Aotearoa New Zealand Code of Practice for Online Safety and Harms

Annual Compliance Report

October 2022

Signatory:	Twitter, Inc.

***If applicable*:** **Relevant Products / Services:**	**Summary** Twitter's mission is to serve the public conversation. Transparency is fundamental to our work in achieving that mission. This inaugural annual compliance report outlines Twitter's commitments and progress under the Aotearoa New Zealand Code of Practice for Online Safety and Harms (the Code) to demonstrate Twitter's efforts to protect the public conversation and uphold the integrity of our service. Twitter is a singular platform and service. As a global platform for public conversation, Twitter 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 committed to providing meaningful transparency reporting to the public. Under the Code, Twitter has made meaningful commitments and progress on all of the outcomes. 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 Twitter Moderation Research Consortium. 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.

4.1 Reduce the prevalence of harmful content online

Signatories commit to implementing policies, processes, products and/or programs that would promote safety and mitigate risks that may arise from the propagation of harmful content online, as it relates to the themes identified in section 1.4.

Outcome 1. Provide safeguards to reduce the risk of harm arising from online **child sexual exploitation & abuse (CSEA)**

Measure 1. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent known child sexual abuse material from being made available to users or accessible on their platforms and services

- Twitter has zero tolerance towards any material that features or promotes child sexual exploitation, one of the most serious violations of the Twitter Rules. This may include media, text, illustrated, or computer-generated images. Regardless of the intent, viewing, sharing, or linking to child sexual exploitation material contributes to the re-victimisation of the depicted children. This also applies to content that may further contribute to victimisation of children through the promotion or glorification of child sexual exploitation. For the purposes of this policy, a minor is any person under the age of 18.

Measure 2. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent search results from surfacing child sexual abuse material

- Twitter aggressively fights online child sexual abuse and has heavily invested in technology and tools to enforce our policy. We use a combination of machine learning and human review — our systems are able to surface content to human moderators who use important context to make decisions about potential rule violations. This work is led by an international, cross-functional team with 24-hour coverage and the ability to cover multiple languages. We also have an appeals process for any potential errors that could occur. The consequence for violating our child sexual exploitation policy is immediate and permanent suspension. In addition, violators will be prohibited from creating any new accounts in the future.

Measure 3. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to adopt enhanced safety measures to protect children online from peers or adults seeking to engage in harmful sexual activity with children (e.g. online grooming and predatory behaviour)

- Twitter tracks and reports on how and when we enforce our policies, and reports of potential violations under the Twitter Rules every six months as a part of the Twitter Transparency Report (see Outcome 10 and 11 for additional information).
- In Twitter's latest Transparency Report from July to December 2021, we suspended 596,997 unique accounts during this reporting period for violating our child sexual exploitation policy. Of these, 91% of suspended accounts were identified proactively by employing internal proprietary tools and industry hash sharing initiatives. These tools and initiatives support our efforts in surfacing potentially violative content for further review and, if appropriate, removal.

For purposes of this report, we have also compiled the **New Zealand-specific approximate data for July to December 2021.**[1] These New Zealand-specific figures and those denoted as such throughout the report should be considered relative to the size and scope of New Zealand's population and market.

- 180 accounts were reported for violations of Twitter's child sexual exploitation policy [2]
- 770 accounts were actioned for violations of Twitter's child sexual exploitation policy [3]
- 769 accounts were suspended for violations of Twitter's child sexual exploitation policy [4]
- 10 pieces of content were removed for violations of Twitter's child sexual exploitation policy.[5]

Measure 4. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to reduce new and ongoing opportunities for the sexual abuse or exploitation of children

- In November 2021, we expanded on our aggressive approach to fighting child sexual exploitation and have enabled for people on Twitter to report CSE via the in-app reporting flow as an additional avenue for us to detect and take down this content. We've always supported CSE reporting via a dedicated form — this additional tool increases our ability to take down content faster and reinforces our commitment to making Twitter a safer place.
- In addition, we partner with governments and law enforcement to facilitate investigations and prosecutions and take an industry-wide approach to tackle this issue. We partner with organisations around the globe in this area, including the National Centre for Missing and Exploited Children (NCMEC) and the International Association of Internet Hotlines (INHOPE). When we remove content, we immediately report it to NCMEC and reports are made available to the appropriate law enforcement agencies around the world to facilitate investigations and prosecutions.

Measure 5. Work to collaborate across industry and with other relevant stakeholders to respond to evolving threats

- Twitter has a long-standing collaboration with the NCMEC. We are active members of several coalitions, such as the Technology Coalition, the ICT Coalition, the WeProtect Global Alliance, INHOPE and the Fair Play Alliance, that bring companies and non-government organisations (NGOs) together to develop solutions that disrupt the exchange of child sexual abuse materials online and prevent the sexual exploitation of children.
- We support companies of all sizes, including those just establishing their child safety protocols and processes. The Tech Coalition brings together companies across the

[1] **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. For definitions of Twitter Rules, see https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec

[2] **"Accounts reported"** reflects the total number of unique accounts that users reported for potentially violating the Twitter Rules.

[3] "**Accounts actioned"** reflects the number of unique accounts that were suspended or had some content removed for violating the Twitter Rules.

[4] "**Accounts suspended"** reflects the number of unique accounts that were suspended for violating the Twitter Rules.

[5] "**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.

technology industry. Together we tackle risks to online child safety through sharing best practices, mentorship, and coordinated efforts to improve the detection and reporting of sexual abuse imagery and other exploitative practices that put children at risk. As the Coalition enters its 16th year, Twitter has continued to collaborate closely with partners to address new and emerging challenges through *Project Protect: A plan to combat online child sexual abuse to prevent and eradicate online CSE*.

Outcome 2: Provide safeguards to reduce the risk of harm arising from online **bullying or harassment**

Measure 6. Implement, enforce and/or maintain policies and processes that seek to reduce the risk to individuals (both minors and adults) or groups from being the target of online bullying or harassment.

- Twitter's mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right – we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives.
- We recognise that if people experience abuse on Twitter, it can jeopardise their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. This includes women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalised, and historically underrepresented communities. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature, and more harmful.
- We are committed to combating abuse motivated by hatred, prejudice, or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalised. For this reason, we prohibit behaviour that targets individuals or groups with abuse based on their perceived membership in a protected category.
- Under our abusive behaviour policy, a user 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. 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. For behaviour targeting people based on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease, this may be in violation of our hateful conduct policy (see Outcome 3).
- At Twitter, we recognise that suicide and self-harm are significant social and public health challenges that require collaboration between all stakeholders – public, private, and civil society – and that we have a role and responsibility to help people access and receive support when they need it.
- Our enforcement approach depends on the type of content being shared, whether or not the reported account is encouraging or promoting self-harm or suicide, and the account's previous history of violations. If a user violates this policy by sharing content that intentionally encourages others to harm themselves, ask others to encourage you to harm yourself, or share detailed information or instructions related to self-harm or suicide methods, we will require the user to remove this content. We will also temporarily lock them out of their account before they can Tweet again. If the user continues to violate this

policy, or if their account is dedicated to promoting or encouraging self-harm or suicide, the account will be permanently suspended. If cases include images or videos related to self-harm or suicide, we will also evaluate this content under our sensitive media policy (see Outcome 5).
- We may also take steps to prevent the spread of instructional material hosted on third-party websites by marking such links as unsafe.

Measure 7. Implement and maintain products and/or tools that seek to mitigate the risk of individuals or groups from being the target of online bullying or harassment.

- Twitter strives to provide an environment where people can feel free to express themselves. If abusive behaviour happens, we want to make it easy for people to report it to us. Multiple Tweets can be included in the same report, helping us gain better context, while investigating the issues to get them resolved faster.
- **Reporting content**: Twitter has options where people can report violations on behalf of another person. In some situations, to help our teams understand the context, we sometimes need to hear directly from the person being targeted to ensure that we have the information needed prior to taking any enforcement action. Individuals do not need to be a member of a specific protected category for us to take action. We will never ask people to prove or disprove membership in any protected category, and we will not investigate this information.
- **Enforcement**: When determining the penalty for violating Twitter's hateful conduct and abusive behaviour policies, 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 hateful conduct as defined in this policy, or who have shared violent threats.
- In our latest Transparency Report from July to December 2021, Twitter globally took action on 940,679 accounts for violating our Abusive Behaviour policy, which prohibits content that harasses or intimidates, or is otherwise intended to shame or degrade others. This is a 10% decrease from our last report and is in line with a 11% decrease in accounts reported under this policy during this period. Additionally, during this reporting period, there was a substantial increase in the volume of accounts suspended (18%), and content removed (23%) under our Suicide and Self-harm policy with 408,143 accounts actioned in total. We attribute these changes to our continued investment in identifying violative content at scale.

  New Zealand-specific approximate data for July to December 2021:
    - 6,762 accounts were reported for violations of the Abusive Behaviour policy
    - 1,541 accounts were actioned for violations of the Abusive Behaviour policy
    - 126 accounts were suspended for violations of the Abusive Behaviour policy

- 2,137 pieces of content were removed for violations of the Abusive Behaviour policy.

New Zealand-specific approximate data for July to December 2021:

- 2,709 accounts were reported for violations of the Suicide and Self-harm policy
- 881 accounts were actioned for violations of the Suicide and Self-harm policy
- 39 accounts were suspended for violations of the Suicide and Self-harm policy
- 1,084 pieces of content were removed for violations of the Suicide and Self-harm policy.

Measure 8. Implement, maintain and raise awareness of product or service related policies and tools for users to report online bullying or harassment content.

- We've recently updated our reporting experience to take a "symptoms-first" approach, providing a welcoming, conversational space for people to describe what's happening with example Tweets to provide as much context and education throughout the process—which we believe will create a safer experience and build trust with reporters. Improving the ability to report is just one step of our work to protect our users and make Twitter safer.
- Based on user feedback, research, and an understanding that our existing reporting process wasn't making enough people feel safe or heard, Twitter overhauled the reporting process to make it easier for people to report harmful behaviour. The new approach, which began testing in December 2021 and became available to all users in June 2022, simplified the reporting process. The new process created a more empathetic experience for people reporting potential policy violations by asking people to describe what happened and by closing the loop to show which Twitter Rules applied to the report and why. This new approach also allows Twitter to collect more first-hand information so we can be more precise with addressing concerns at scale. After experimenting, we saw the number of actionable reports increased by 50% through the new reporting flow. This symptoms-first method, where Twitter first asks the person what's going on, has ultimately lifted the burden from the individual to be the one who has to interpret the violation at hand.

Measure 9. Support or maintain programs, initiatives or features that seek to educate and raise awareness on how to reduce or stop online bullying or harassment.

- Through Twitter's owned and operated accounts, blogs, Help Centre, and partnerships with key organisations in the Twitter Trust & Safety Council, Twitter runs ongoing educative and engagement campaigns to educate users and raise awareness on how to utilise Twitter's safety products and how Rules apply on the service. Some highlights include working with the Trust & Safety Council partners, a global group of independent expert organisations, as they advise internal teams to develop products, programs, and rules, and assist with socialising updates to the platform. Another example includes the account @TwitterSafety that Tweets information about the latest safety tools, resources, and updates from Twitter. Additionally, Twitter's Common Thread blog frequently shares information about the shared future of a healthier conversation, discusses broader ethical questions, and takes an honest look at what's happening on Twitter and across the internet.
- In Aotearoa New Zealand, Twitter has worked with Netsafe NZ on the development and annual updates of their "Staying safe online" booklet, which includes relevant information on Twitter's tools, policies and enforcement. Additionally, Netsafe NZ has been a part of

Twitter's [Trust & Safety Council] since its formation in 2016, and continues to provide important guidance and feedback for Twitter's service.

CASE STUDY 1: Using nudges to have healthier conversations

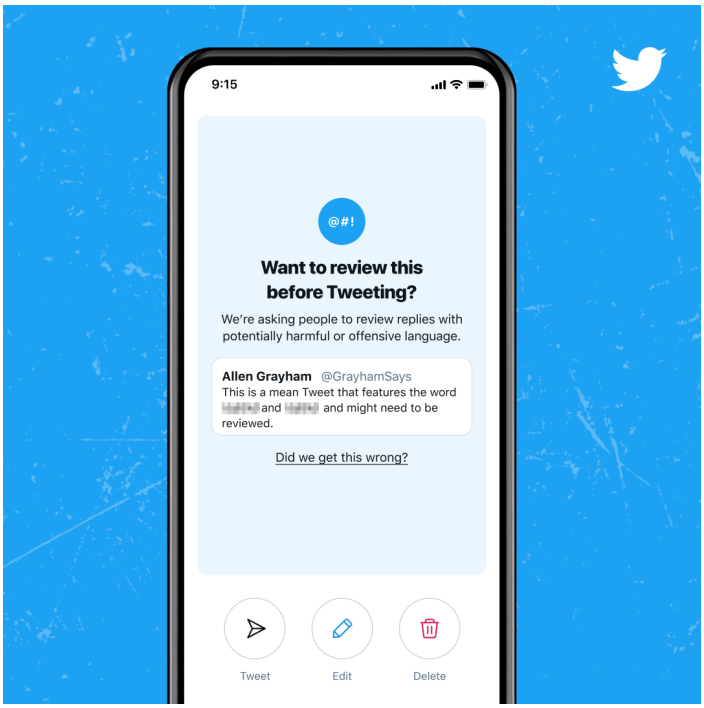

Image 1: The prompt that appears to encourage a user to reconsider their Tweet

Twitter has been experimenting with nudges to help create a healthier online environment that reflects what we expect when we interact with one another in real life. Not only has the team seen positive results in the impact these nudges can have on helping facilitate healthier conversations in English but they are now being expanded to languages like Portuguese, Spanish, and Turkish.

A new feature that prompts people to reconsider Tweet replies containing harmful language is seeing promising results, with people changing or deleting their replies over 30% of the time when prompted for English users in the U.S. and around 47% of the time for Portuguese users in Brazil. Twitter designed the intervention to try to bring more awareness to the moments where people who may get caught in what they call a "hot state" — when they're about to use words they may later regret.

The company began testing the latest version of the offensive Reply prompt in February 2021. For six weeks, the team studied how well these interventions worked, compared to a control group that received no prompts. The team behind the experiment — Matthew Katsaros, who is a part-time research advisor at Twitter and the director of the Social Media Governance Initiative at Yale's Justice Collaboratory, and Twitter data scientists Kathy Yang and Lauren Fratamico — compiled their findings in an academic paper. The findings suggest that nudges can encourage less offensive speech online without hindering participation in online conversations. People who

were prompted to reconsider their Replies cancelled them 9% of the time and revised them 22% of the time (37% of these were to a less offensive alternative). Overall, those who were prompted posted 6% fewer offensive Tweets. The team also observed a decrease in both the number of future offensive Tweets written by users who got the nudge, and the number of offensive replies they themselves received.

Outcome 3: Provide safeguards to reduce the risk of harm arising from online **hate speech**

Measure 10. Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of hate speech.

- Under Twitter's Hateful conduct policy, we do not allow users to promote violence against, directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories. We will review and take action against reports of accounts targeting an individual or group of people with any of the following behaviour, whether within Tweets or Direct Messages. This includes:
    - Violent threats.
    - Wishing, hoping, or calling for serious harm on a person or group of people.
    - References to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims.
    - Incitement against protected categories.
    - Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone.
    - Hateful imagery.
    - Using insults or profanity with the purpose of harassing or intimidating others.
    - Encouraging or calling for others to harass an individual or group of people.
    - Unwanted sexual advances or graphic objectification.

Measure 11. Implement and maintain products and tools that seek to prohibit or reduce the prevalence of hate speech.

- 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 hateful conduct as defined in this policy, or who have shared violent threats.

- We expanded our Hateful conduct policy in December 2021 to prohibit dehumanising speech on the basis of gender, gender identity and sexual orientation. In our latest Transparency Report from July to December 2021, during this period Twitter globally suspended 104,565 accounts under this policy.

  New Zealand-specific approximate data for July to December 2021:
  - 7,641 accounts were reported for violations of the Hateful conduct policy
  - 1,327 accounts were actioned for violations of the Hateful conduct policy
  - 158 accounts were suspended for violations of the Hateful conduct policy
  - 1,903 pieces of content were removed for violations of the Hateful conduct policy

Measure 12. Implement, maintain and raise awareness of product or service related policies and tools for users to report potential hate speech.

- With regards to our Hateful conduct policy, we are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalised. The policy makes clear that no one on Twitter may promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.
- Over the past two years, we conducted a multi-stage consultation process, to expand this policy and encompass the evolving nature of conversations online. We engaged multiple global stakeholders on the phased updates. Through this process we continued to expand on the policy and prohibit behaviour that targets individuals or groups with abuse based on their perceived membership in a protected category. The consultation sought a variety of perspectives and voices when considering how to balance harm reduction and freedom of expression, as well as how to avoid any unintended consequences of removing legitimate speech from marginalised groups.

Measure 13. Support or maintain programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of online hate speech.

- Twitter is a member of the Online Hate Observatory working on developing a better understanding of the mechanics behind online hate to build better answers in cooperation with NGOs, researchers, and relevant governments. Twitter is also actively involved in the development and consultation of the joint Australia-New Zealand government funded Organisation for Economic Co-operation and Development (OECD) Voluntary Transparency Reporting Protocols.
- At an EU level, we have worked closely with the Radicalisation Awareness Network (RAN) and have been part of the Civil Society Empowerment Program, supporting organisations across the EU in countering violent extremism.

Measure 14. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online hate speech.

- We believe at Twitter that wider efforts on promoting safe use of online services that are focused on bolstering the voices of NGOs and nonprofits would facilitate the desired consultation and cooperation in the private sector. Many of these nonprofit and non-governmental groups do critical work, and policy makers should continue to find ways to broaden support for these efforts and initiatives that promote best practices concerning

the safe use of services.

Outcome 4: Provide safeguards to reduce the risk of harm arising from online **incitement of violence**

Measure 15. Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of content that potentially incites violence.

- Twitter addresses the risk of harm arising from online incitement of violence through a range of policies, enforcement actions and product solutions.
- ***Glorification of violence***
    - Glorifying violent acts could inspire others to take part in similar acts of violence. Additionally, glorifying violent events where people were targeted on the basis of their protected characteristics (including: race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease) could incite or lead to further violence motivated by hatred and intolerance. For these reasons, we have a policy against content that glorifies acts of violence in a way that may inspire others to replicate those violent acts and cause real offline harm, or events where members of a protected group were the primary targets or victims.
    - The consequences for violating our glorification of violence policy depends on the severity of the violation and the account's previous history of violations. The first time a user violates this policy, we will require them to remove the content. We will also temporarily lock them out of their account before they can Tweet again. If a user continues to violate this policy after receiving a warning, their account will be permanently suspended.
- ***Perpetrators of violent attacks policy***
    - We want Twitter to be a place where people can find reliable information and express themselves freely and safely without feeling burdened by unhealthy content. In the aftermath of terrorist, violent extremist and mass violent attacks, we know many want to express compassion for victims, condemn the attacks and/or the perpetrators, and discuss how these incidents impact people and their communities. Some might also wish to share manifestos or other similar content produced by the attack's apparent perpetrator or an accomplice, either to express outrage or condemnation of the perpetrator's possible motives. Under our perpetrators of violent attacks policy, we will remove any accounts maintained by individual perpetrators of terrorist, violent extremist, or mass violent attacks, and may also remove Tweets disseminating manifestos or other content produced by perpetrators. The consequences for violating our violent events policy depends on the severity of the violation. Accounts maintained by perpetrators of terrorist, violent extremist, or mass violent attacks will be permanently suspended.
    - Violent attacks claimed by a violent organisation or by a member of such organisations are covered under our violent organisations policy. We do not require that a person have been confirmed as members of terrorist organisations or other violent groups, nor that they have any official affiliation with any group, organisation, or ideology, for us to enforce on content under this aspect of our policies.
- ***Violent threats policy***

- Healthy conversation is only possible when people feel safe from abuse and don't resort to using violent language. For this reason, we have a policy against threatening violence on Twitter. We define violent threats as statements of an intent to kill or inflict serious physical harm on a specific person or group of people.
- We will immediately and permanently suspend any account found to be posting violent threats. In rare cases, we may not suspend an account immediately. For example, if the reported content is a form of hyperbolic speech. In such cases, we may require the user to remove this content. We may also temporarily lock a user out of their account before they can Tweet again. If a user continues to violate this policy after receiving a warning, their account will be permanently suspended.

Measure 16. Implement and maintain products and tools that seek to prohibit or reduce the prevalence of content that potentially incites violence.

- Twitter takes robust measures in the detection and removal of terrorist or violent extremist content on the service. In our latest Transparency Report from July to December 2021, globally Twitter suspended 41,386 accounts, and we took action on 70,229 unique pieces of content for sharing content that threatens violence against an individual, a group of people, or the glorification of violence. Twitter globally suspended 33,693 unique accounts for violations of the terrorism and violent extremism policy during this reporting period. Of those accounts, 92% were proactively identified and actioned. Our current methods of surfacing potentially violating content for review include leveraging the shared industry hash database supported by the Global Internet Forum to Counter Terrorism (GIFCT).

  New Zealand-specific approximate data for July to December 2021:
  - 2,898 accounts were reported for violations related to various categories of violence which includes, among other policies, violations of the Terrorism and Violent Extremism policy[6]
  - 38 accounts were suspended for violations of the Terrorism and Violent Extremism policy.[7] In addition, 107 accounts were suspended and 169 pieces of content were removed for other categories of violence that are not related to terrorism and violent extremism (e.g. Violent Threats policy).

Measure 17. Implement, maintain and raise awareness of product or service related policies and tools for users to report content that potentially incites violence.

- In 2022, Twitter rolled out an overhauled reporting process for organic reports, which lifts the burden from the individual to be the one who has to interpret the violation at hand. Instead it asks them what happened. This method is called symptoms-first, where Twitter first asks the person what's going on, where users can better describe the context of the reported Tweets. Users will be able to select who the report is for, with the following options:
  - Myself
  - Someone else
  - A specific group, or
  - Everyone on Twitter.

[6] "**Accounts reported**" reflects the total number of unique accounts that users reported for potentially violating the Twitter Rules.
[7] "**Accounts actioned**" reflects the number of unique accounts that were suspended or had some content removed for violating the Twitter Rules. "**Accounts suspended**" reflects the number of unique accounts that were suspended for violating the Twitter Rules.

- Users are then asked to provide more context to the report, with the option to select additional Tweets from the account they are reporting.
- Based on user feedback, research, and an understanding that our existing reporting process wasn't making enough people feel safe or heard, Twitter overhauled the reporting process to make it easier for people to report harmful behaviour. The new approach, which began testing in December 2021 and became available to all users in June 2022, simplified the reporting process. The new process created a more empathetic experience for people reporting potential policy violations by asking people to describe what happened and by closing the loop to show which Twitter Rules applied to the report and why. This new approach also allows Twitter to collect more first-hand information so we can be more precise with addressing concerns at scale. After experimenting, we saw the number of actionable reports increased by 50% through the new reporting flow. This symptoms-first method, where Twitter first asks the person what's going on, has ultimately lifted the burden from the individual to be the one who has to interpret the violation at hand.
- Twitter encourages those that witness conduct on the platform that may violate the hateful conduct policy to report it. Twitter leverages a combination of keyword as well as behaviour-related signal scanning to identify accounts similar to those previously detected to be sharing violating content. We also leverage proprietary technology to detect variations of image material, as well as utilise ban-evasion detection methods which include automated suspension of accounts using certain signals associated with previously identified violators.

Measure 18. Support or maintain programs and initiatives that seek to educate users on how to reduce or stop the spread of online content that incites violence.

- Twitter is part of a number of existing multi-stakeholder organisations, forums, and initiatives in the technology space, and we have a long history of involvement in a number of international initiatives to combat serious online threats. For example, we are members and signatories of many coalitions and organisations, including but not limited to, the Global Internet Forum to Counter Terrorism (GIFCT), the Aqaba Process, the Christchurch Call to Action, and the Australian Taskforce to Combat Terrorism and Extreme Violent Material Online. We are also invested in the Global Research Network on Terrorism and Technology (GRNTT) to develop research and policy recommendations designed to prevent terrorist exploitation of technology. As the GIFCT Operating Board chair for 2021, Twitter endeavoured to support the work of the GIFCT as it deploys annual programs, training, and expands on its transparency efforts as an independent nonprofit.

Measure 19. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online content that incites violence.

- In furtherance of our work as signatories to the Christchurch Call to Action and through our Data for Good program, Twitter partnered with academics at the University of Otago's National Centre for Peace and Conflict Studies to analyse the ways in which Twitter and social media more broadly was used for both positive and negative purposes. Preliminary research just a week after the Christchurch massacre analysed data generated from tens of thousands of public Tweets anchored to the violence, highlighting a local and global outpouring of support for victims, solidarity with the citizens of New Zealand, the affirmation of democratic ideals, pushback against terrorism, and unequivocal condemnation of the perpetrator. The research helped demonstrate key dynamics of different communication patterns prior to the Christchurch attacks, but also enabled us to better understand ways in

which differences get turned into polarised divisions.
- Additionally at the UN General Assembly in September 2022 Twitter alongside Microsoft, the US Government, and NZ Government announced its support of the 'Christchurch Call initiative on Algorithmic Outcomes'. Twitter has provided both financial support and datasets for OpenMined, the NGO that is driving the first phase of the project to develop specific privacy enhancing technology (PET) to enable external researchers to investigate the role algorithms play, in a privacy-protective environment. Twitter's leadership in this sphere remains focused on enabling independent research and creating a better understanding of this issue that is increasingly at the heart of policy conversations around the world.

Outcome 5: Provide safeguards to reduce the risk of harm arising from online **violent or graphic content**

Measure 20. Implement, enforce and/or maintain policies and processes that seek to prohibit and/or reduce the spread of violent or graphic content online.

- People use Twitter to show what's happening in the world, often sharing images and videos as part of the conversation. Sometimes, this media can depict sensitive topics, including violent and adult content. We recognise that some people may not want to be exposed to sensitive content, which is why we balance allowing people to share this type of media with helping people who want to avoid it to do so.
- Violent or graphic content fall under Twitter's Sensitive Media policy. Our aim is to limit exposure to sensitive images and videos and to prevent the sharing of potentially disturbing types of sensitive media. For this reason, we differentiate our enforcement approach depending on the type of media that has been shared and where it has been shared. Adult content that was created or shared without the consent of those depicted is reviewed under our non-consensual nudity policy.

Measure 21. Implement and maintain products and tools that seek to and/or reduce the spread of violent or graphic content.

- ***Graphic violence, adult content, and hateful imagery***
  - A user can't target people with unsolicited images or videos that contain graphic violence, adult content, or hateful imagery; and
  - A user can't include graphic violence, adult content, or hateful imagery within live video, profile, header, List banner images, or Community banner images.
- ***Violent sexual conduct and gratuitous gore***
  - Twitter prohibits violent sexual conduct to prevent the normalisation of sexual assault and non-consensual violence associated with sexual acts. We prohibit gratuitous gore content because research has shown that repeated exposure to violent content online may negatively impact an individual's wellbeing. For these reasons, a user can't share images or videos that depict violent sexual conduct or gratuitous gore on Twitter. Note that very limited exceptions may be made for gory media associated with newsworthy events.
  - Enforcement action under the Sensitive Media policy depends on the type of media the user has shared, and where they have shared it. Additional information on enforcement actions can be found on the Twitter Help Centre.

- ***Non-consensual nudity***
    - Twitter prohibits people from posting or sharing intimate photos or videos of someone that were produced or distributed without their consent. Sharing explicit sexual images or videos of someone online without their consent is a severe violation of their privacy and the Twitter Rules. Sometimes referred to as revenge porn, this content poses serious safety and security risks for people affected and can lead to physical, emotional, and financial hardship.
    - Under this policy, you can't post or share explicit images or videos that were taken, appear to have been taken or that were shared without the consent of the people involved. Examples of the types of content that violate this policy include, but are not limited to:
        - hidden camera content featuring nudity, partial nudity, and/or sexual acts;
        - creepshots or upskirts - images or videos taken of people's buttocks, up an individual's skirt/dress or other clothes that allows people to see the person's genitals, buttocks, or breasts;
        - images or videos that superimpose or otherwise digitally manipulate an individual's face onto another person's nude body;
        - images or videos that are taken in an intimate setting and not intended for public distribution; and
        - offering a bounty or financial reward in exchange for intimate images or videos.
    - We recognise that it can be difficult for those impacted to report this type of content for review. To reduce the burden on those affected, anyone can report the following types of content: creepshots or upskirts; content where a bounty or financial reward is offered in exchange for non-consensual nudity media; and intimate images or videos that are accompanied by text that wishes/hopes for harm to come to those depicted or information that could be used to contact those depicted. For other types of content, because Twitter allows some types of adult content, we need to evaluate context to assess if reported content has been created or shared without the consent of those involved, and we may need to hear directly from the individual(s) featured (or an authorised representative, such as a lawyer) to ensure that we have sufficient context before taking any enforcement action.
- In our latest Transparency Report from July to December 2021, Twitter globally removed a total of 1.1 million unique pieces of content under our Sensitive Media policy during this period.

  New Zealand-specific approximate data for July to December 2021:

    - 4,021 accounts were reported for violations of the Sensitive media policy
    - 1,204 accounts were actioned for violations of the Sensitive media policy
    - 134 accounts were suspended for violations of Sensitive media policy
    - 1,213 pieces of content were removed for violations of the Sensitive media policy.

Measure 22. Implement, maintain and raise awareness of product or service related policies and tools for users to report potential violent and graphic content.

- Twitter maintains detailed updates about new features, helpful tips, and changes to our rules and guidelines on our global Help Centre available at help.twitter.com. Information about our range of enforcement options and our approach to policy development and enforcement is also provided for any person that visits our Help Centre.

Outcome 6: Provide safeguards to reduce the risk of harm arising from online **misinformation**

Measure 23. Implement, enforce and/or maintain policies, processes and/or products that seek to reduce the spread of online misinformation.

- Twitter is committed to keeping people informed about what's happening in the world. As such we care deeply about the issues of misinformation as well as disinformation, and their potentially harmful effects on the civic and political discourse that is core to our mission. Twitter addresses misinformation and disinformation through a range of policies, enforcement actions, and product solutions.
- 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.
- 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*.
- Twitter's main policies that focus on misleading content include: (1) COVID-19 misleading information policy; (2) Civic integrity policy; (3) Crisis misinformation policy; and (4) Synthetic and manipulated media policy:
    - ***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.
    - 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, 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

- As of March 2021, we incorporated a five-strike system meant to address repeated violations of the COVID-19 misleading information policy. After the fifth strike, the user is eligible for suspension under the policy. Since the launch of the strike system we invested in and increased our proactive detection efforts to surface and mitigate the harm related to COVID-19 misinformation. In our latest Twitter Transparency Report from July to December 2021, globally we suspended 1,376 accounts for violations of the COVID-19 misinformation policy.

New Zealand-specific approximate data for July to December 2021:

- 118 accounts were actioned for violations of the COVID-19 misleading information policy
- 12 accounts were suspended for violations of the COVID-19 misleading information policy
- 144 pieces of content were removed for violations of the COVID-19 misleading information policy.

- ***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 the user to remove this content. We will also temporarily lock the user out of their account before they can Tweet again. Tweet deletions accrue 2 strikes.
        - **Profile modifications:** If the user violates this policy within their profile information (e.g. your bio), we will require them to remove this content. We will also temporarily lock the user out of their account before they can Tweet again. If they violate this policy again after their first warning, their 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: (1) Apply a label and/or warning message to the content where it appears in the Twitter product; (2) Show a warning to people before they share or like the content; (3) Turn off people's ability to reply, Retweet, or like the Tweet; or (4) Reduce the visibility of the content on Twitter and/or prevent it from being recommended.

- ***Crisis misinformation policy***
    - We will take action on accounts that use Twitter's services to share false or misleading information that could bring harm to crisis-affected populations. A crisis is any situation in which there is a widespread threat to life, physical safety, health, or basic subsistence that is beyond the coping capacity of individuals and the communities in which they reside. Currently, the scope of this policy includes international armed conflict. In these contexts, we're focusing on misleading information with the capacity to:
        - Serve as a pretext for further aggression by armed actors, belligerents, or combatants,
        - Trigger forced or anticipatory displacement of vulnerable populations, or lead to increased humanitarian needs,
        - Negatively impact the ability of humanitarian protection, human rights, or relief organisations to provide assistance or access affected populations,
        - Incite the targeting or surveillance of groups that can be identified based on their political, religious, ethnic or ideological affiliation or membership, or organisations and actors protected by international humanitarian law;

        - Disrupt potential ceasefire agreements, peacekeeping operations, or diplomatic solutions to conflict or insecurity, among other matters.
- ***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. In order for content with **misleading media** (including images, videos, audios, gifs, and URLs hosting relevant content) to be labelled or removed under this policy, it must:
        - Include media that is significantly and deceptively altered, manipulated, or fabricated, or
        - Include media that is shared in a deceptive manner or with false context, and
        - Include media likely to result in widespread confusion on public issues, impact public safety, or cause serious harm.

Measure 24. Implement, enforce and/or maintain policies and processes that seek to penalise users who repeatedly post or share misinformation that violates related policies.

- 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 via our Help Centre.
    - **Launching prebunks**: During important events, 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.

Measure 25. Support or maintain media literacy programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of misinformation.

- Globally, our partners UNESCO, Common Sense Media, the National Association for Media Literacy, the Family Online Safety Institute, and Connect Safely (amongst others) have helped us develop materials and conduct workshops to help people who use Twitter to learn how to better process online information and discern between sources of news.
- In August 2020, we worked on a campaign around conspiracy theories / disinformation, partnering with UNESCO, the European Commission and the World Jewish Congress. The aim of the campaign was to inform people of how to spot conspiracy theories online and how to make smarter choices when you do see this content circulating.
- This also includes the Teaching and Learning with Twitter handbook, launched in partnership with UNESCO. We focus on elements like verification of sources, critical thinking, active citizenship online, and the breaking down of digital divides.
- We also partner with journalistic NGOs for training and media literacy initiatives, including Reporters Without Borders, the Committee to Protect Journalists, and the Reporters Committee for Freedom of the Press.

Measure 26. Support or maintain programs and/or initiatives that seek to support civil society, fact-checking bodies and/or other relevant organisations working to combat misinformation.

- Twitter has worked with The Associated Press (AP) and Reuters to expand our efforts to identify and elevate credible information on Twitter. We are committed to making sure that when people come to Twitter to see what's happening, they are able to easily find reliable information. Twitter will be able to expand the scale and increase the speed of our efforts to provide timely, authoritative context across the wide range of global topics and conversations that happen on Twitter every day. The scope of this program is independent of the work that Twitter's Trust & Safety teams do to determine whether Tweets are in violation of the Twitter Rules. AP and Reuters will not be involved in enforcement decisions.
- Twitter's Curation team helps give people context to make informed decisions about what they see on Twitter. When large or rapidly growing conversations happen on Twitter that may be noteworthy, controversial, sensitive, or may contain potentially misleading information, Twitter's Curation team sources and elevates relevant context from reliable sources. People currently see this added context and reliable information in the following places on Twitter:
    - **Trends.** To help explain a top Trend, we often attach additional context to the Trend in the form of a Moment, a single Tweet, or via a written description.
    - **Explore Tab.** Explore catches you up on a range of subjects, curated just for you. This is where Twitter shows you what's happening right now. It includes Moments created by Twitter and third parties like news organisations, as well as Trends.
    - **Search.** When people search for a hashtag or a phrase on Twitter, certain keywords determined by Twitter will automatically show content at the top of the results from trusted resources or Moments that debunk misinformation.
    - **Prompts.** During the highest visibility events, such as elections or public health emergencies, Twitter will show prompts in the Explore tab or the Home Timeline that link to a public service announcement (PSA) Moment. These can include information from trusted sources on topics like how to vote safely in a pandemic or trustworthy information about getting vaccinated.

    - **Labels.** Sometimes a Tweet violates our Synthetic and Manipulated Media, COVID-19 or Civic Integrity misinformation rules but may remain visible on Twitter. In those instances, a label may be added to the Tweet that links to a Moment with informative context on the topic or to the Twitter Rules.

Measure 27. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from misinformation.

- Additionally, at Twitter we are 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.
- ***Twitter misleading information prompts:*** 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.
- ***Misleading information reporting beta***: In the face of misleading information, Twitter is aiming 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.
    - Currently, Twitter is testing out a misleading information reporting flow that 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 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."

CASE STUDY 2: 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 behaviour 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.

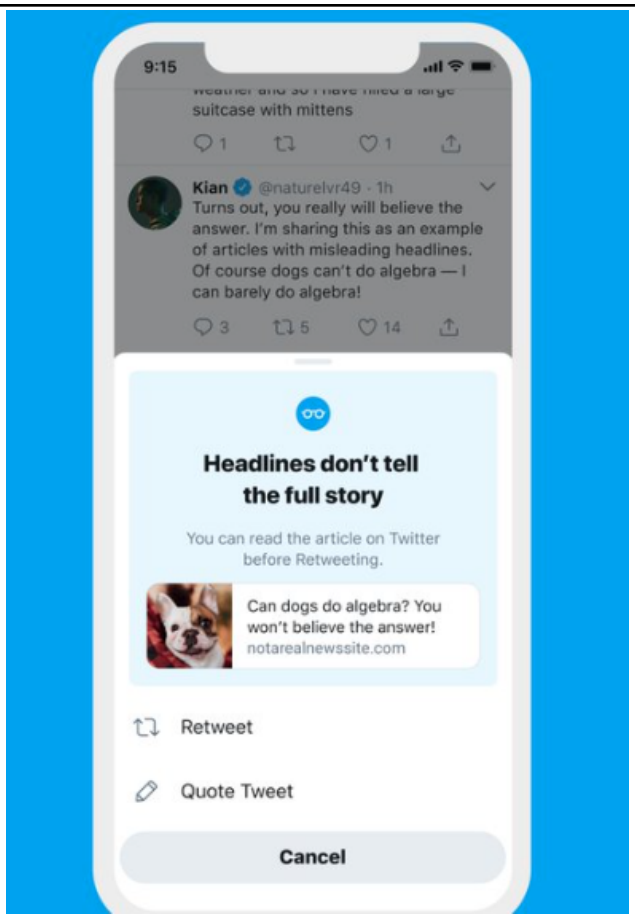

Image 2: The prompt that appears when a user Retweets an article they haven't opened on Twitter.

CASE STUDY 3: Birdwatch

Twitter is currently testing a feature called Birdwatch in the United States. Birdwatch aims to create a better informed world by empowering people on Twitter to collaboratively add notes to potentially misleading Tweets. Pilot 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.

The program began testing in 2021 and is regularly being updated and improved thanks to analysis from our research team and feedback from our academic advisory board and Birdwatch contributors. Now, we're rolling out a new onboarding process and expanding the visibility of notes to increase the positive impact of Birdwatch and enable healthier Twitter conversations.

Our research indicates that Birdwatch is an effective way to keep people better informed on Twitter. According to the results of three surveys, people who see a Birdwatch note are, on average, 20-40% less likely to agree with the substance of a potentially misleading Tweet than someone who sees the Tweet alone. By analysing our internal data, we also estimate that people on Twitter who see notes are, on average, 15-35% less likely to Like or Retweet a Tweet than someone who sees the Tweet alone.

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."

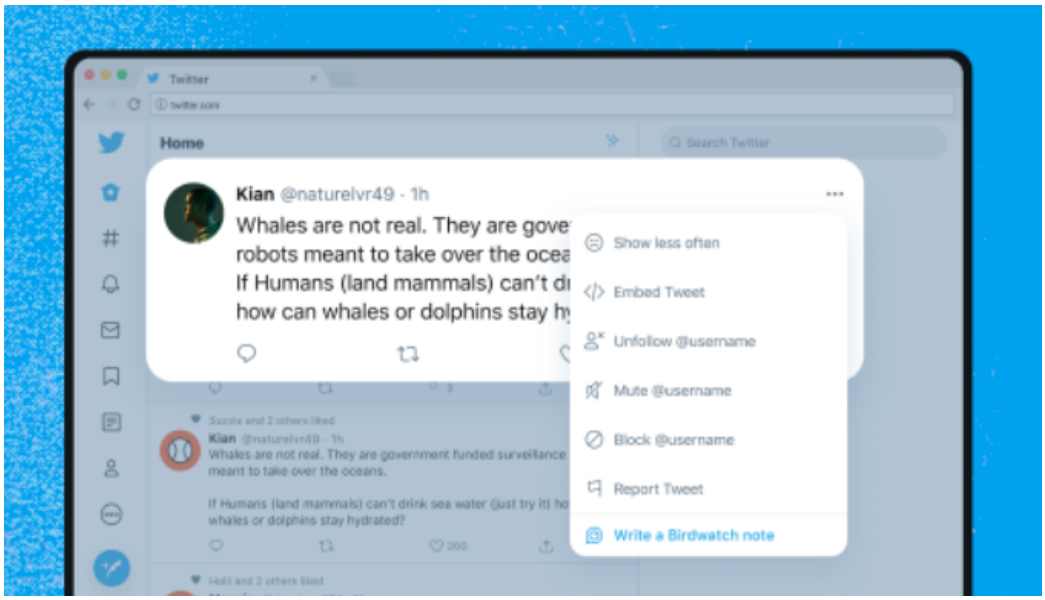

Image 3: Depicts the drop-down menu available for Contributors to write a Birdwatch Note.

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

Outcome 7: Provide safeguards to reduce the risk of harm arising from online **disinformation**

Measure 28. Implement, enforce and/or maintain policies, processes and/or products that seek to suspend, remove, disable, or penalise the use of fake accounts that are misleading, deceptive and/or may cause harm.

- 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.
- **Misleading and deceptive identities policy**
    - Deceptive identities may feature the likeness of another person or organisation in a manner that confuses others about the account's affiliation. Fake identities, which may use stolen or computer-generated photos and fabricated names to pose as a person or organisation that doesn't exist, are also considered deceptive when they engage in disruptive or manipulative behaviour.

- One of the main elements of an identity on Twitter is an account's profile, which includes a username (@handle), account name, profile image, and bio. An account's identity is deceptive under this policy if it uses false profile information to represent itself as a person or entity that is not associated with the account owner, such that it may mislead others who use Twitter.
- The consequences for violating the policy depend on the severity and type of violation, as well as an account's history of previous violations. The actions we take may include the following:
  - Profile modifications: If an account is potentially confusing in terms of its affiliation, we may require the user to edit the content on their profile. If the user violates this policy again after their first warning, their account will be permanently suspended.
  - Temporary account suspension: If we believe the user may be in violation of this policy, we may require the user to provide government issued identification (such as a driver's licence or passport) in order to reinstate their account.
  - Permanent suspension: If the user is engaged in impersonation or is using a misleading or deceptive fake identity, we may permanently suspend their account.
- Anyone can report a suspected misleading or deceptive identity directly from the account's profile. In cases where an account is suspected of misusing a specific individual or entity's identity, we may need more information to determine whether the account is run or authorised by the entity portrayed in the profile. To ensure we have enough context, we may need a report from the portrayed party or their authorised representative in order to take action.
- In our latest Transparency Report from July to December 2021, globally we actioned 181,644 accounts and suspended 169,396 accounts for violations of the impersonation policy.

New Zealand-specific approximate data for July to December 2021:

- 1,348 accounts were reported for violations of the the impersonation policy
- 180 accounts were actioned for violations of the the impersonation policy
- 154 accounts were suspended for violations of the impersonation policy
- 29 pieces of content were removed for violations of impersonation policy.

Measure 29. Implement, enforce and/or maintain policies, processes and/or products that seek to remove accounts, (including profiles, pages, handles, channels, etc) that repeatedly spread disinformation.

- **Platform manipulation and spam policy**
  - 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.
  - A large part of this work involves fighting platform manipulation (including spam and malicious automation) strategically and at scale. Users may not 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.

- Platform manipulation can take many forms and our rules are intended to address a wide range of prohibited behaviour, 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 behaviour 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 may result in harsher enforcements, including 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 or provide distinct purposes for each account. The remaining accounts will remain 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 behaviour 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.
    - 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 behaviour.

Measure 30. Implement, enforce and/or maintain policies, processes and/or products that seek to provide information on public accounts (including profiles, pages, handles, channels, etc) that empower users to make informed decisions (e.g. date a public profile was created, date of changes to primary account information, number of followers).

- **Government and State-affiliated media account labels** (State media labels)
    - Twitter provides an unmatched way to connect with, and directly speak to public officials and representatives. This direct line of communication with leaders and officials has helped to democratise political discourse and increase transparency and accountability. State-affiliated account labels provide additional context about accounts that are controlled by certain official representatives of governments, state-affiliated media entities and individuals 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.
    - For Government account labels, 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, defence officials, and key diplomatic leaders. Where accounts do not play a role as a geopolitical or official Government communication channel, we will not label the account. As of October 2022, 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 journalists, associated with state-affiliated media entities.
    - State-affiliated media are 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 prominent staff may be labelled. In the case of state-affiliated media entities, Twitter will not recommend or amplify accounts or their Tweets with these labels to people. In limited circumstances where there is heightened risk for harm, including situations where governments block access to information on the internet in the context of an armed conflict, Twitter will also not recommend or amplify certain government accounts or their Tweets with these labels to people. We will also add labels to Tweets that share links to state-affiliated media websites and will not recommend or amplify these Tweets to people. 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. As part of the development of this process, we consulted with a number of expert groups, including members of the Digital and Human Rights Advisory group in Twitter's Trust & Safety Council.

- **Automated account labels**
    - Automated accounts (often referred to as "bots") perform programmed actions through the Twitter API. Examples of automated accounts you might see on Twitter include bots that help you find vaccine appointments and disaster early warning systems. When these accounts let you know they're automated, you get a better understanding of their purpose when you're interacting with them. Automated labels provide transparency by helping you identify if an account is a bot or not. Automated accounts are created and managed by other people on Twitter. Our Automation rules require these accounts to display labels and remain connected to a human-run account. When someone who manages an automated account sets that account to display the identifying automatic account label, it also connects their human-run account to let everyone know who's managing it. Once an account accepts the invitation to our test group, an automated account label will appear on their account profile under their profile name and handle. The label may also appear on their Tweets.

Measure 31. Implement, enforce and/or maintain policies, processes and/or products that seek to provide transparency on paid political content (e.g. advertising or sponsored content) and give users more context and information (e.g. paid political or electoral ad labels or who paid for the ad).

- Measure 31 is not applicable to Twitter.

Measure 32. Implement, enforce and/or maintain policies, processes and/or products that seek to disrupt advertising and/or reduce economic incentives for users who profit from disinformation.

- ***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 are outlined below.
    - *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. This includes misleading advertisements on Twitter that contradict the scientific consensus on climate change.

- *Unacceptable business practices policy:* Twitter prohibits the promotion of unacceptable business practices globally. Examples of unacceptable business practices include:
  - Potentially deceptive, misleading, or harmful business propositions.
  - Making misleading, false, or unsubstantiated claims during the promotion of a product or service.
  - Promoting misleading information or omitting vital information on pricing, payment terms, or expenses the customer will incur.
  - Promoting offers or deals that are not available or easily located on the landing page.
  - Use of misleading 'before and after' text or images.
  - Use of content which could be reasonably considered to 'body-shame' the customer.
  - Encouragement, glamorisation, or promotion of unhealthy or unsafe eating behaviours or eating disorders.
- *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 advertising:* 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.

Measure 33. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from disinformation.

- Twitter engages and collaborates with a wide range of stakeholders to mitigate detected threats and also empower customers with credible context on important issues. For example, we collaborate with the Electoral Council of Australia and New Zealand (ECANZ) to protect and promote electoral integrity online. This collaboration led to a Statement of Intent being established between the Australian Electoral Commission (AEC), Twitter and other online media platforms ahead of the 2022 Australian Federal Election. The Statement of Intent established a framework for detailed operational arrangements allowing the AEC to refer harmful electoral content to online platforms for consideration and removal, where content was in breach of relevant legislation or the platform's own policies.

- As part of Twitter's commitment to transparency, we work with a global group of expert academics, members of civil societies and NGOs, and journalists through the Twitter Moderation Research Consortium (TMRC). The goal of the TMRC is to prioritise transparency by sharing data on issues that are relevant to those who are studying content moderation. These combined efforts have given researchers access to 52 datasets spanning nine terabytes of media and more than 220 million Tweets, all with one goal in mind: empowering an unprecedented level of empirical research into state-backed attacks on the integrity of the conversation on Twitter. For example, earlier this year we shared data from about 15 information operations as a pilot for the TMRC. This data has already enabled critical, independent research by TMRC's partners at the Stanford Internet Observatory, the Cazadores de Fake News, and the Australian Strategic Policy Institute.

Outcome 8. Users are empowered to **make informed decisions** about the content they see on the platform

Measure 34. Implement, enforce and/or maintain policies, processes, products and/or programs that helps users make more informed decisions on the content they see

- We are committed to increasing the collective health on the platform, and this includes promoting content and behaviour that addresses people's needs, or that users will find important or useful. This also includes supporting user choice in how you moderate content you would like to see, and how you'd like to cultivate your own experience on the platform. We are committed to giving users control of what they see and who they interact with on our platform
- ***Sensitive Media Warnings***
    - People use Twitter to show what's happening in the world, often sharing images and videos as part of the conversation. Sometimes, this media can depict sensitive topics, so we give users a way to add Sensitive Media warnings on photos and videos they Tweet out. Sensitive Tweet Warnings gives users the ability to mark their tweets as containing sensitive content and consumers the option of whether or not to view the tweet. Sensitive media includes photos or videos with nudity, violence, or any content that might be triggering for some but doesn't violate our Rules and Policies. In December 2021, we began testing sensitive tweet warnings, and in February 2022 the feature became available globally across all devices.

Measure 35. Implement, enforce and/or maintain policies, processes, products and/or programs that seek to promote accurate and credible information about highly significant issues of societal importance and of relevance to the digital platform's user community (e.g. public health, climate change, elections)

- Twitter seeks to amplify accurate and credible information during times of crisis, emergency, or civic events. Over the years, Twitter has become a critical communication tool for responding to natural disasters. Our teams have a longstanding commitment to working alongside global partners and developers to share important information, provide real-time updates, facilitate relief efforts, and much more.
- We also take steps to address misleading information that can surface during these crises. When people search for keywords on Twitter related to active emergency situations, we work

to ensure they're met with authoritative and credible information first. Search prompts are one important way we do this. For example, when heavy flooding occurred in Pakistan beginning in June 2022, Twitter launched a #ThereIsHelp prompt to redirect our Pakistani audience to an emergency helpline at the Pakistani Red Crescent Society. Similarly, during this year's Taal Volcano eruption in the Philippines and Typhoon Noru in Southeast Asia, search prompts were created to help quickly deliver credible information to people on Twitter impacted by these extreme weather events. During natural disasters, Twitter has tools such as Moments and Lists that can be used to create a centralised source of credible information.

Measure 36. Launch programs and/or initiatives that educate or raise awareness on disinformation, misinformation and other harms, such as via media/digital literacy campaigns

- Additional information on how Twitter promotes accurate and credible information about highly significant issues of societal importance can be found under Outcomes 6 and 7.

Outcome 9. Users are **empowered with control** over the content they see and/or their experiences and interactions online

Measure 37. Implement, enforce and/or maintain policies, processes, products and/or programs that seek to provide users with appropriate control over the content they see, the character of their feed and/or their community online.

- We aim to increase consumer trust by giving people proactive control over their conversations. People may feel safer when they can choose to only engage with audiences they trust and avoid abusive replies.
- Using conversation settings enables a user to engage in more meaningful conversations and avoid unwelcome replies. This feature can help users proactively manage a conversation without having to use Block and can prevent spam. While everyone can see the Tweet and conversation, depending on your setting, not everyone would be able to reply.
- **Before a Tweet, users can choose who can reply with three options:**
    - Everyone (standard Twitter, and the default setting)
    - Only people the user follows
    - Only people the user mentions
    - Tweets with the latter two settings will be labelled and the reply icon will be greyed out for people who can't reply. People who can't reply will still be able to view, Retweet, Retweet with Comment, share, and like these Tweets.
- **Other control settings available to users include:**
    - **Mute:** Mute is a feature that allows people to remove an account's Tweets from their timeline without unfollowing or blocking that account. Muted accounts will not know that you've muted them and you can unmute them at any time.
    - **Block:** Block is a feature that helps people control how they interact with other accounts on Twitter. This feature helps people in restricting specific accounts from contacting them, seeing their Tweets, and following them.
    - **Remove followers**: This feature allows people to limit interactions with someone without having to block them. It is intended for everyone but can be specifically helpful for:
        - People with public accounts who don't receive "accept" or "reject" follow requests.

        - People who are moving from Public to Protected Tweets.
        - People who want to curtail interactions with others without the severity of blocking.
    - **Mention controls:** Mention Controls gives people the ability to remove themselves from a Tweet, thread, or conversation. Unmentioning lets people who try to @mention others see they've left the conversation and can't be tagged, changing the incentives for how people abuse @mentions.
    - In December 2021 we launched a Digital Safety Playbook in the Twitter for Good section of about.twitter.com. The Digital Safety Playbook is a resource that includes available safety tools to help users feel safer, be in control, and manage their digital footprint.

Measure 38. Launch and maintain products that provide users with controls over the appropriateness of the ads they see.

- We are committed to offering Twitter users meaningful privacy choices when it comes to advertising, and outline how users can turn off the interest-based ads feature. Twitter users also have the option to "dismiss" Promoted Ads and/or accounts that you're not interested in seeing.

CASE STUDY 4: Twitter Circle

Engaging in healthy conversations may sometimes warrant the need to select who views your Tweets. Twitter Circle helps reduce harassment and abuse by giving people more choice over who gets to see their content. We saw promising findings from the experiment phase that Twitter Circle Tweets received far fewer safety actions like blocks and mutes, and Twitter Circle replies were less abusive than non-Twitter Circle replies.

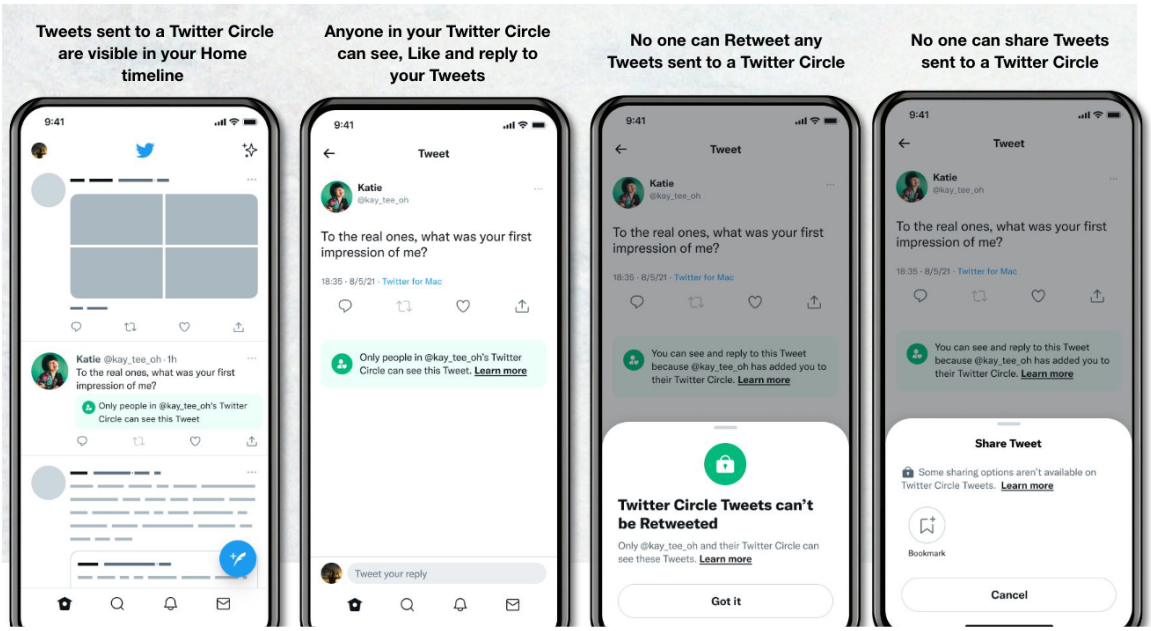

Image 4: Depicts what users will see when Tweeting to their Twitter Circle.

- Twitter Circle allows people to select who can see their Tweets on a Tweet-by-Tweet basis. With Twitter Circle, people can choose up to 150 people to have access to view the Tweets they publish to their Timeline when they select to Tweet to their Circle.
- This provides the user with the choice of who can see and engage with their content on their own terms without having to switch to an alt account or have a private account.
- We first started experimenting with the product in May 2022 with a select group of users globally on iOS, Android and Web. The response was overwhelmingly positive. The feature was then made available to everyone on iOS, Android, and Twitter.com globally in August 2022.

Outcome 10. Transparency of policies, systems, processes and programs that aim to reduce the risk of online harms

Measure 39. Publish and make accessible for users Signatories' safety and harms-related policies and terms of service.

- The Twitter Service is intended to allow users to create and share ideas and information instantly. It is intended to facilitate communication and free expression, which requires representation of a diverse range of perspectives.
- These are important goals for the company, which underpin our provision of the Twitter service to hundreds of millions of daily active users worldwide, including in New Zealand. But there are clear limits imposed by Twitter as to what users can do on the Service via the Twitter User Agreement, which every user agrees to when they create an account.
- In New Zealand, the Twitter User Agreement comprises the Terms of Service, our Privacy Policy, the Twitter Rules, and all incorporated policies.
    - Our Terms of Service govern user's access to and use of our services, including our website, APIs, email notifications, applications, buttons, widgets, ads, commerce services, and our other covered services, and any information, text, links, graphics, photos, audio, videos, or other materials or arrangements of materials uploaded, downloaded or appearing on the Services (collectively referred to as "Content"). By using Twitter's Services, a user agrees to be bound by these Terms.
    - Our Privacy Policy describes how we handle the information you provide to us when you use our Services, including the collection, use, storage, and processing of this information.
    - Twitter's purpose is to serve the public conversation, and the Twitter Rules are to ensure all people can participate in the public conversation freely and safely. We want people to be able to freely express their opinions while ensuring they are protected from harm. If people don't feel their conversations are safe from abuse and harassment, we know they won't feel comfortable participating in the public conversation.
- Twitter provides its terms and policies that users must follow in order to continue to use our service.

Measure 40. Publish and make accessible information (such as via blog posts, press releases and/or media articles) on relevant policies, processes, and products that aim to reduce the spread and prevalence of harmful content online.

- Twitter provides regular updates on our products, policies, and health and safety features through a range of channels such as the Twitter blog and Twitter accounts including @Twitter, @TwitterSafety, @Policy, @TwitterGov, and @TwitterSupport.

Outcome 11. Publication of regular **transparency reports** on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics

Measure 41. Publish periodic transparency reports with KPIs/metrics showing actions taken based on policies, processes and products to reduce the spread or prevalence of harmful content (e.g. periodic transparency reports on removal of policy-violating content).

- Transparency is central to Twitter's mission of serving the public conversation through open and free exchange of ideas. We publish a biannual Twitter Transparency Report that shares a wide variety of information about metrics and processes to earn our users trust, to enable robust engagement with our systems, and to strengthen our systems.
- Through initiatives such as the Twitter Transparency Centre (discussed further under Outcome 11), we are committed to reaching beyond Twitter to integrate diverse perspectives that make our service better for everyone.
- 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.
- Twitter updates our policies as our rationales, approaches, or enforcement options evolve, we 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.
- In July 2022, Twitter released its 20th Transparency Report, which marks a decade of transparency reporting. Meaningful transparency helps people understand the rules of online services and hold governments accountable for their actions, and in turn, helps keep us accountable for principled content moderation and responsiveness to government demands. Since our original report in 2012, our transparency reporting has evolved into a more comprehensive Twitter Transparency Centre covering a broader array of our transparency efforts.
- Now we include sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, state-backed information operations, and more. We've also worked to make this complex subject matter increasingly interactive and intuitive over the years. This has led to a number of improvements, including the use of data visualisations, localising the reports into seven major languages, and continuing to iterate on examples and contextualising the data.

- Twitter is committed to data-driven transparency. 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. In 2022, we then launched the Twitter Moderation Research Consortium (TMRC). For more information on the TMRC see Outcome 7, Measure 33.

Measure 42. Submit to the Administrator an annual compliance report, as required in section 5.4, that set out the measures in place and progress made in relation to Signatories' commitments under the Code.

- As signatories to the NZ Code of Practice for Online Safety and Harms, we will share this report as required under the relevant section for review.

Outcome 12. Independent research to understand the impact of safety interventions and harmful content on society and/or research on new technologies to enhance safety or reduce harmful content online.

Measure 43. Support or participate, where appropriate, in programs and initiatives undertaken by researchers, civil society and other relevant organisations (such as fact-checking bodies). This may include broader regional or global research initiatives undertaken by the Signatory which may also benefit Aotearoa New Zealand.

Twitter API for Academic 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 Centre 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): See Outcome 7, Measure 33 for an overview of the TMRC.
    - 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 in 2021 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 NGOs 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 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.

Twitter’s Trust & Safety Council

- As we work to improve the health of the public conversation, we’re committed to reaching beyond Twitter’s virtual walls to integrate diverse perspectives that make our service better for everyone. That’s why we regularly collaborate with trusted partners, including on our Trust & Safety Council, to develop products and programs, and to improve the Twitter Rules.
- We know the best version of Twitter is the one that people who use it help build. In 2021, we engaged with the Trust and Safety Council on thirteen projects early in the development process. We distilled and put to use their feedback on ways we can offer a better and safer experience for people using Twitter. Their feedback directly informed our approach on several products, including:

- **Communities**: We incorporated feedback on the need to manage expectations on the role that moderators play by limiting the number of responsibilities and building tools to help them manage potential harassment.
- **Tips**: We incorporated feedback on the need to emphasise that the people on our service are responsible for transactions in a user-friendly way by asking people to agree to terms of service when enabling the feature.
- **Safety Mode**: To mitigate the effect on limiting counter-speech, a concern raised by the council particularly for people in positions of power, we decided to automatically time out interventions for seven days.
- **Conversation settings**: We started testing a notification that reminds people they can change who can reply to their Tweets to increase awareness and adoption as a direct recommendation from the council.
- **Parental resources**: Working with partners, we developed a Digital Safety Playbook to help parents learn about the tools available to help them feel safer, be in control, and manage digital footprints.
- **Education assets**: Working with UNESCO, we developed digital assets that teach how Twitter can be used in the classroom and help people get UNESCO's guidance on the best practices for media and information literacy.
- **UN Envoy on Youth**: In partnership with the UN, we developed an accessible booklet about digital safety and online protection for young people with a checklist produced in collaboration with the Office of the UN Secretary-General's Envoy on Youth.
- **Digital Safety Playbook:** In December 2021, we launched a consolidated Twitter safety guide for NGOs and others to reference and share with their community members - especially women journalists and youth. We developed a visually inspired, one-stop resource covering select Twitter tools that were designed to help people on Twitter feel safer, more in control, and empowered to manage their digital footprint.
- We also worked with a number of partners to build on our mental health resources and support. As we continue to grapple with the weight and broad reaching effects of an unprecedented public health crisis through the pandemic, it is our job to ensure that Twitter remains a safe space for anyone interested in mental health tips and resources, or opening up about their individual mental health concerns. For example, in 2020 for World Suicide Prevention Day, we worked with various mental health partners across the globe to raise awareness and encourage honest conversation around the emotional challenges experienced amid the unprecedented COVID-19 crisis. We've consistently and continuously expanded our work with NGOs focused on mental health. In particular, we've continued to engage suicide prevention organisations and counseling services to ensure that people on Twitter feel safe and have access to support when they need it most.

Measure 44. Support or convene at least one event per year to foster multi-stakeholder dialogue, particularly with the research community, regarding one of the key themes of online safety and harmful content, as outlined in section 4. This may include broader regional or global events undertaken by the Signatory which involve Aotearoa New Zealand.

- In 2020, Twitter partnered with the University of Otago's National Centre for Peace and Conflict Studies (@NCPACS), in a New Zealand first through our #DataforGood program. Our shared goal was to use Twitter data to study the ways online conversations can be used to promote tolerance and inclusion instead of division and exclusion. In the aftermath of the

horrific Christchurch terror attacks on 15 March 2019, Twitter and NCPACS embarked on a joint research project aimed at analysing the ways in which Twitter and social media more broadly was used for both positive and negative purposes.

- In tandem with our multilateral efforts with industry and as signatories to the Christchurch Call to Action, we've been resolutely focused on the ways Twitter can tackle extremism while promoting dialogue and inter-community understanding. Our goal is to harness the unique power of our service by providing data that empowers research, and hopefully gain a better understanding of the risks and the opportunities available through fostering an open, public conversation. Along with Governments of New Zealand and the USA, and Microsoft, we are investing in a technology innovation initiative under the banner of the Christchurch Call. The Christchurch Call Initiative on Algorithmic Outcomes will support the creation of new technology to understand the impacts of algorithms on people's online experiences.

Outcome 13. Support independent evaluation of the systems, policies and processes that have been implemented in relation to the Code.

Measure 45. Commit to selecting an independent third-party organisation to review the annual compliance reports submitted by Signatories, and evaluate the level of progress made against the Commitments, Outcomes and Measures, as outlined in section 4, as well as commitments made by Signatories in their Participation Form (see Appendix 2).

- As a signatory to the Code, Twitter supports and commits to selecting an independent third-party organisation to review compliance with the Code.

Conclusion

We are committed to protecting the health of the conversation on Twitter. We're constantly working to ensure our service is a place where all people can safely participate in the public conversation. We are committed to increasing the collective health on the platform, and this includes promoting content and behaviour that addresses people's needs, or that users will find important or useful.

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 Centre.

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. 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, non-profit 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.