



Australian Code of Practice on Disinformation and Misinformation Twitter Annual Transparency Report Reporting Period: 2022

Summary

Twitter’s objective is to be an authentic, informative, entertaining and trusted platform which empowers citizens’ free speech, an essential bedrock of a functional democracy. In parallel, we believe we must tackle bad actors and prevent manipulation.

In Q4 2022, the company embarked on transformational change. New approaches are vital so our service and company can thrive. That work is reflected in our ongoing commitments to protecting user rights and safety.

Moving beyond a leave-up or take-down binary, protecting the space for public debate and providing full transparency are key to the Twitter service going forward.

As a signatory to the Australian Code of Practice on Misinformation and Disinformation (“the Code”), this latest report reflects that major transformation underway. Since the technical reporting period also covers a substantial window prior, some examples and data may be unavailable at this time or less relevant given the new foundation forward for the Twitter service.

We endeavour to provide useful updates and examples of new products and policies related to enforcement, transparency, knowledge-sharing and authenticity for this report. These include but are not limited to Community Notes, the Twitter Blue subscription, and our Open-sourced algorithms.

Our updates to the Twitter rules set clear guidelines on what is allowed and they will continue to evolve as behaviors and threats change. We have added information about these that is also relevant to 2023 context and clarity given the above changes. For example, our Freedom of Speech not Reach philosophy announced in Q4 continues to see additions pertinent to topics of and commitments under the Code. By Q2 2023 we were applying our first *visibility filters* on content, combining our approaches to enforcement with our commitments to transparency at Twitter 2.0.

Twitter’s Commitments under the Code

Outcome	Status
Objective 1: Provide Safeguards against Harms that may arise from Disinformation and Misinformation	
Outcome 1a: Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures.	Opt-in
Outcome 1b: Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.	Opt-in



Outcome 1c: Users can report content and behaviours to Signatories that violates their policies under 5.10 through publicly available and accessible reporting tools	Opt-in
Outcome 1d: Users will be able to access general information about Signatories actions in response to reports made under 5.11.	Opt-in
Outcome 1e: Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.	Opt-in
Objective 2: Disrupt advertising and monetisation incentives for Disinformation	
Outcome 2: Advertising and/or monetisation incentives for Disinformation are reduced.	Opt-in
Objective 3: Work to ensure the security and integrity of services and products delivered by digital platforms	
Outcome 3: The risk that Inauthentic User Behaviours undermine the integrity and security of Services and Products is reduced.	Opt-in
Objective 4: Empower consumers to make better informed choices of digital content	
Outcome 4: Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.	Opt-in
Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms	
Outcome 5: Users are better informed about the source of Political Advertising.	Not applicable in Australia
Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research	
Outcome 6: Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.	Opt-in
Objective 7: Signatories publicize the measures they take to combat Disinformation and Misinformation	
Outcome 7: The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.	Opt-in

Objective 1: Provide Safeguards against Harms that may arise from Disinformation and Misinformation.

Outcome 1a: Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures.



At Twitter we continually evolve our policies and products to address new challenges and online behaviours and we've adopted a range of measures to reduce the risks of harms.

- Launched, continue to expand the ability for users to meaningfully participate in content being more informative on Twitter or with additional context via Community Notes.
- Launched and continue to expand a new system of verification, or checkmarks, on Twitter. The Twitter Blue subscription is one example.
- Continued investment tackle manipulation¹ and spam on our service with updates to our policies and enforcement².

We focus on Community Notes (“notes”) as one the most important and scalable ways to address and combat misinformation on Twitter³. We know that misleading information is complex, evolving, and sometimes cloaked behind questions or opinions. To ensure that people are better informed on Twitter we launched Community Notes, our approach to offering context and surfacing credible information⁴. As Community Notes rapidly evolve on platform, within Twitter, and in public, this product presents a profound shift for our company and people who use our service. It is a priority area of development, grounded in more than a decade and half of Twitter, the platform and content moderation experiments, policies, and products. It is also grounded in ongoing research, evaluation and consultation.

¹ <https://help.twitter.com/en/rules-and-policies/platform-manipulation>

² <https://help.twitter.com/en/rules-and-policies/manipulated-media>

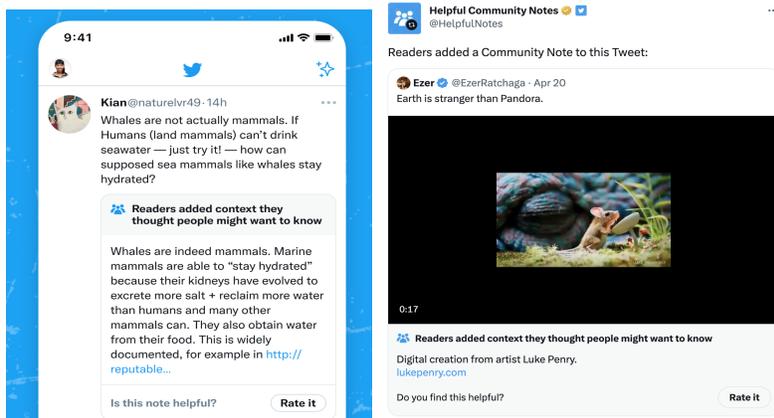
³ <https://help.twitter.com/en/using-twitter/community-notes>

⁴ *Id.*



CASE STUDY 1: Community Notes

Community Notes' goal is to show helpful context to people when they come across potentially misleading information on Twitter. When a note is rated Helpful by contributors, it starts being shown directly on the Tweet. Here's how they show up:



Contributors in Australia can leave notes on any Tweet and if enough contributors from different points of view rate that as helpful, the note will be publicly shown on a Tweet. Community Notes *do not* represent Twitter's viewpoint and cannot be edited or modified by our teams. A Tweet with a Community Note will not be labeled, removed, or addressed by Twitter unless it is found to be violating the Twitter Rules,⁵ Failure to abide by the rules can result in one's removal from accessing Community Notes, and/or other remediations.

In November 2022 we announced the initial Birdwatch experiment previously reported was expanding and evolving to become "Community Notes"⁶.

It is made up of independent contributors, and individual notes are never written by us. This is intentional, as it helps ensure our efforts to address potentially misleading information are informed by a diverse group of people who use our service. It is designed to surface *notes* that are informative and helpful to as many people as possible thanks in large part to what's known as a bridging algorithm. Most notes contain additional sources that can be clicked for an even deeper dive into a subject or conversation. They would also have the ability to rate the notes they see to help us understand if they're helpful or not.

We published a white paper⁷ setting out the intention and functionality of the product, along with making data available to researchers on the use of the feature.

⁵ <https://help.twitter.com/en/rules-and-policies/twitter-rules.html>

⁶ https://blog.twitter.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us

⁷ https://github.com/twitter/communitynotes/blob/main/birdwatch_paper_2022_10_27.pdf

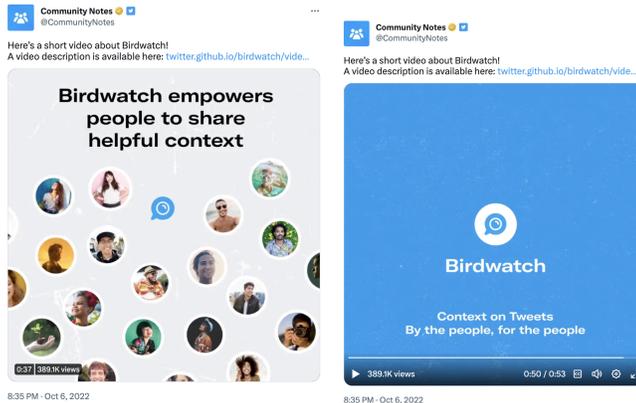


Table 1: Birdwatch/Community Notes on Twitter 2022⁸

Users can now also see notes on Tweets that are embedded in articles and websites and get more context wherever they are reading Tweets⁹.



Step 1 of 3: Your account meets the program requirements

To join the Community Notes pilot program, your Twitter account must have:

- No Twitter Rules violations since Jan 1, 2023
- Joined Twitter at least 6 months ago
- A verified phone number

That phone number must be:

- A trusted phone carrier
- Not associated with other Community Notes accounts

Next



Step 2 of 3: Agree to uphold the Community Notes values

As a member of the pilot program, you agree to:

- Contribute to build understanding 
- Act in good faith 
- Be helpful, even to those who disagree 

Next



Step 3 of 3: Agree to public contributions

We want everyone to be able to understand and evaluate how Community Notes works. So all your notes and rankings are publicly available, even if your account is protected. [Find out how Community Notes shares data.](#)

We may contact you for feedback by Tweet or Direct Message. Taking part in this research is optional.

Contributions are also subject to the [Twitter Rules](#), [Terms of Service](#), and [Privacy Policy](#).

Agree and finish

Community Notes



You signed up for Community Notes. We'll reach out when you're ready to go. For now, follow along @CommunityNotes.

Table 2: How to sign up for Community Notes

Community Notes were made visible around the world, including, in Australia¹⁰, and Twitter started admitting contributors from Australia in January 2023¹¹. We admit new contributors in batches, growing the contributor base by ~10% per week as we are monitoring quality and continuing to expand over time. Notes are evolving both globally and in Australia, and one positive trend is that we are seeing more contributors signing up.

⁸ <https://twitter.com/CommunityNotes/status/1578001121855012864>
⁹ <https://twitter.com/CommunityNotes/status/1652079596110594049>
¹⁰ <https://twitter.com/CommunityNotes/status/1601753552476438528>
¹¹ <https://twitter.com/CommunityNotes/status/1616504999844130816>



Community Notes help users to make more informed choices about the source of news and factual content on Twitter by empowering people on Twitter to collaboratively add context to potentially misleading Tweets¹². It aims to create a better-informed world.

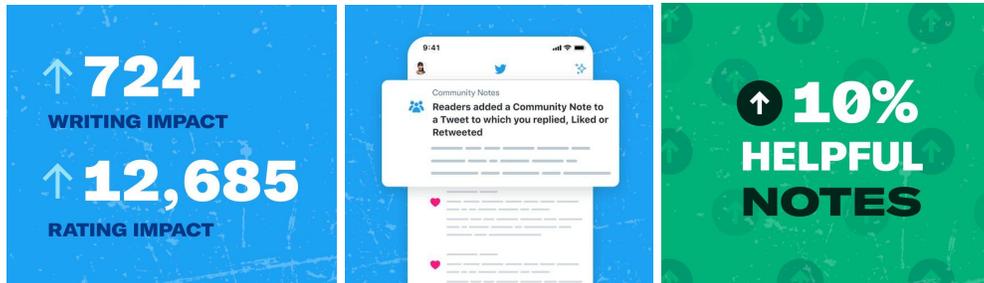


Table 3: Community Notes help users get well-informed on Twitter.

Key dates

- In January 2023, we launched an algorithm update that keeps note statuses and contributor impact scores more stable as we expand Community Notes. In addition to helping going forward, it boosts existing contributor impact scores by better recognizing helpful past contributions¹³. [See Table 3](#)
- In February 2023, users received a heads up if a Community Note starts showing on a Tweet they have replied to, Liked or Retweeted. This helps give people extra context that they might otherwise miss¹⁴ [See Table 3](#)
- In April 2023, Twitter shared the result of our algorithm update which boosts the number of Helpful Notes by 10%+. It builds change that uses confidence scores to identify notes that are broadly found helpful with high precision¹⁵. [See Table 3](#)

CASE STUDY 2: Twitter Blue and Expanded Verification

Twitter applies visual identity signals like labels and checkmarks on account profiles to provide more context about — and help distinguish — different types of accounts¹⁶ Some of these indicators are applied by Twitter, while others are triggered by user action. These show one of Twitter’s efforts to combat inauthentic accounts, misinformation and disinformation.

- In December 2022 Twitter Blue subscriptions became available in Australia¹⁷. Twitter Blue is one of a range of scalable measures to elevate quality conversations¹⁸. It is an opt-in, paid subscription. Tweets from verified users will be prioritized in places — helping to fight scams and spam.¹⁹ Only Twitter accounts created more than 30 days prior can sign up. Once subscribed to Twitter Blue, changes to their profile photo, display name, or username (@handle) will result in the loss of the blue checkmark until

¹² <https://help.twitter.com/en/using-twitter/community-notes>

¹³ <https://twitter.com/CommunityNotes/status/1616641911502303234>

¹⁴ <https://twitter.com/CommunityNotes/status/1628158167006994436>

¹⁵ <https://twitter.com/CommunityNotes/status/1649473048947597312>

¹⁶ <https://help.twitter.com/en/rules-and-policies/profile-labels>

¹⁷ <https://verified.twitter.com/en>

¹⁸ <https://help.twitter.com/en/using-twitter/twitter-blue>

¹⁹ <https://verified.twitter.com/en>



the account is validated as continuing to meet our requirements, and no further changes will be allowed during this review period²⁰.



Table 4: Twitter’s different profile labels in Blue, Gold and Grey colors

- **Blue checkmark:** The blue checkmark means that an account has an active subscription to Twitter Blue and meets our eligibility requirements. In April 2023 we removed legacy verified checkmarks²¹ and communicated that to remain verified on Twitter, individuals can sign up for Twitter Blue²².
- **Gold checkmark and square profile picture:** The gold checkmark indicates that the account is an official business account through Twitter Verified Organizations²³.
- **Grey checkmark:** The grey checkmark indicates that an account represents a government/multilateral organization or a government/multilateral official. Additional government and multilateral accounts can receive grey checkmarks through Verified Organizations²⁴.
 - December 2022: Twitter users started seeing additional icons that provide context for accounts on Twitter. In addition to blue and gold checks, Twitter introduced grey checks for government and multilateral accounts and square affiliation badges for select businesses²⁵.
 - March 2023: Twitter accepts applications for grey checkmarks for eligible government and multilateral accounts²⁶.

Outcome 1b: Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.

On Twitter, users are informed about the types of behaviours and types of content that are prohibited and/or managed. The Twitter rules²⁷ set the guidelines for what is allowed on our platform and continue to evolve as threats and behaviours change.

Here is an overview of key related policies enforced during the reporting period:

²⁰ <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts>
²¹ <https://twitter.com/verified/status/1648764138452299778>
²² https://twitter.com/i/twitter_blue_sign_up
²³ <https://help.twitter.com/en/rules-and-policies/profile-labels>
²⁴ <https://help.twitter.com/en/rules-and-policies/profile-labels>
²⁵ <https://twitter.com/TwitterSupport/status/1604955466727047168>
²⁶ <https://twitter.com/TwitterSafety/status/1638998382562910208>
²⁷ <https://help.twitter.com/en/rules-and-policies/twitter-rules>



- **Misleading and Deceptive Identity policy:** Twitter accounts that pose as another person, brand, or organisation in a confusing or deceptive manner may be permanently suspended under this policy.
- **Platform Manipulation and Spam policy:** People on Twitter 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 experiences on Twitter.
- **Synthetic and manipulated media policy:** People on Twitter may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm.

Examples of updates to the above:

Synthetic and manipulated media policy²⁸ The policy added information regarding out-of-context media. You 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.

COVID-19 Misleading Information Policy Twitter was no longer enforcing a separate and dedicated COVID-19 misleading information policy on its service from late November 2022²⁹. Much has changed since the globally declared pandemic of 2020 to present day. We note the exiting of emergency phases, lifted travel, quarantine and removed vaccination reporting requirements. There was recognition that important issues of public debate had been stifled. Twitter will continue to evaluate the viability of rapidly changing, single-issue policies going forward for our unique service, particularly in crisis situations.

September 2022	Total Since January 2020
520 accounts challenged	11.72M accounts challenged
173 accounts suspended	11,230 accounts suspended
602 content removed	97,674 content removed
In the month of August, we challenged 520 accounts, suspended 173 accounts, and removed 602 pieces of content globally.	Since introducing our COVID-19 guidance last year, we have challenged 11.72 million accounts, suspended 11,230 accounts, and removed over 97,674 content worldwide as of September 2022.

Table 4: Twitter’s Total accounts vs August 2022, COVID-19 pandemic

A scalable approach that aligns the enforcement philosophy of Twitter 2.0, **Freedom of Speech not Reach**³⁰, with our commitment to transparency is relevant. The former focuses on, where appropriate, making content that may violate our policies less discoverable. Going forward, we’re bringing increased visibility to those actions on our platform. While this is rolling out in a

²⁸ <https://help.twitter.com/en/rules-and-policies/manipulated-media>

²⁹ <https://transparency.twitter.com/en/reports/covid19.html#2021-jul-dec>

³⁰ https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy



phased approach it is one that allows Twitter to show how and where its policies are being enforced where doing so will not inform and arm bad actors.

In terms of enforcement, we apply our Rules³¹ consistently for all people on our service and regularly update our policies. We also use a combination of manual review and technology to help us enforce our rules. We continue to strengthen our service by building new defences such as improving our auto-detection technology against attempted manipulation, which includes malicious automated accounts, spam, as well as other activities that violate our TOS.

Outcome 1c: Users can report content and behaviours to Signatories that violate their policies under 5.10 (1b) through publicly available and accessible reporting tools.

We have a range of dedicated tools available for all of our users³² to report content that may violate our rules and policies and now meaningfully contribute directly to the service via Community Notes by adding or rating Notes for tweets which may be made public. As reported, Community Notes were visible in Australia in Q4 2022 and were made available for Australian contributors in Q1 2023.

On reporting Twitter has publicly available and accessible robust reporting forms, both in-app, on web and via our Help Centre where users can report 24/7, and they will be notified once our team has reviewed and taken enforcement action, where appropriate³³. Users can report Tweets, Lists, and Direct Messages that are in violation of our Rules or our TOS³⁴.

For Community Notes, Twitter created the ability for tweet authors to request additional review if they disagree that a Community Note is “helpful” or provides important context to their tweet³⁵. We’ve made publicly available that process for review, with simple information for how and where to report³⁶.

Our safety and security features, Help Centre pages and FAQs about in-app services³⁷ are in place to minimise end-users’ exposure to harmful content, empower end-users to manage their safety on Twitter and mitigate the impact that may arise from the propagation of misinformation and disinformation. When Twitter takes enforcement actions, we may do so either on a specific piece of content (e.g., an individual Tweet or Direct Message), on an account, or employ a combination of these options. In some instances, this is because the behaviour violates the Twitter Rules. Other times, it may be in response to a valid legal request from an authorised entity in a given country³⁸.

³¹ <https://help.twitter.com/en/rules-and-policies/twitter-rules>

³² <https://help.twitter.com/en/safety-and-security>

³³ <https://help.twitter.com/en>

³⁴ <https://help.twitter.com/en/safety-and-security/report-a-tweet>

³⁵ <https://help.twitter.com/en/using-twitter/community-notes>

³⁶ <https://communitynotes.twitter.com/guide/en/contributing/additional-review>

³⁷ <https://help.twitter.com/en>

³⁸ <https://help.twitter.com/en/rules-and-policies/enforcement-options>

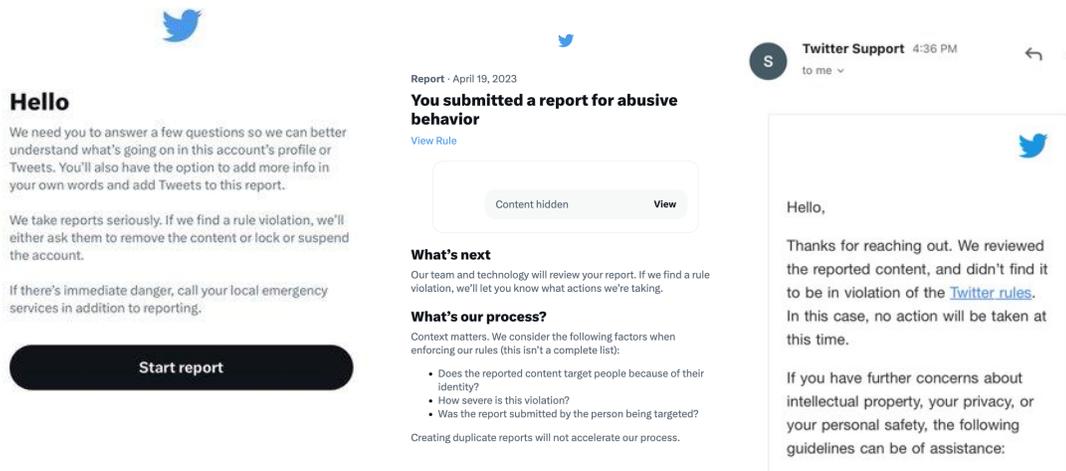


Table 5: Screenshots of how to report and user notifications from Twitter

Outcome 1d: Users will be able to access general information about Signatories actions in response to reports made under 5.11 (1c)

Transparency is fundamental to Twitter. Users will be able to access general information about Twitter’s actions in response to their reports. We are reviewing our approach to transparency reporting in light of innovations in content moderation and changes in the regulatory landscape. Twitter went ahead however and published our 21st report, with data on our policy enforcement for the first half of 2022 summarizing our health & safety efforts³⁹.

Over the reporting period, Twitter required users to remove 6,586,109 pieces of content that violated the Twitter Rules, an increase of 29% from H2 2021. We took enforcement action on 5,096,272 accounts during this period, a 20% increase, and 1,618,855 accounts were suspended for violating the Twitter Rules, a 28% increase⁴⁰.

Twitter continues to take action on content that violates our Rules and protects users’ rights in response to government legal requests. We intend to share more about our path forward for transparency reporting in later 2023. In the meantime, we will continue to give users insights into Twitter’s work to promote entertaining, informative and healthy conversation.

³⁹ <https://twitter.com/TwitterSafety/status/1650952198451499008>

⁴⁰ https://blog.twitter.com/en_us/topics/company/2023/an-update-on-twitter-transparency-reporting



Policy	Accounts actioned	Accounts suspended	Content removed
Abuse/Harassment	1,083,788	96,284	1,524,067
Child Sexual Exploitation	696,015	691,704	11,927
Hacked Materials	65	0	135
Hateful Conduct	1,085,651	111,056	1,527,442
Illegal or Certain Regulated Goods or Services	399,297	249,328	1,365,341
Impersonation	266,034	249,572	19,798
Misleading and Deceptive Identities	2	0	2
Non-Consensual Nudity	68,714	16,670	115,226
Perpetrators of Violent Attacks	381	0	1,578
Private Information	45,844	2,536	78,357
Promoting Suicide or Self Harm	439,555	11,776	547,377
Sensitive Media	1,315,670	150,757	1,352,155
Terrorism/Violent Extremism	30,616	30,616	0
Violence	28,753	19,838	35,240

Table 6: Twitter’s transparency report for 1 January - 30 June 30, 2022

In December 2022, we also reported that that over a 10 day period, our efforts to reduce spam on Twitter led to 85% reduction in reports of spammy group Direct Messages ~90% reduction in spammy Direct Messages sent by accounts users do not follow⁴¹.

Outcome 1e: Users will be able to access general information about Signatories’ use of recommender systems and have options relating to content suggested by recommender systems.

On Twitter, users have a range of options for controlling their experience especially via the For you, Following timelines⁴², Subscribed accounts, and via Twitter Lists⁴³.

- **For you** shows Tweets based on recommendations of accounts users follow or topics they are interested in;
- **Following** timeline displays Tweets from only the accounts users follow;
- Accounts that users are **Subscribed** to; and
- Users can also pin their favorite **Twitter Lists** to the top of their own timeline, giving an additional control of their home timeline

Users of Twitter can easily access information about these via the app and our Help Center.

⁴¹ <https://twitter.com/TwitterEng/status/1606083185502089217>

⁴² <https://help.twitter.com/en/using-twitter/twitter-timeline>

⁴³ <https://help.twitter.com/en/using-twitter/twitter-lists>



Home

For you

Following

Subscribed

Twitter List

Table 7: Home timelines with For You, Following, Subscribed, Twitter Lists)

When users choose Twitter on the **For you** or **Following** tabs, users will return to whichever timeline they had open last. Users also see content such as promoted Tweets or Retweets in their timeline. Users also see features that help them manage the For you timeline. We make recommendations to make it easier and faster to find content that contributes to the conversation in a meaningful way, such as content that is relevant, credible, and safe. This means users will sometimes see Tweets from accounts they do not follow. We recommend Tweets to users based on who they already follow and Topics they follow, and do not recommend content that might be abusive or spammy. We share recommendations via push notifications, Notifications tab, and by adding them to users' For you timeline.

Twitter Lists allows users to customize, organize and prioritize the Tweets users see in their timeline. Users can also choose to join Lists created by others on Twitter, or from your own account you can choose to create Lists of other accounts by group, topic or interest. Viewing a List timeline will show users a stream of Tweets from only the accounts on that List⁴⁴. In Users' Home timeline on Twitter for iOS and Android apps, users might see a prompt to Discover new Lists. If we suggest a List to users that's of interest, they can simply tap Follow. From the prompt, users can also tap Show more to browse through our Lists discovery page. There, we will show users more Lists we might think they will like to follow and they can search for additional Lists in the search box at the top of the page. We will also show you recommendations from the Lists they follow right in their For you timeline.

One of the most significant changes to Twitter is our transparency here is about our **recommendations algorithm**. Twitter published a blog to introduce how the algorithm selects Tweets for the user's timeline⁴⁵. Our recommendation system is composed of many interconnected services and jobs. While there are many areas of the app where Tweets are recommended—Search, Explore, Ads—we focused on the home timeline's For You feed: Every day, we serve over 150 billion Tweets to people's devices. Ensuring that we are delivering the best content possible to our users is both a challenging and an exciting problem. We are working on new opportunities to expand our recommendation systems—new real-time features, embeddings, and user representations—and we have one of the most interesting datasets and user bases in the world to do it with. We are building the town square of the future. This requires a recommendation algorithm to distill the roughly 500 million Tweets posted daily down to a handful of top Tweets that ultimately show up on their device's For You timeline.

This was first shared in March 2023, when Twitter announced a new era of transparency opening much of our source code to the global community and how on GitHub, users can find two new repositories containing the source code for many parts of Twitter, including our **recommendations algorithm**, which controls the Tweets users see on the For You timeline. *

⁴⁴ <https://help.twitter.com/en/using-twitter/twitter-lists>

⁴⁵ https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm



- We shared more information on our recommendation algorithm post on our Engineering Blog⁴⁶. For this release, we aimed for the highest possible degree of transparency, while excluding any code that would compromise user safety and privacy or the ability to protect our platform from bad actors, including undermining our efforts at combating child sexual exploitation and manipulation.
- We also took additional steps to ensure that user safety and privacy would be protected, including our decision not to release training data or model weights associated with the Twitter algorithm at this point.

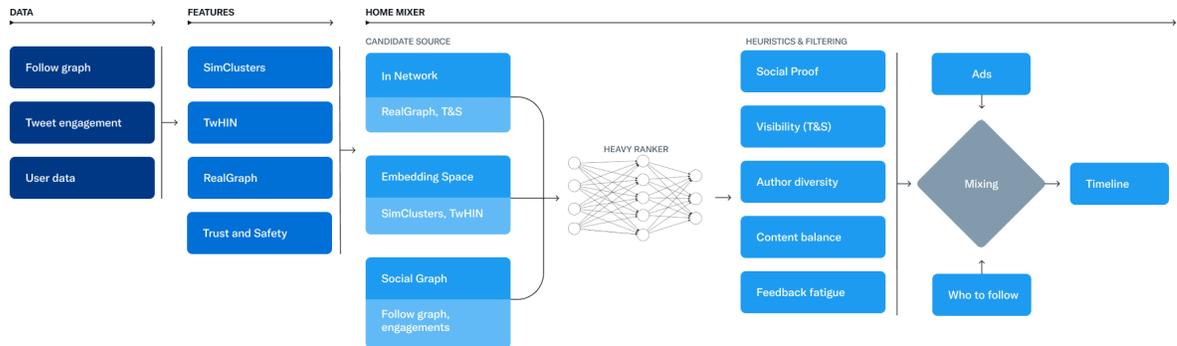


Table 8: The major components used to construct a timeline

Additional detail: Twitter has several Candidate Sources that we use to retrieve recent and relevant Tweets for a user. For each request, we attempt to extract the best 1500 Tweets from a pool of hundreds of millions through these sources. We find candidates from people users follow (In-Network) and from people they do not follow (Out-of-Network). As of March 2023, the For You timeline consists of 50% In-Network Tweets and 50% Out-of-Network Tweets on average, though this may vary from user to user.

The goal of the For You timeline is to serve users relevant Tweets. At this point in the pipeline, we have ~1500 candidates that may be relevant. Scoring directly predicts the relevance of each candidate Tweet and is the primary signal for ranking Tweets on users' timeline. At this stage, all candidates are treated equally, without regard for what candidate source it originated from. Ranking is achieved with a ~48M parameter neural network that is continuously trained on Tweet interactions to optimize for positive engagement (e.g. Likes, Retweets, and Replies). This ranking mechanism takes into account thousands of features and outputs ten labels to give each Tweet a score, where each label represents the probability of an engagement. We rank the Tweets from these scores.

At this point, Home Mixer has a set of Tweets ready to send to the user's device. As the last step in the process, the system blends together Tweets with other non-Tweet content like Ads, Follow Recommendations, and Onboarding prompts, which are returned to their device to display. The pipeline above runs approximately 5 billion times per day and completes in under 1.5 seconds on average. A single pipeline execution requires 220 seconds of CPU time, nearly 150x the latency users perceive on the app.

⁴⁶ *Id.*



The goal of our open source endeavor is to provide full transparency to our users about how our systems work. We have released the code powering our recommendations so that users can view to understand our algorithm in greater detail⁴⁷, and we are also working on several features to provide users greater transparency within our app. Some of the new developments we have planned include: A better Twitter analytics platform for creators with more information on reach and engagement, *Greater transparency into any safety labels applied to their Tweets* or accounts, and *Greater visibility into why Tweets appear on users' timeline*. The latter two are reflected in launch and first updates to Freedom of Speech not Reach and visibility filtering⁴⁸ which are underway (see further down in report).

Objective 2: Disrupt advertising and monetisation incentives for Disinformation and Misinformation

Outcome 2: Advertising and/or monetisation incentives for Disinformation and Misinformation are reduced.

As reported above, as we evolve, we are giving people greater transparency and control over their experience on the platform, and this includes our advertisers. We have also been working to improve the advertising experience on Twitter by making ads more relevant. Underpinning these efforts is our work to ensure ads appear in brand-suitable environments. Ensuring that the context in which ads appear does not conflict with a brand's message and values is foundational to delivering a safe, relevant, and informative experience for everyone on Twitter⁴⁹.

In December 2022, we announced Adjacency Controls for advertisers, and shared an update on our partnerships with Integral Ad Science and DoubleVerify to provide independent reporting on the context in which ads appear on Twitter⁵⁰:

- **Adjacency Controls:** We started giving advertisers pre-bid controls to prevent their ads from appearing adjacent to Tweets that use keywords they'd like to avoid in relevance-ranked Home Timelines (the vast majority of timelines on Twitter). To start, these controls apply to adjacent Tweets in English only, and we'll roll out to other languages shortly. Empowering brands to customize their campaigns to prevent their ads from appearing adjacent to unsuitable content is an important step towards increased ad relevance on Twitter. This feature builds on Twitter's longstanding suite of brand safety protections and controls for advertisers, and we'll continue to evolve these solutions over time⁵¹.
- **3rd-Party Brand Safety Measurement:** We expanded our partnerships with industry-leading brand safety partners DoubleVerify and Integral Ad Science. In early Q1 2023, DoubleVerify and Integral Ad Science are offering their post-bid brand safety reporting for Tweets in our Home Timeline at scale to our advertising partners. This reporting will give advertisers transparency on the context in which their ads served, according to the GARM Brand Safety & Suitability Framework. These partnerships will provide independent validation of Twitter's efforts to uphold the GARM Brand Safety

⁴⁷ <https://github.com/twitter/the-algorithm> and <https://github.com/twitter/the-algorithm-ml>

⁴⁸ https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy

⁴⁹ <https://business.twitter.com/en/blog/adjacency-controls-third-party-measurement.html>

⁵⁰ <https://twitter.com/TwitterBusiness/status/1604584937788739585>

⁵¹ <https://business.twitter.com/en/blog/adjacency-controls-third-party-measurement.html>



Floor and prevent placements unsafe for all advertisers, in addition to giving advertisers reassurance on the effectiveness of Adjacency Controls to ensure brand suitability⁵².

Promoted content on Twitter must adhere to the Twitter rules and our advertising policies. People using Twitter can also make reports related to Twitter Ads that might potentially violate our policies⁵³. These will be assessed against the Twitter Ads Policy⁵⁴, the Twitter Rules⁵⁵ and TOS⁵⁶ and any enforcement action will be taken in line with these policies. As mentioned above, Twitter uses a combination of human review and technology to help us enforce our rules. Our specially trained team reviews and responds to reports 24/7; they have the capacity to review within context and respond to reports in multiple languages. In addition, we publish specific policies for advertisers that share standards for that are outlined below.

- **Political content advertising policy:** During the reporting period we continued to prohibit political advertising in Australia and as reflected under our political content advertising policy⁵⁷. We'll continue to review and update these.
- **Inappropriate content advertising policy:** Our policy on inappropriate content advertising⁵⁸ prohibits advertising deemed to be dangerous or exploitative, misrepresentative, along with misleading synthetic or manipulated content and content engaged in coordinated harmful activity.
- **Quality advertising policy:** Our quality advertising policy⁵⁹ 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.

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

Outcome 3: The risk that Inauthentic User Behaviours undermine the integrity and security of Services and Products is reduced.

Transparency is core to our mission. Our goal with these changes is to provide more transparency about more issues, while grappling with the considerable safety, security, and integrity challenges in this space. We want Twitter to be a place where people can make human connections, find reliable information, and express themselves freely and safely. To make that possible, we do not allow spam or other types of platform manipulation⁶⁰. Twitter has focused on verifications, anti-spam efforts and our renewed investment in proactive detection and disruption.

As shared in **Case Study 2 (Outcome 1a)**, Twitter launched Twitter Blue Subscription as a part of new and expanded verification to help reduce inauthentic user behaviours that are trying to undermine the integrity and security of Twitter. Only Twitter accounts created more than 30 days

⁵² *Id.*

⁵³ <https://help.twitter.com/en/safety-and-security/reporting-twitter-ads>

⁵⁴ <https://business.twitter.com/en/help/ads-policies.html>

⁵⁵ <https://help.twitter.com/en/rules-and-policies/twitter-rules>

⁵⁶ <https://twitter.com/en/tos>

⁵⁷ <https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html>

⁵⁸ <https://business.twitter.com/en/help/ads-policies/ads-content-policies/inappropriate-content.html>

⁵⁹ <https://business.twitter.com/en/help/ads-policies/ads-content-policies/quality-policy.html>

⁶⁰ <https://help.twitter.com/en/rules-and-policies/platform-manipulation>



ago can sign up for Twitter Blue. All Twitter Blue subscribers are required to confirm their phone number as part of sign up. Once subscribed to Twitter Blue, changes to users' profile photo, display name, or username (@handle) will result in the loss of the blue checkmark until the account is validated as continuing to meet our requirements, and no further changes will be allowed during this review period⁶¹.

The consequences for violating Twitter Rules and policies depend on the severity of the violation as well as any previous history of violations. Our action is also informed by the type of spammy activity that we have identified. The actions we take may include the following⁶²:

- **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.
- **Limiting the visibility of Tweets:** Tweets which we believe to be in violation of these policies may not appear in certain parts of the Twitter product, and/or may not be recommended or amplified by Twitter.
- **Tweet deletion and temporary account locks:** If the platform manipulation or spam offense is an isolated incident or first offense, 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 offenses will result in permanent suspension. In the case of a violation centering around the use of multiple accounts, users may be asked to choose one account to keep. The remaining accounts will be permanently suspended.
- **Temporary loss of access to Twitter features or products:** We may temporarily limit or restrict access to Twitter features or products, including (but not limited to) Tweets, Edit Tweet, Direct Messages, Spaces, Communities, or Live.
- **Permanent suspension:** For severe violations, accounts will be permanently suspended at first detection.

In April 2023 we announced we would be adding more transparency to the enforcement actions we take on Tweets. Restricting the reach of Tweets, also known as visibility filtering, is one of our existing enforcement actions that allows us to move beyond the binary "leave up versus take down" approach to content moderation. However, like other social platforms, we have not historically been transparent when we've taken this action⁶³.

As a first step, we reported that users will start to see labels on some Tweets identified as potentially violating our rules around Hateful Conduct letting them know we have limited their visibility⁶⁴. These actions will be taken at tweet level only, will not affect a user's account. While these labels initially only apply to a set of Tweets that potentially violate our Hateful Conduct policy, we plan to expand them to other applicable policy areas in the coming months. This change is designed to result in enforcement actions that are more proportional and transparent for everyone on our platform.

These labels bring a new level of transparency to enforcement actions by displaying which policy the Tweet potentially violates to both the Tweet author and other users on Twitter. Tweets

⁶¹ <https://help.twitter.com/en/using-twitter/twitter-blue>

⁶² <https://help.twitter.com/en/rules-and-policies/platform-manipulation>

⁶³ https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy

⁶⁴ <https://twitter.com/TwitterSafety/status/1647996597211484160>



with these labels will be made less discoverable on the platform. Additionally, we will not place ads adjacent to content that we label⁶⁵.

What remains unchanged with this approach our commitment to keeping Twitter a safe place for conversation. We continue to remove illegal content and suspend bad actors from our platform.

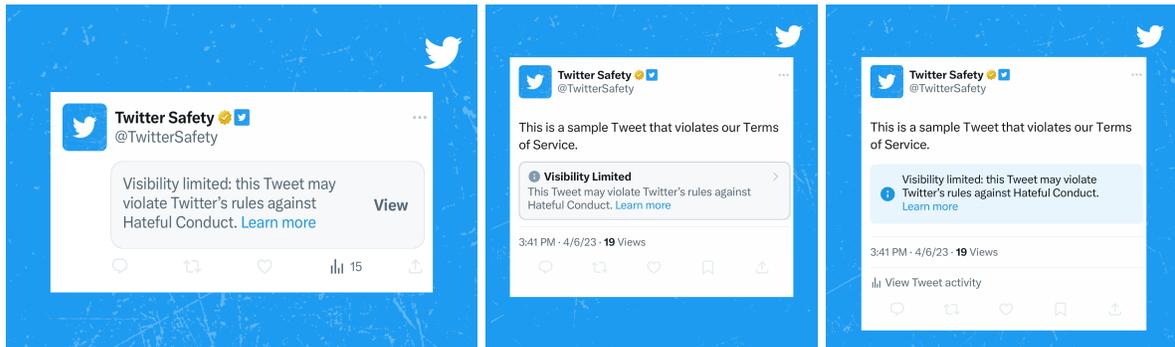


Table 9: Twitter Safety account shares display of visibility filtering on Twitter

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

Outcome 4: Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation

There are a range of ways users on Twitter are enabled to make more informed choices about content sources including news and factual content.

- As shared in detail in Outcome 1a, Community Notes is one of the most important ways we’re enabling users to make more informed choices about content they see on Twitter. We’re also empowering people on Twitter to collaboratively do so by adding context to potentially misleading Tweets⁶⁶.
- The Twitter checkmarks⁶⁷ provide visual identity signals on account profiles to provide more context about — and help distinguish — different types of accounts. For example, the blue checkmark means that an account has an active subscription to [Twitter Blue](#) and meets our eligibility requirements. The Gold checkmark indicates it is an official business account whereas the grey checkmark indicates an account represents a government/multilateral organization, or a government/multilateral official.
- We first reported in our 2021 report under the Code that Twitter’s read-before-you-tweet prompt, or nudge was underway. It continued into 2022 and we added languages to the experiment. Twitter started the prompt in 2020⁶⁸ “when people Retweet an article that they haven’t opened on Twitter, we may ask if they would like to open it first”. The prompt explored how users can make more self-informed choices on Twitter. Insights into the

⁶⁵ <https://help.twitter.com/en/rules-and-policies/enforcement-options>

⁶⁶ <https://help.twitter.com/en/using-twitter/community-notes>

⁶⁷ <https://help.twitter.com/en/rules-and-policies/profile-labels>

⁶⁸ <https://twitter.com/twittersupport/status/1270783537667551233?lang=en>



effectiveness of the prompt, and how it changed user behavior when they're shown the alert demonstrated:

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

[As outlined above relating to the Transparency report in Outcome 1d](#), Twitter recognises the importance of helping users identify trusted information and make informed choices in today's information ecosystem. We have adopted a number of strategies to signal authenticity, accuracy and more contexts to our users to get informed on the Twitter service with our Transparency report⁶⁹. As we review our approach to transparency reporting in light of innovations in content moderation and changes in the regulatory landscape, we shared the update to users of our 21st report, with data on our policy enforcement for the first half of 2022⁷⁰.

Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.

Outcome 5: Users are better informed about the source of Political Advertising.

Objective 5 is not applicable. For the reporting period and as of May 2023, there was no change in Twitter's political content advertising policy in Australia. Twitter began implementation of this Political Content Advertising Policy⁷¹ in 2019 and has previously reported on it under the Code. Twitter defines this as political content that references a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome. As with other policies, Twitter will continue to review and update.

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

Outcome 6: Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation

Over the years, hundreds of millions of people have sent over a trillion Tweets, with billions more every week. Twitter data are among the world's most powerful data sets. As of May 2023 the new Free, Basic, Pro and Enterprise Tiers of Access to the Twitter API are available, also to independent researchers. We're also looking at new ways to continue serving this community as a new company.

In mid-December 2022 we deprecated or paused existing projects and put them under review as part of the company transition. We reassured of our commitment to the Twitter Developer Platform⁷² and continued investments, especially the Twitter API. By March 30th 2023 we began relaunching, with announcements made also via @TwitterDev.

⁶⁹ <https://transparency.twitter.com/>

⁷⁰ <https://twitter.com/TwitterSafety/status/1650952198451499008>

⁷¹ <https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html>

⁷² <https://twitter.com/TwitterDev/status/1603823066496147456>



We continue to rapidly expand on these. For example by 3 May 2023, we announced that Verified Government or publicly owned services who tweet weather alerts, transport updates and emergency notifications may use the API for those critical purposes, for free⁷³.

Twitter is committed to the success of our Developer ecosystem. We will continue to build on these efforts and inform the public as we improve Twitter in the open. Below are recent notable efforts that Twitter has supported where independent researchers are adding to improve public understanding. As reported above, Twitter's review, relaunch and considerations for continued partnership are live and underway.

Name	Overview and description of the products/research
Community Notes transparency	<p>Twitter made the Community Notes algorithm publicly available on GitHub, along with the data that powers it, so anyone can audit, analyze, or suggest improvements.⁷⁴</p> <p>We accepted the first Community Notes code change from the public and reported the first time ever that we scored and displayed notes using code written by people outside the company⁷⁵. This change optimized a function that identifies explanatory tags that describe why raters found a note helpful or not. Twitter welcomes improvements like this, and would love to see contributions that strengthen note quality, adversarial resistance or other core elements of the system⁷⁶</p> <p>All Community Notes contributions are publicly available on the Download Data page of the Community Notes site so that anyone has free access to analyze the data, identify problems, and spot opportunities to make Community Notes better.⁷⁷</p>

⁷³ <https://twitter.com/TwitterDev/status/1653492584176656384>

⁷⁴ <https://twitter.com/CommunityNotes/status/1578004584320172034>

⁷⁵ <https://twitter.com/CommunityNotes/status/1629229535337058305>

⁷⁶ <https://twitter.com/CommunityNotes/status/1629229897691324416>

⁷⁷ <https://communitynotes.twitter.com/guide/en/under-the-hood/download-data>



<p>Independent assessment of hate speech on Twitter with Sprinklr⁷⁸</p>	<p>New independent reports using Twitter data: We reported in March 2023 a recent partnership with Sprinklr⁷⁹ for an independent assessment of hate speech on Twitter⁸⁰. Sprinklr’s AI-powered model found the reach of hate speech is even lower than our own model quantified⁸¹. Twitter announced the finding and Sprinkler’s report is publicly available⁸².</p> <table border="1"> <thead> <tr> <th colspan="2">Toxic Tweets</th> <th colspan="2">Non-Toxic Tweets</th> </tr> </thead> <tbody> <tr> <td> 79.83K Tweets</td> <td> 58.9K Distinct Users</td> <td> 472.64K Tweets</td> <td> 299.35K Distinct Users</td> </tr> <tr> <td> 12.72M Impressions</td> <td> 224.36 Average Impressions</td> <td> 277.37M Impressions</td> <td> 684.35 Average Impressions</td> </tr> <tr> <td> 229.98K Engagements</td> <td> 2.88 Average Engagements</td> <td> 4.28M Engagements</td> <td> 9.06 Average Engagements</td> </tr> </tbody> </table> <p>Table 10: Comparing Toxic Tweets with Non-Toxic Tweets</p> <p>When compared to non-toxic tweets in the dataset containing slur keywords, toxic tweets received 3 times fewer impressions on average.</p>	Toxic Tweets		Non-Toxic Tweets		79.83K Tweets	58.9K Distinct Users	472.64K Tweets	299.35K Distinct Users	12.72M Impressions	224.36 Average Impressions	277.37M Impressions	684.35 Average Impressions	229.98K Engagements	2.88 Average Engagements	4.28M Engagements	9.06 Average Engagements
Toxic Tweets		Non-Toxic Tweets															
79.83K Tweets	58.9K Distinct Users	472.64K Tweets	299.35K Distinct Users														
12.72M Impressions	224.36 Average Impressions	277.37M Impressions	684.35 Average Impressions														
229.98K Engagements	2.88 Average Engagements	4.28M Engagements	9.06 Average Engagements														
<p>Open-source algorithm</p>	<p>Twitter is constantly experimenting with the open-source algorithms that select which notes to show. So everyone can follow along, they can easily see in “<i>Note Details</i>” which model computed the current status of a note⁸³. In April 2023, Twitter shared another update on the changes we have made to our open source repos this week and a preview of what’s next⁸⁴.</p>																

Objective 7: Signatories publicise the measures they take to combat Disinformation and Misinformation.

Outcome 7: The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.

There are multiple locations where users can access information about Community Notes and its live development or actions being taken to address Disinformation and Misinformation. We put extensive work into updating, developing, and educating users on Twitter’s rules and enforcement actions. As described in detail in **Outcome 1a, 1c, 1d** and **Outcome 4** regarding measures to combat misinformation the public can access information about steps Twitter is taking now and plans. Twitter is committed to transparency in principle and in practice. We aim to improve the accessibility and usefulness to the public as well e.g. through the enforcement philosophy of Twitter 2.0, **Freedom of Speech not Reach⁸⁵** and bringing increased visibility to

⁷⁸ <https://twitter.com/TwitterSafety/status/1638255718540165121>
⁷⁹ <https://twitter.com/Sprinklr>
⁸⁰ <https://partners.twitter.com/en/partners/sprinklr>
⁸¹ <https://twitter.com/TwitterSafety/status/1638262108650348545>
⁸² <https://www.sprinklr.com/blog/identify-toxic-content-with-leading-analytical-ai/>
⁸³ <https://twitter.com/CommunityNotes/status/1636852370809257984>
⁸⁴ <https://twitter.com/TwitterEng/status/1652049665184137216>
⁸⁵ https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy



the actions on our platform. Publicising our disclosures in ways that are meaningful to the public and users of Twitter will continue to evolve as recently reported on the Twitter blog⁸⁶.

Conclusion

As a townsquare of the internet, we are constantly working to ensure our service is a place where all people can safely participate to find information that is important, useful or entertaining. This work is core to Twitter. Our approaches at Twitter 2.0, as outlined in this report, remain aligned with the guiding principles of the Code, including protecting freedom of expression and the commitment to keeping users safe and tackling bad actors. We trust this report provides an understanding of the serious resolve with which our company and teams approach this commitment.

One key aspect that changed in the reporting period and for the company is our approach to experimentation. As the public have seen over the past several months, Twitter is embracing public testing and substantial change. We believe that this open and transparent approach to innovation is healthy, as it enables us to move faster and gather user feedback in real-time. We believe that a service of this importance will benefit from feedback at scale, and that there is value in being open about our experiments and what we are learning. We do all of this work with one goal in mind: to improve Twitter for our customers, partners, and the people who use it in Australia and across the world. We look forward to continuing our work with Government, partners, civil society and industry to improve understanding of these complex issues.

⁸⁶ https://blog.twitter.com/en_us/topics/company/2023/an-update-on-twitter-transparency-reporting