company values and the guiding principles of the Code, particularly protecting freedom of expression. We trust this overview of our work to date and future commitments under the Code, provide an understanding of the serious resolve with which our teams approach protecting the integrity of public conversation.

We look forward to continuing our work with government and academic partners, as well as across industry, to improve public understanding of these complex issues and to take meaningful steps that protect our service, the people who use it, and the Open Internet.