



Annual Compliance Report September 2023

Aotearoa New Zealand Code of Practice for Online Safety and Harms

Signatory: X Corp

If applicable: Relevant Products / Services

More than half a billion people from around the world gather on X to talk about their interests in real-time. Our mission is to promote and protect the public conversation—to be the town square of the internet. X enables people to directly engage on important topics with elected representatives, local or national leaders and fellow citizens¹.

In 2022, the company embarked on transformational change. New approaches are vital so our service and company can thrive. We are dedicated to our part in addressing the complex, evolving risks that misinformation and disinformation harms can pose, with innovations and approaches right-sized to and recognizing our unique platform. X works to get in front of a range of tactics that people use to target the process. To do this we hire the right people, update our policies and evolve our product.

As a signatory to the Aotearoa New Zealand Code of Practice for Online Safety and Harms (“the Code”), this latest report reflects that major transformation underway, while providing examples of new products and policies related to enforcement, transparency, knowledge-sharing and authenticity. These include, but are not limited to, Community Notes, the X Premium subscription, and our Open-source algorithms.

Our updates to the X Rules set clear guidelines on what is allowed and they will continue to evolve as behaviors and threats change. For example, our Civic Integrity Policy² provides an extra layer of protection that is applied for a limited period of time before and during an election. We updated this policy to make sure we strike the right balance between tackling the most harmful types of content—those that could intimidate or deceive people into surrendering their right to participate in a civic process—and not censoring political debate.

One of our new approaches to enforcement, Freedom of Speech, Not Reach, where we restrict the reach of violating content, is, we believe, one of the most important ways we can combine our commitment to addressing content in a proportionate manner and with critical transparency. To this effect, we have started adding publicly visible labels to posts identified as potentially violating our Rules and policies, letting people know when their reach has been restricted.

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.

¹ https://blog.twitter.com/en_us/topics/company/2023/supporting-peoples-right-to-accurate-and-safe-political-discourse-on-x

² <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>



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

There is no place for abuse of minors on X. Over the past year we have strengthened our policies, deployed new automated technology, and increased the number of cybertips we send to the National Center for Missing & Exploited Children³. And while we are proud of what we have achieved, we know this is an area which is highly adversarial and bad actors change their approach in a rapid manner and therefore, requires constant updates and iteration to enforcement strategies. We are moving faster than ever to make X safer and keep child sexual exploitation (CSE) material off our platform.

- In September 2023, we announced that in addition to our partnerships with the Tech Coalition and Independent Women's Forum, we are investing in a partnership with Thorn to further strengthen our efforts. By utilizing the *Safer*, Built by Thorn tool, X has been able to significantly expand the scope of our proactive detection.
 - As a result of all these investments and updates, 95% of the accounts we suspend we find before any user reports, up from 75%. We will continue to share updates on our efforts in this critical area and remain unwavering in our determination to find and hold to account those who engage in this heinous crime.
- Our goal is to entirely remove CSE materials from X. Not only are we detecting more bad actors faster, we are building new defenses that proactively reduce the discoverability of posts that contain this type of content.
- In June 2023, we renewed our commitment with Tech Coalition to better prevent, detect, report, and remove online CSE materials⁴.
- In February 2023, we updated our approach to be more aggressive in that we are proactively and severely limiting the reach of any content that we detect may contain CSE material. This includes moving swiftly to remove the content and suspend the bad actor(s) involved⁵.
- In December 2022, we shared an update on our efforts to make Twitter safer and remove bad actors who create, distribute, or engage with CSE material. We improved our detection and enforcement methods and expanded our partnerships with organizations that help prevent the trafficking of CSE material. This resulted in 57% more CSE suspensions in November, which was significantly more than any other month year to date in 2022⁶.

³ <https://X.com/Safety/status/1700253217504862340>

⁴ <https://x.com/GlobalAffairs/status/1674528878184759296?s=20>

⁵ <https://twitter.com/Safety/status/1620908366062305280>

⁶ <https://x.com/Safety/status/1601439984292360193>



Table 1: Showing CSE Account Suspension from February 2022 to January 2023

At X, we train our internal enforcement teams on X's Terms Of Services⁷, with a focus on policies related to Child Sexual Exploitation⁸. Training takes place on a regular cadence and we also conduct refresher training to ensure that we retain important knowledge about our Rules, policies and enforcement options⁹ on X.

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

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

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

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

On 7 April 2023, we updated our Abuse and Harassment policy¹⁰ to clarify how we define targeted harassment. We believe in free speech and we also believe users have a right to use and enjoy our platform without being subjected to targeted and repeated harassment. We

⁷ <https://twitter.com/en/tos#:~:text=The%20X%20Entities%20make%20no,from%20your%20access%20to%20or>

⁸ <https://help.twitter.com/en/rules-and-policies/sexual-exploitation-policy>

⁹ <https://help.twitter.com/en/rules-and-policies/enforcement-options>

¹⁰ <https://help.X.com/en/rules-and-policies/abusive-behavior>



define “targeted harassment” as behavior that is repeated, unreciprocated, and intended to humiliate or degrade an individual(s). This includes targeting people based on gender, race, religion, or sexual orientation¹¹. You can learn more about this policy.¹²

Our mission at X is to promote and protect the public conversation, as a townsquare of the internet. We believe X users have the right to express their opinions and ideas without fear of censorship. We also believe it is our responsibility to keep users on our platform safe from content violating our Rules. These beliefs are the foundation of Freedom of Speech, not Freedom of Reach - our enforcement philosophy which means, where appropriate, restricting the reach of posts that violate our policies by making the content less discoverable.

On 17 April 2023, we shared an update on our approach to policy enforcement that better aligns this philosophy with our commitment to transparency. Restricting the reach of posts, also known as visibility filtering, is an enforcement action that allows us to move beyond the binary “leave up versus take down” approach to content moderation. We started adding publicly visible labels to a set of posts that potentially violate our Hateful Conduct policy to let users know we have restricted the reach of their post and plan to expand their application to other applicable policy areas in the coming months¹³.

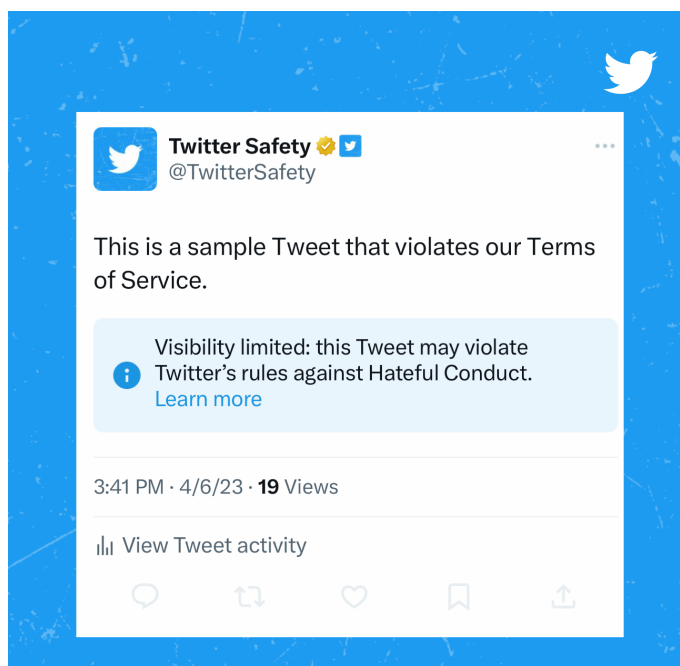


Table 2: Displaying how Visibility filtering works

We remain committed to maintaining free speech on X, while equally maintaining the health of our platform. For content that meets the threshold for enforcement under our Terms of Service we apply proportionally appropriate enforcement action. We will continue to remove the most serious violations of our Rules, such as content inciting or calling for violence, persistent

¹¹ <https://X.com/Safety/status/1644213174764449792>

¹² <https://help.twitter.com/en/rules-and-policies/abusive-behavior>

¹³ https://blog.X.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy



harassment and suspend bad actors from our platform¹⁴.

The team at X works hard every day to fight hate in all forms, and we invite all of our partners to participate in an open dialogue to help us make a difference. As we have said before, our doors are always open to organizations that want to actively partner with us to strengthen our policies and improve our approach. And we hope that all groups can realize the potential of working more productively with X – so we can all work together to protect free expression and simultaneously keep our platform safe for everyone¹⁵.

In December 2022, we updated our recent efforts to reduce the reach of hateful speech on X¹⁶. Counting the number of posts that contain a specific slur is not an accurate way to measure hateful speech. Context matters, and not all occurrences of slur words are used in a hateful way. Slur words may be used in counterspeech, reclaimed phrases, and song lyrics, for example. People will still see slur words in posts when they follow an account that uses them. However, we will not amplify posts containing hate speech or slurs when used in a hateful way, and we will not serve ads adjacent to those posts.

We have more work to do. We are working to introduce in-app transparency when we limit the reach of a post. We will continue to invest in the moderation of illegal and harmful content to ensure a safe platform for everyone.

Below are a few notable highlights of changes we have made since the beginning of 2023.

- **April 2023:** we updated our *Abuse and Harassment policy*¹⁷ to clearly define harassment. We believe in free speech and we also believe users have a right to use and enjoy our platform without being subjected to targeted and repeated harassment. We define “targeted harassment” as behaviour that is repeated, unreciprocated, and intended to humiliate or degrade an individual(s). This includes targeting people based on gender, race, religion, or sexual orientation¹⁸.
- **March 2023:** we updated our enforcement related to our *Violent Speech policy*¹⁹ by taking a more aggressive approach towards accounts sharing such content, which prohibits violent threats, wishes of harm, glorification of violence, and incitement of violence. X has a zero-tolerance approach towards Violent Speech, and in most cases, we will suspend any account violating this policy. For less severe violations, we may require accounts to delete the content before they can access their account again²⁰.
- **February 2023:** Another change we have made is to the *Violent and Hateful Entities policy*.²¹ There is no place on X for violent and hateful entities, including (but not limited to) terrorist organisations, violent extremist groups, perpetrators of violent attacks, or individuals who affiliate with and promote their illicit activities. The violence

¹⁴ https://blog.X.com/en_us/topics/product/2023/freedom-of-speech-not-reach--new-updates-and-progress

¹⁵ <https://twitter.com/Safety/status/1700253217504862340>

¹⁶ <https://x.com/Safety/status/1601619357188038656?s=20>

¹⁷ <https://help.X.com/en/rules-and-policies/abusive-behavior>

¹⁸ <https://X.com/XSafety/status/1644213178337984515>

¹⁹ <https://help.X.com/en/rules-and-policies/violent-speech>

²⁰ <https://X.com/XSafety/status/1630660504992489475>

²¹ <https://help.X.com/en/rules-and-policies/violent-entities>



and hate these entities engage in and/or promote jeopardises the physical safety of those targeted.

- In addition, we updated our *Perpetrators of Violent Attacks policy*.²² We will remove any accounts maintained by individual perpetrators of terrorist, violent extremist, or mass violent attacks, as well as any accounts glorifying the perpetrator(s), or dedicated to sharing manifestos and/or third party links where related content is hosted. We may also remove posts disseminating manifestos or other content produced by perpetrators.

Outcome 6 and 7. Provide safeguards to reduce the risk of harm arising from online misinformation and disinformation

We continually evolve our policies and products to address new challenges and online behaviours. We have adopted a range of measures to reduce the risks of harms arising from misinformation and disinformation including Community Notes, X Premium subscription, updated our policy and communication of our rules against manipulation and spam.

Supporting people's right to accurate and safe political discourse on X

More than half a billion people from around the world gather on X to talk about their interests in real-time, and that includes elections. X enables people to directly engage on important topics with elected representatives, local or national leaders and fellow citizens. During elections, X works to get in front of a range of tactics that people use to target the process. To do this we hire the right people, update our policies and evolve our product²³.

- **Our people:** We are currently expanding our safety and elections teams to focus on combating manipulation, surfacing inauthentic accounts and closely monitoring the platform for emerging threats.
- **Our policies:** We have Rules²⁴ in place to help protect the safety and authenticity of conversations on X. During elections, our *Civic Integrity policy*²⁵ provides an extra layer of protection that is applied for a limited period of time before and during an election.
 - In August 2023, we updated this policy to make sure we strike the right balance between tackling the most harmful types of content—those that could intimidate or deceive people into surrendering their right to participate in a civic process—and not censoring political debate.
 - The policy will also be aligned with our updated enforcement philosophy, Freedom of Speech, Not Reach²⁶. We will add publicly visible labels to posts identified as potentially violating the Civic Integrity Policy, letting people know when their reach has been restricted.

²² <https://help.X.com/en/rules-and-policies/perpetrators-of-violent-attacks>

²³ https://blog.twitter.com/en_us/topics/company/2023/supporting-peoples-right-to-accurate-and-safe-political-discourse-on-x

²⁴ <https://help.twitter.com/en/rules-and-policies/x-rules>

²⁵ <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>

²⁶ https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy



- Building on our commitment to free expression, we are also going to allow political advertising. Starting in the U.S., we will continue to apply specific policies to paid-for promoted political posts²⁷. This will include prohibiting the promotion of false or misleading content, including false or misleading information intended to undermine public confidence in an election, while seeking to preserve free and open political discourse.
- We will also provide a global advertising transparency center so that everyone can review political posts being promoted on X, in addition to robust screening processes to ensure only eligible groups and campaigns are able to advertise.
- **Our product:** We continue to scale Community Notes, an innovative tool that empowers a vetted and growing group of people to add context to posts when they see something that could be wrong, misleading or requires another point of view. X shouldn't determine the truthfulness of disputed information; rather, we should empower our users to express their opinions and openly debate during elections, in line with our commitment to protecting freedom of expression.

Our work is ongoing. These increased investments in people, policy and product will further ensure our communities have access to open, accurate and safe political discourse on X.

Community Notes

Community Notes are one the most important and scalable ways to address and combat misinformation on X²⁸. As Community Notes rapidly evolve on platform, within X, and in public, this product presents a profound shift for our company and people who use our service. It is one of our big bets, grounded in more than a decade and half of X, the platform and content moderation experiments, policies, and products. It is also grounded in ongoing research, evaluation and consultation.

Community Notes is available globally with contributors in 44 countries, and is fully open-sourced. We are already seeing real impact—people are on average 30% less likely to agree with the substance of a post after reading a Community Note about it, and they are also less likely to reshare it. All promoted posts are eligible for Community Notes, including promoted political posts.

We know that misleading information is complex, evolving, and sometimes cloaked behind questions or opinions. To ensure that people are better informed on X, X launched Community Notes, our approach to offering context and surfacing credible information²⁹.

Contributors in New Zealand can leave notes on any post and if enough contributors from different points of view rate that as helpful, the note will be publicly shown on a post. Community Notes *do not* represent X's viewpoint and cannot be edited or modified by our teams. A post with a Community Note will not be labeled, removed, or addressed by X unless it is found to be

²⁷ <https://business.twitter.com/en/help/ads-policies.html>

²⁸ <https://help.X.com/en/using-X/community-notes>

²⁹ *Id.*



violating the X Rules,³⁰ Terms of Service (TOS),³¹ or our Privacy Policy.³² 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 to Community Notes³³ and moved it out of pilot state. 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 are helpful or not.

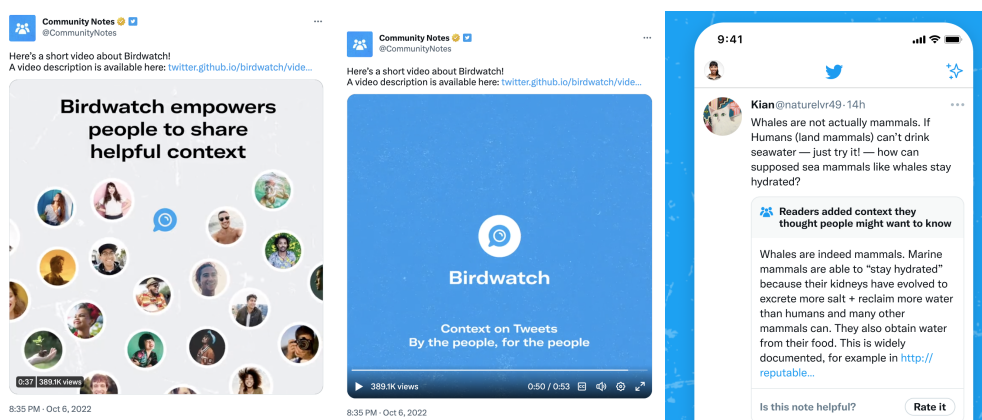


Table 3: Demonstrating how Birdwatch/Community Notes work³⁴

Community Notes were made visible around the world, including in New Zealand³⁵; X started admitting contributors from New Zealand in January 2023³⁶. 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.

Users can now also see notes on posts that are embedded in articles and websites and get more context wherever they are reading posts³⁷.

³⁰ <https://help.X.com/en/rules-and-policies/X-rules.html>

³¹ <https://X.com/en/tos>

³² <https://X.com/en/privacy>

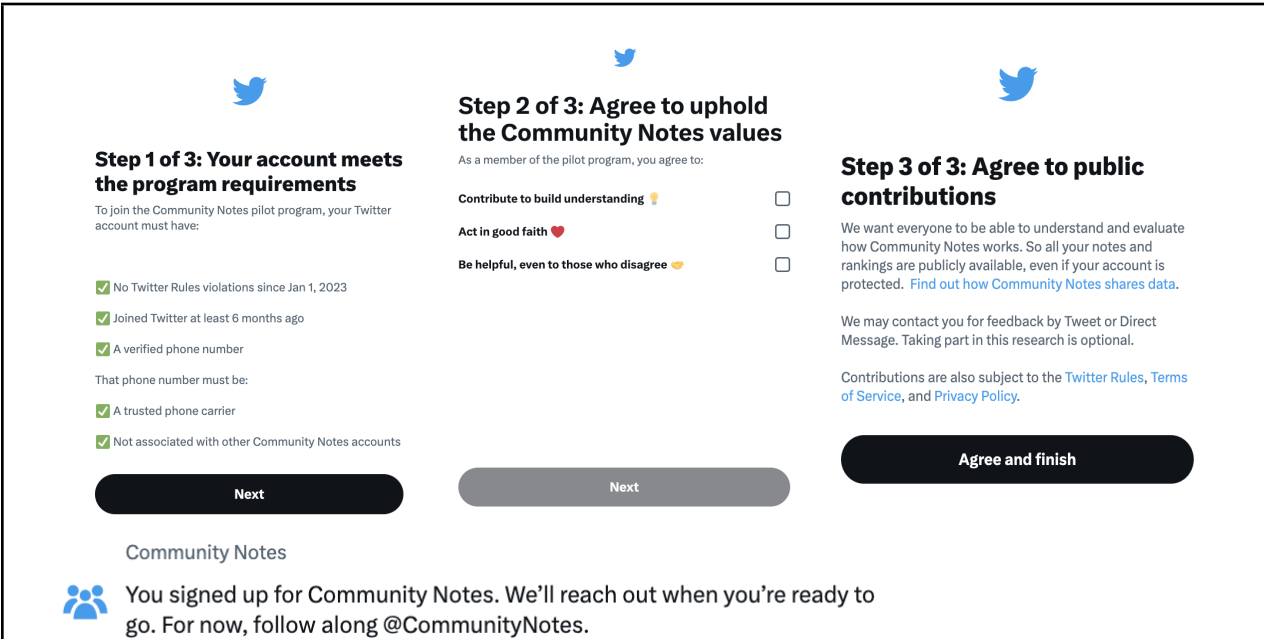
³³ https://blog.X.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-X-us

³⁴ <https://X.com/CommunityNotes/status/1578001121855012864>

³⁵ <https://X.com/CommunityNotes/status/1601753552476438528>

³⁶ <https://X.com/CommunityNotes/status/1616504999844130816>

³⁷ <https://X.com/CommunityNotes/status/1652079596110594049>

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

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

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](#).

Community Notes

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

Table 4: Demonstrating how to sign up for Community Notes

As Community Notes are still evolving globally and in New Zealand. We are seeing a positive trend where more contributors are signing up and it is a promising new start.

Community Notes help users to make more informed choices about the source of news and factual content on X by empowering people on X to collaboratively add context to potentially misleading posts³⁸. It aims to create a better-informed world.

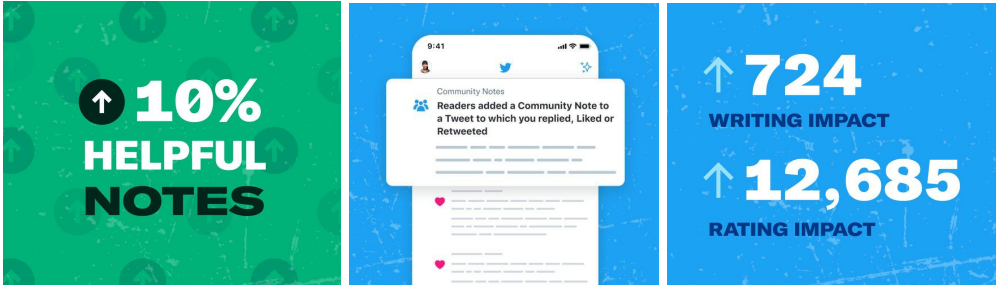


Table 5: Demonstrating how Community Notes help users get well-informed on X

- In April 2023, X 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 precision³⁹. [See Table 5](#)
- In February 2023, users received a heads up if a Community Note starts showing on a post they have replied to, Liked or Reposted. This helps give people extra context that they might otherwise miss⁴⁰. [See Table 5](#)

³⁸ <https://help.X.com/en/using-X/community-notes>

³⁹ <https://X.com/CommunityNotes/status/1649473048947597312>

⁴⁰ <https://X.com/CommunityNotes/status/1628158167006994436>



- 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 contributions⁴¹. [See Table 5](#)

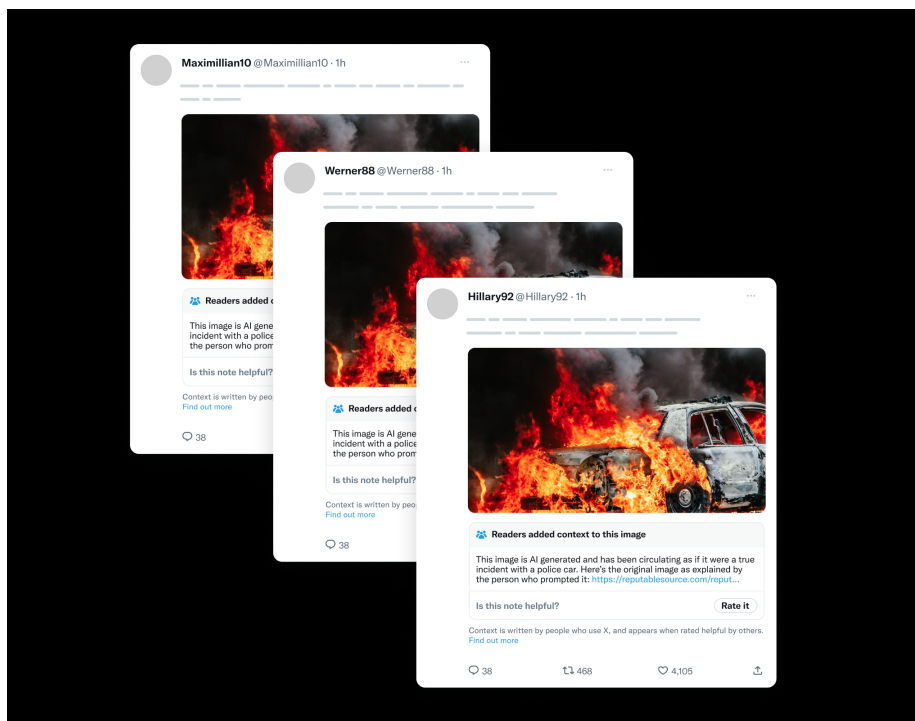


Table 6: Demonstrating how Notes on Media works

- In May 2023, from AI-generated images to manipulated videos, it's common to come across misleading media. We piloted a feature that puts a superpower into contributors' hands: **Notes on Media**. Notes attached to an image will automatically appear on recent and future matching images⁴². In many cases, these Notes can provide valuable context, not just for a single post, but for any post containing the same media⁴³.
 - Tagging notes as “about the image” makes them visible on all posts that our system identifies as containing the same image. These Notes, when deemed Helpful, accumulate view counts from all the posts they appear in, but only count as one Writing and Rating Impact for the author and raters.
 - To ensure transparency, raters will see a disclaimer indicating that notes about the image may appear across multiple posts. This way, they'll be aware that the context provided by these notes extends beyond the specific post they're currently viewing. When someone rates a media note, the rating is associated with the Post on which the Note appeared. This allows Community Notes to identify cases where a note may not apply to a specific post.

⁴¹ <https://X.com/CommunityNotes/status/1616641911502303234>

⁴² <https://x.com/CommunityNotes/status/1663609484051111936>

⁴³ <https://communitynotes.twitter.com/guide/en/contributing/notes-on-media>



- Currently, this feature is experimental and only supports posts with a single image. We're actively working on expanding it to support posts with multiple images, GIFs, and videos.
- In August 2023, we updated that Notes on Media is available for videos and to all Top Writers. Notes written on videos will automatically show on other posts containing matching videos. A highly-scalable way of adding context to edited clips, AI-generated videos, and more⁴⁴.

X Premium and Expanded Verification

X applies visual identity signals like labels and checkmarks on account profiles to provide more context about — and help distinguish — different types of accounts⁴⁵ Some of these indicators are applied by X, while others are triggered by user action. This is one of X's efforts to combat inauthentic accounts, misinformation and disinformation.

- In October 2022, X tested Edit post which went well and made it available to X Premium members in New Zealand among other countries.⁴⁶
- In December 2022 X Premium subscriptions became available in New Zealand⁴⁷. X Premium is one of a range of scalable measures to elevate quality conversations⁴⁸. It is an opt-in, paid subscription. Posts from verified users will be prioritized in places — helping to fight scams and spam.⁴⁹ Once subscribed to X Premium, changes to their 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 period⁵⁰.

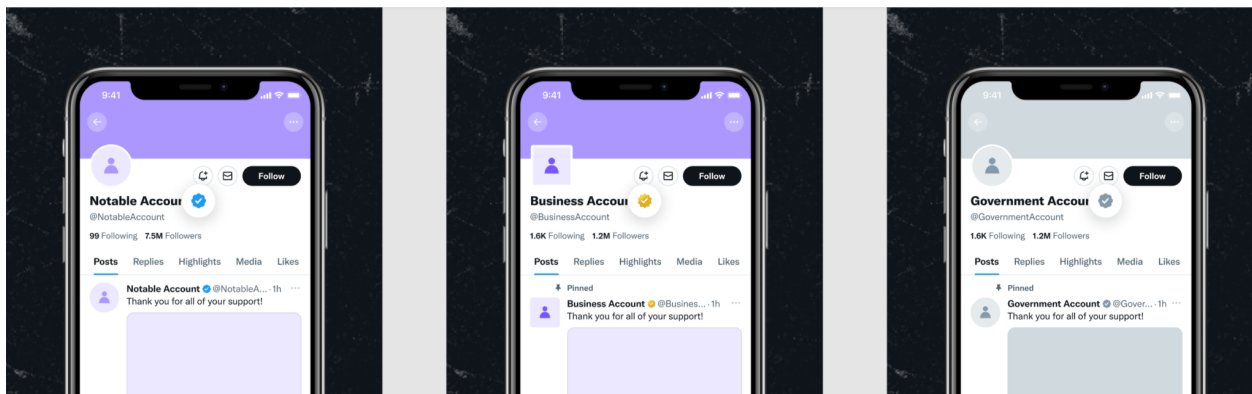


Table 7: Demonstrating X's different profile labels in Blue, Gold and Grey colors

- Blue checkmark: The blue checkmark means that an account has an active subscription to X Premium and meets our eligibility requirements. In April 2023, we removed legacy

⁴⁴ <https://x.com/CommunityNotes/status/1699123989438947373>

⁴⁵ <https://help.X.com/en/rules-and-policies/profile-labels>

⁴⁶ <https://X.com/XBlue/status/1576980429814759424>

⁴⁷ <https://verified.X.com/en>

⁴⁸ <https://help.X.com/en/using-X/X-blue>

⁴⁹ <https://verified.X.com/en>

⁵⁰ <https://help.X.com/en/managing-your-account/about-X-verified-accounts>



verified checkmarks⁵¹ and communicated that to remain verified on X, individuals can sign up for X Premium⁵².

- Gold checkmark and square profile picture: The gold checkmark indicates that the account is an official business account through X Verified Organizations⁵³.
- 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 Organizations⁵⁴.
 - December 2022: X users start seeing additional icons that provide context for accounts on X. In addition to blue and gold checks, X introduced grey checks for government and multilateral accounts and square affiliation badges for select businesses⁵⁵.
 - March 2023: X accepts applications for grey checkmarks for eligible government and multilateral accounts⁵⁶.

Update on X's Rate Limits

In July 2023, we updated our rate limits.⁵⁷ To ensure the authenticity of our user base we must take extreme measures to remove spam and inauthentic accounts from our platform. That's why we temporarily limited usage so we could detect and eliminate inauthentic accounts and other bad actors that are harming the platform. Any advance notice on these actions would have allowed bad actors to alter their behavior to evade detection⁵⁸.

At a high level, we are working to prevent these accounts from 1) scraping people's public Twitter data to build AI models and 2) manipulating people and conversation on the platform in various ways. Currently, the restrictions affect a small percentage of people using the platform. As it relates to our customers, effects on advertising have been minimal. While this work will never be done, we're all deeply committed to making X a better place for everyone.

Spam Reduction in Direct Messages

In July 2023, we added a new messages setting that should help reduce the number of spam messages in DMs. With the new setting enabled, messages from users who you follow will arrive in your primary inbox, and messages from verified users who you don't follow will be sent to your message request inbox. Users who previously had their permissions set to allow message requests from everyone will be migrated to this new setting, but can switch back at any time⁵⁹. As a result, we saw a 70% reduction in spam in Direct Messages compared to the previous week. This work is ongoing, and we will continue to make changes to fight spam to make Twitter better for everyone⁶⁰.

⁵¹ <https://X.com/verified/status/1648764138452299778>

⁵² https://X.com/i/X_blue_sign_up

⁵³ <https://help.X.com/en/rules-and-policies/profile-labels>

⁵⁴ <https://help.X.com/en/rules-and-policies/profile-labels>

⁵⁵ <https://X.com/XSupport/status/1604955466727047168>

⁵⁶ <https://X.com/XSafety/status/1638998382562910208>

⁵⁷ <https://x.com/XBusiness/status/1676278441774292992>

⁵⁸ <https://business.twitter.com/en/blog/update-on-twitters-limited-usage.html>

⁵⁹ <https://twitter.com/Support/status/1679529814212894723>

⁶⁰ <https://x.com/Support/status/1682561887588995074>



4.2 Empower users to have more control and make informed choices

Signatories recognise that users have different needs, tolerances, and sensitivities that inform their experiences and interactions online. Content or behavior that may be appropriate for some will not be appropriate for others, and a single baseline may not adequately satisfy or protect all users. Signatories will therefore empower users to have control and to make informed choices over the content they see and/or their experiences and interactions online. Signatories will also provide tools, programs, resources and/or services that will help users stay safe online.

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

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

We have a range of dedicated tools that are available for all of our users⁶¹ to control their experiences and make informed choices, as well as ways for users to report content that may violate those policies. X has publicly available and accessible robust reporting forms, both in-app 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 appropriate⁶². Users can report posts, Lists, and Direct Messages that are in violation of our Rules or our TOS⁶³.

Our safety and security features, Help Centre pages and FAQs about in-app services⁶⁴ are in place to minimise end-users' exposure to harmful content, empower end-users to manage their safety on X and mitigate the impact on end-users that may arise from the propagation of misinformation and disinformation.

When X takes enforcement actions, we may do so either on a specific piece of content (e.g., an individual post or Direct Message), on an account, or employ a combination of these options. In some instances, this is because the behaviour violates the X Rules. Other times, it may be in response to a valid legal request from an authorised entity in a given country⁶⁵.

⁶¹ <https://help.X.com/en/safety-and-security>

⁶² <https://help.X.com/en>

⁶³ <https://help.X.com/en/safety-and-security/report-a-Post>

⁶⁴ <https://help.X.com/en>

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

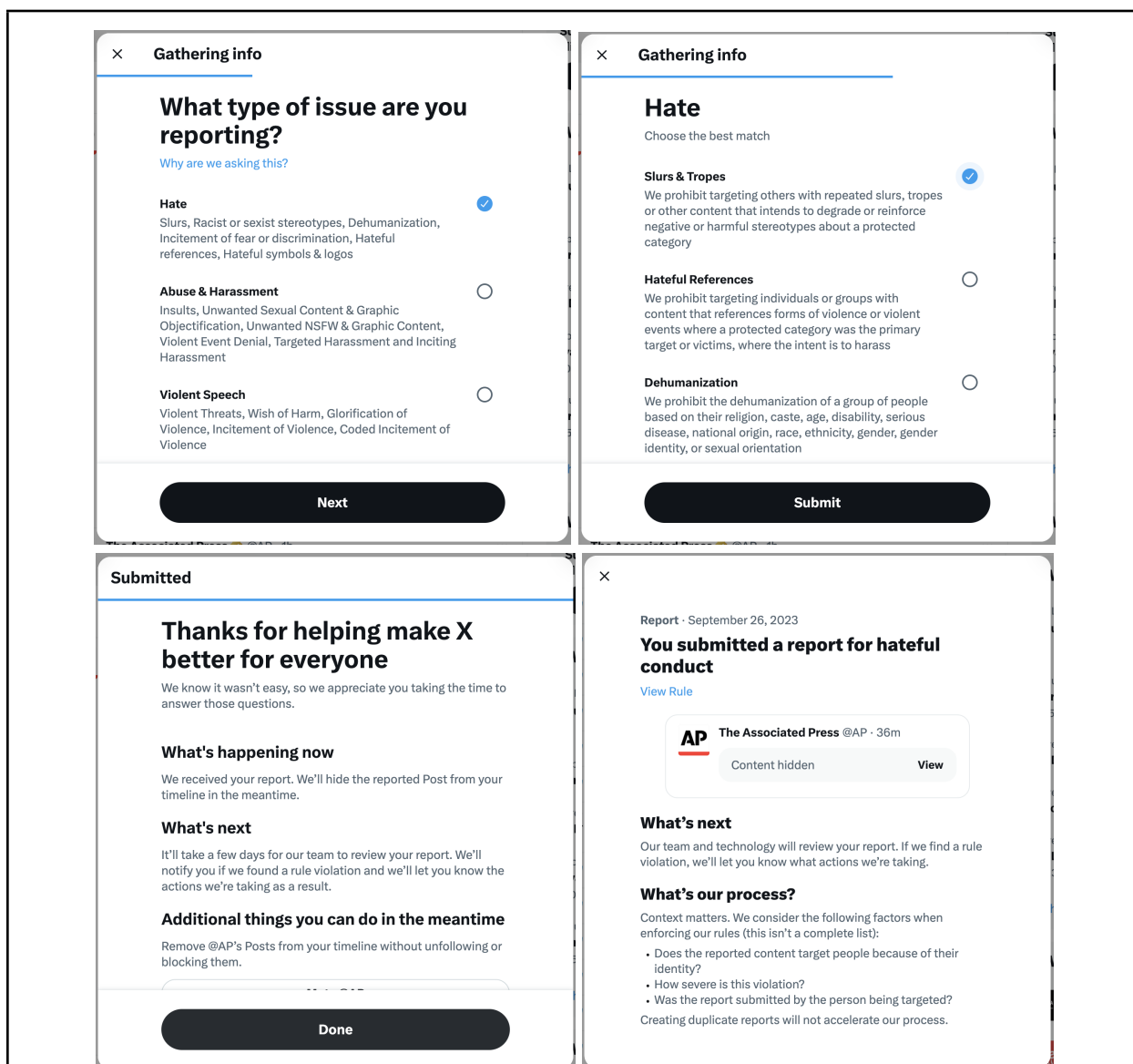


Table 8: Screenshots of how to report and user notifications from X

On X, users have a range of options for controlling their experience especially via the For you and Following timelines⁶⁶, Subscribed accounts, and via Lists⁶⁷.

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

⁶⁶ <https://help.X.com/en/using-X/X-timeline>

⁶⁷ <https://help.X.com/en/using-X/X-lists>



Home

For you

Following

Subscribed

Twitter List

Table 9: controlling their Home (For You, Following, Subscribed and X Lists)

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

When users choose X on the **For you** or **Following** tabs, users will return to whichever timeline you had open last. Users also see content such as promoted posts or Reposts 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 Posts from accounts they do not follow. We recommend posts 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.

X Lists allows users to customize, organize and prioritize the posts users see in their timeline. Users can also choose to join Lists created by others on X, 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 Posts from only the accounts on that List⁶⁸. In Users' Home timeline on X 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 X is our transparency about our recommender systems. X published a blog to introduce how the algorithm selects posts for the user's timeline⁶⁹. Our recommendation system is composed of many interconnected services and jobs. While there are many areas of the app where posts are recommended—Search, Explore, Ads—this post will focus on the home timeline's For you feed. Every day, we serve over 150 billion posts 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 posts posted daily down to a handful of top posts that ultimately show up on their device's For you timeline.

⁶⁸ <https://help.X.com/en/using-X/X-lists>

⁶⁹ https://blog.X.com/engineering/en_us/topics/open-source/2023/X-recommendation-algorithm



4.3 Enhance transparency of policies, processes and systems

Transparency helps build trust and facilitates accountability. Signatories will provide transparency of their policies, processes and systems for online safety and content moderation and their effectiveness to mitigate risks to users. Signatories, however, recognise that there is a need to balance public transparency of measures taken under the Code with risks that may outweigh the benefit of transparency, such as protecting people’s privacy, protecting trade secrets and not providing threat actors with information that may expose how they may circumvent or bypass enforcement protocols or systems.

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

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

Transparency is fundamental to everything we do at X.

In March 2023, X announced a new era of transparency opening much of our source code to the global community. On GitHub, users can find two new repositories containing the source code for many parts of X, including our recommendations algorithm, which controls the posts users see on the For you timeline.

- We shared more information on our recommendation algorithm in this post on our Engineering Blog⁷⁰. 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 X algorithm at this point.

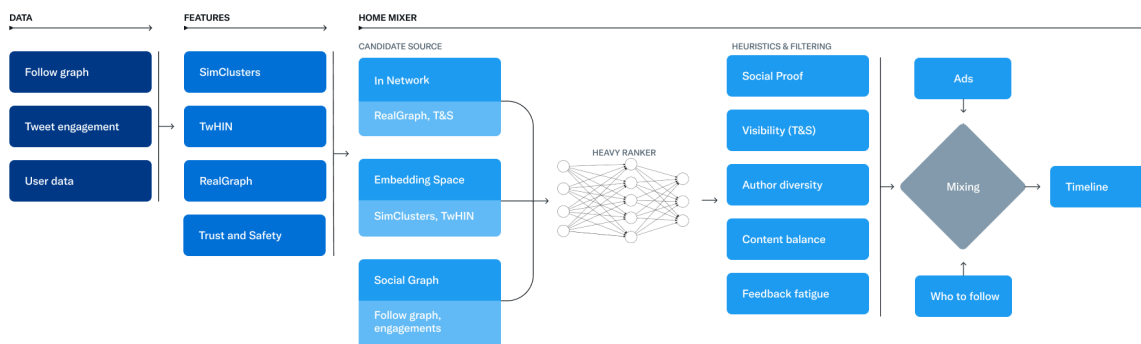


Table 10: Demonstrating the major components used to construct a timeline

⁷⁰ *Id.*



X has several Candidate Sources that we use to retrieve recent and relevant posts for a user. For each request, we attempt to extract the best 1500 posts 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 Posts and 50% Out-of-Network Posts on average, though this may vary from user to user.

The goal of the For you timeline is to serve users relevant posts. At this point in the pipeline, we have ~1500 candidates that may be relevant. Scoring directly predicts the relevance of each candidate post and is the primary signal for ranking posts 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 post interactions to optimize for positive engagement (e.g. Likes, Reposts, and Replies). This ranking mechanism takes into account thousands of features and outputs ten labels to give each post a score, where each label represents the probability of an engagement. We rank the posts from these scores.

At this point, Home Mixer has a set of posts ready to send to the user's device. As the last step in the process, the system blends together Posts with other non-Post 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

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 that users can view to understand our algorithm in greater detail⁷¹, 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 X analytics platform for creators with more information on reach and engagement, Greater transparency into any safety labels applied to their Posts or accounts, and Greater visibility into why posts appear on users' timeline.

In April 2023, we published our 21st report, with data on our policy enforcement for the first half of 2022⁷². As we review our approach to transparency reporting in light of innovations in content moderation and changes in the regulatory landscape, we believe it's important to share data from H1 2022 on our trust and safety efforts.

⁷¹ <https://github.com/X/the-algorithm> and <https://github.com/X/the-algorithm-ml>

⁷² <https://X.com/Safety/status/1650952198451499008>



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 11: showing X's transparency report between 1 January - June 30, 2022

X continues to take action on content that violates our Rules⁷³ and protects users' rights in response to government legal requests. Over the reporting period, X required users to remove 6,586,109 pieces of content that violated the X 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 X Rules (a 28% increase).

Around the world, X received approximately 53,000 legal requests to remove content from governments during the reporting period. X's compliance rate for these requests varied by requester country. The top requesting countries were Japan, South Korea, Turkey and India.

X received over 16,000 government information requests for user data from over 85 countries during the reporting period. Disclosure rates vary by requester country. The top five requesting countries seeking account information in H1 2022 were India, the United States, France, Japan, and Germany. We intend to share more about our path forward for transparency reporting later this year. In the meantime, we will continue to give you insights into X's work to promote entertaining, informative and healthy conversation⁷⁴.

4.4 Support independent research and evaluation

⁷³ <https://help.X.com/en/rules-and-policies/x-rules>

⁷⁴ https://blog.X.com/en_us/topics/company/2023/an-update-on-X-transparency-reporting



Independent local, regional or global research by academics and other experts to understand the impact of safety interventions and harmful content on society, as well as research on new content moderation and other technologies that may enhance safety and reduce harmful content online, are important for continuous improvement of safeguarding the digital ecosystem. Signatories will seek to support or participate in these research efforts. Signatories may also seek to support independent evaluation of the systems, policies and processes they have implemented under the commitments of the Code. This may include broader initiatives undertaken at the regional or global level, such as independent evaluations of Signatories' systems.

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.

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

Over the years, hundreds of millions of people have sent over a trillion posts, with billions more every week. X data are among the world's most powerful data sets. The new Free, Basic and Enterprise Tiers of Access to the X API are available to independent researchers. We are 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 X Developer Platform⁷⁵ and continued investment, especially the X API. By March 30th we began the relaunch with announcements - as with earlier updates - made also via @XDev.

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

X is committed to the success of our Developer ecosystem. We will continue to build on these efforts and inform the public as we improve X in the open. Below are recent notable efforts that X has supported independent researchers to improve public understanding.

Community Notes transparency: X made the Community Notes algorithm publicly available on GitHub, along with the data that powers it, so anyone can audit, analyze, or suggest improvements.⁷⁶

We have 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 company⁷⁷. This change optimized a function that identifies explanatory tags that describe

⁷⁵ <https://X.com/XDev/status/1603823066496147456>

⁷⁶ <https://X.com/CommunityNotes/status/1578004584320172034>

⁷⁷ <https://X.com/CommunityNotes/status/1629229535337058305>



why raters found a note helpful or not. X welcomed improvements like this, and would love to see contributions that strengthen note quality, adversarial resistance or other core elements of the system⁷⁸

Independent assessment of hate speech on X with Sprinklr⁷⁹: In March 2023, we reported the recent partnership with Sprinklr⁸⁰ for an independent assessment of hate speech on X⁸¹. Sprinklr’s AI-powered model found that the reach of hate speech on X is even lower than our own model quantified⁸². X announced the finding and Sprinklr’s report is publicly available⁸³.

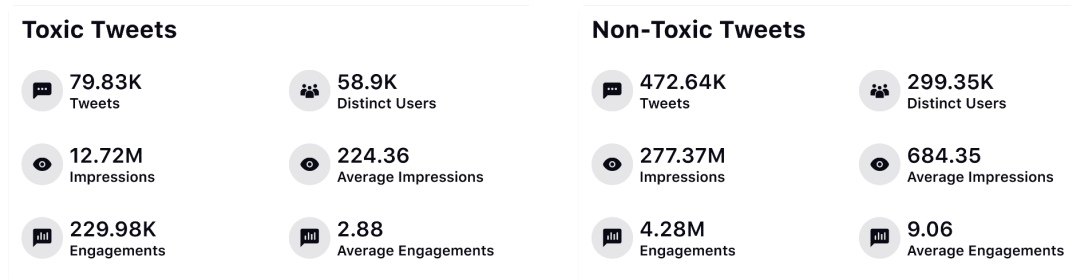


Table 12: Comparing Toxic Posts with Non-Toxic Posts

Open-source algorithm: when compared to non-toxic Posts in the dataset containing slur keywords, toxic posts received 3 times fewer impressions on average.

X is constantly experimenting with the open-source algorithms that select which notes to show. So everyone can follow along, they can easily see in “Note Details” which model computed the current status of a note⁸⁴. In April 2023, X shared another update on the changes we have made to our open source repos this week and a preview of what’s next⁸⁵.

⁷⁸ <https://X.com/CommunityNotes/status/1629229897691324416>

⁷⁹ <https://X.com/XSafety/status/1638255718540165121>

⁸⁰ <https://X.com/Sprinklr>

⁸¹ <https://partners.X.com/en/partners/sprinklr>

⁸² <https://X.com/XSafety/status/1638262108650348545>

⁸³ <https://www.sprinklr.com/blog/identify-toxic-content-with-leading-analytical-ai/>

⁸⁴ <https://X.com/CommunityNotes/status/1636852370809257984>

⁸⁵ <https://X.com/XEng/status/1652049665184137216>