



Annual Compliance Report 2024

Aotearoa New Zealand Code of Practice for Online Safety and Harms

Signatory: X Corp

If applicable: Relevant Products / Services

At X, 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.

As a signatory to the Aotearoa New Zealand Code of Practice for Online Safety and Harms (the “**Code**”), this report outlines our overall approach to online safety on X, as well as the measures and activities to combat harmful and inappropriate content during the period from 1 October 2023 to 30 September 2024, for New Zealand.

Our Approach to Online Safety

X's purpose is to serve the public conversation. In line with our mission to promote open conversation, we encourage a variety of perspectives on the platform. This is central to our [Freedom of Speech, Not Reach philosophy](#), that moves us away from a binary take down/leave up approach to a more reasonable, proportionate and effective moderation process. Violence, harassment, and other similar types of behaviour discourage people from expressing themselves, and ultimately diminish the value of global public conversation. We thus have clear [Rules and Policies](#) in place that are designed to ensure all people can participate in the public conversation freely and safely. They apply globally, including to New Zealand, and are easily accessible on our [Help Center](#).

Our [Rules and Policies](#) are dynamic, and we continually review them to ensure that they are up-to-date, necessary and proportionate. Creating a new policy or making a policy change requires in-depth research around trends in online behaviour, developing clear external language that sets expectations around what's allowed, and creating enforcement guidance for reviewers that can be scaled across millions of pieces of content and accounts.

While we aim to enable open discussion of differing opinions and viewpoints, we are committed to the objective, timely, and consistent enforcement of our Rules. To enforce our [Rules](#), we use a combination of machine learning and human review.

Our content moderation systems are designed and tailored to mitigate potential harms without unnecessarily restricting the use of our platform and fundamental rights, especially freedom of expression. Content moderation activities are implemented and anchored on principled policies and leverage a diverse set of interventions to ensure that our actions are reasonable, proportionate and effective. Our content moderation systems blend automated and human review paired with a robust appeals system that enables our users to quickly raise potential moderation anomalies or mistakes. This work is led by an international, cross-functional team with 24-hour coverage and the ability to cover multiple languages. These moderation activities are supplemented by scaled human investigations into the tactics, techniques and procedures that bad actors use to circumvent our rules and policies.

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



X strives to provide an environment where people can feel free to express themselves. If abusive behaviour happens, we want to make it easy for people to report it to us. When we take [enforcement actions](#), we may do so either on a specific piece of content (e.g., an individual post or Direct Message) or on an account. We may employ a combination of enforcement actions against prohibited behaviour, in line with the X Rules.

We always aim to exercise moderation with transparency. Where our systems or teams take action against content or an account as a result of violating our Rules or in response to a valid and properly scoped request from an authorised entity in a given country, we strive to provide context to users.

X is a place for users to share ideas and information, connect with communities, and see the world around them. In order to protect the very best parts of that experience, we provide [tools](#) designed to help users control what they see and what others can see about them, so that they can express themselves on X with confidence. Our diverse product-level safety features allow users to modify their experience and engagement on X to ensure each user is able to participate on the platform in a safe and meaningful way.

For further details on how we're making X safer please refer to the following: <https://help.x.com/en/resources/a-safer-twitter>.

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 of the Code.

X has clear [Rules and Policies](#) in place that are designed to ensure all users on X can participate in the public conversation freely and safely. They apply globally, including to New Zealand, and are easily accessible on our [Help Center](#). All users are required by our [Terms of Service](#) to use X in compliance with these Rules and Policies.

Our mission is to empower people to express their opinions and beliefs without barriers, and we recognize that experiencing online abuse jeopardises an individual's ability to freely participate in the public conversation. Therefore, we have detailed rules related to [violence](#), [harassment](#) and other similar types of behaviour that discourage people from expressing themselves, and ultimately diminish the value of global public conversation. [Our rules](#) are to ensure all people can participate in the public conversation freely and safely.

Our approach to moderation has remained consistent during the reporting period. X employs a combination of heuristics and machine learning algorithms to automatically detect content that violates the [X Rules and Policies](#). We use combinations of natural language processing models, image processing models and other sophisticated machine learning methods to detect potentially violative content. These models vary in complexity and in the outputs they produce.

For example, the model used to detect abuse on the platform is trained on abuse violations detected in the past. Content flagged by these machine learning models are either reviewed by human content reviewers before an action is taken or, in some cases, automatically actioned based on model output.

Heuristics are typically utilised to enable X to react quickly to new forms of violations that



emerge on the platform. Heuristics are common patterns of text or keywords that may be typical of a certain category of violations. Pieces of content detected by heuristics may also get reviewed by human content reviewers before an action is taken on the content. These heuristics are used to flag content for review by human agents and prioritise the order such content is reviewed.

Our Safety Team works every day to make this platform a better, safer space for everyone – users, partners, and clients alike. In April 2024, we updated our X Safety leadership². Our Safety team is a vital component of X's success, and we are committed to investing in and growing this team.

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

We recognise that minors are a more vulnerable group by virtue of their age. Users below the age of 13 are not permitted to sign up for the service, as stipulated by our Terms of Service. Users who do not meet our age requirements have their account locked. In addition, parents and guardians are able to access our [Rules and Policies](#) to learn more about how to keep their child's account and experience on X safe, secure and welcoming. This includes a form permitting them to [report](#) accounts holders who they suspect as being underage.

Nevertheless, although minors represent a minimal fraction of X's user base, we are fully committed to the protection of this group. We have a number of specific tools and policies to protect younger audiences on our platform, in addition to the suite of tools ([here](#) and [here](#)) which are designed to help our users control what they see on X and what others can see about them on X, so that they can express themselves on X with confidence. Accounts belonging to known minors will be defaulted to "Protected posts". This means that known minors will receive a request when new people want to follow them (which they can approve or deny), that their posts will only be visible to their followers, and that their posts will only be searchable by them and their followers (i.e. they will not appear in public searches). Accounts belonging to known minors will be restricted to receiving DMs from accounts they follow by default. Post location is off by default, and users need to opt in to the service. More details are provided [here](#).

X policies and tools apply uniformly across all of our users and operate to minimise end users' exposure to harmful content.

In **May 2024**, X consolidated all child safety related policies—Child Sexual Exploitation, Physical Child Abuse Media, and Media of Minors in Physical Altercations³—under a new standalone [Child Safety](#) umbrella policy. This structural change did not lead to any enforcement changes and was primarily meant to improve user transparency with enhanced and distinct Help Center articles.

We encourage our users to come to X to share stories, raise awareness, and speak their mind - including calling attention to the exploitation of children and minors. However, even when shared with the intent to bring awareness or justice, to express outrage or sharing content in a humoristic context, posting media of children experiencing sexual abuse or certain types of physical abuse can contribute to their revictimization and may even lead to the normalization of sexual or physical violence against children. When this content is shared on X, we may remove it even if it was shared with good intent. Our priority is to

² <https://x.com/Safety/status/1775202997284282543>

³ Physical Child Abuse Media and Media of Minors in Physical Altercations were previously subpolicies under the Sensitive Media policy umbrella.



protect minors from physical and psychological harm irrespective of the context in which the content may be shared.

Child Sexual Exploitation

X has zero tolerance towards any material that features or promotes child sexual exploitation. This may include real media, text, illustrated, or computer-generated media - including generative AI media. Regardless of the intent, anyone viewing, sharing, linking, or engaging with any kind of child sexual exploitation material contributes to the re-victimization of the depicted children and puts children at an extreme risk of being harmed. This also applies to content that may further contribute to victimisation of children through the promotion or glorification of child sexual exploitation.

Physical Child Abuse

We will remove most instances of media depicting physical child abuse, even if shared to raise awareness or express outrage in order to prevent revictimization or normalization of violence against children. When assessing the best course of action to take, we consider: whether the child is nude, partially clothed, or fully clothed; the severity of harm the child is experiencing; and whether the media was shared with abusive, non-abusive, or newsworthy context.

Media of Minors in Physical Altercation

We aim to protect the wellbeing of minors involved in physical altercation, regardless of their status (i.e victim or aggressor), while balancing the need to raise awareness of these issues. When assessing the best course of action to take, we may consider whether: the content is shared with an abusive or non-abusive context; we have received a report from the minor or an authorised representative; and the content is excessively graphic.

How to Report

Anyone can report violations of this policy using our dedicated in-app reporting flow or via our Help Centre⁴. An X account is not needed to report potential violations. In the majority of cases, the consequences for violating our CSE policy is immediate and permanent suspension from the platform. Additionally, when we're made aware of content depicting or promoting child sexual exploitation, including links to third party sites where this content can be accessed, the accounts sharing this content will be reported to the National Center for Missing & Exploited Children (NCMEC) and, for certain violations, we may instead require post removal. For example, we may ask someone to remove the violating content. Subsequent violations may lead to account suspension.

Protection of minors

X collects each account holder's date of birth through the neutral presentation of a date of birth prompt. Users must be over the age of 13 to use the site and will be automatically off-boarded if their date of birth indicates that they are under 13⁵.

Account holders who enter a date of birth that makes them under the age of 18 will not be permitted to see known adult content on X. Furthermore, once these accounts enter a date of birth that makes them under the age of 18, they will be stopped from re-entering a new date of birth.

X also prohibits knowingly marketing or advertising certain products and services to minors as detailed in the following policy: [Prohibited advertising content for minors](#). This policy applies to monetization on X and X's paid advertising products and advertisements containing age-inappropriate content will be tagged as "not family safe" and will be restricted from being shown to users under the age of 21 and signed-out users. If an account does not

⁴ <https://help.x.com/en/forms/safety-and-sensitive-content/cse>

⁵ <https://x.com/en/tos>



have a date of birth associated with it, X infers the user's age based on their interactions with the site.

Partnerships

X is committed to utilising the full extent of our capabilities to eradicate child sexual exploitation. We have also strengthened our enforcement with more tools, and technology to prevent bad actors from distributing, searching for, or engaging with CSE content across all forms of media. Along with taking action under our Rules, we're proud to partner with NCMEC to keep children safe⁶. We send a mix of automated and human reviewed reports to NCMEC. We will continue using the full extent of our capabilities to eliminate these harmful practices⁷.

Another key partnership is our work with Thorn. Through our ongoing partnership with Thorn, X is doing more to create a safe platform. X was involved in testing Thorn's solution during its beta phase to proactively detect text-based child sexual exploitation. This work builds on our relentless efforts to combat child sexual exploitation online, with the specific goal of expanding our capabilities in fighting high-harm content where a child is at imminent risk⁸. Learn more about Thorn's solutions for platforms to mitigate risk and scale child safety⁹

Our work to stop child sexual exploitation online remains a top priority for X. We are steadfast in continuing this important work and committed to deepening partnerships that allow us to continually improve our approach to safeguarding our platform.

Enforcement

In January 2024, we updated our work our work to tackle Child Sexual Exploitation on X¹⁰. In 2023, as a result of our investment in additional tools and technology to combat CSE, X suspended 12.4 million accounts for violating our CSE policies. This is up from 2.3 million accounts in 2022. Along with taking action under our rules, we also work closely with NCMEC. In 2023, X sent 850,000 reports to NCMEC, including our first ever fully-automated report, over eight times more than Twitter sent in 2022.

Not only are we detecting more bad actors faster, we're also building new defenses that proactively reduce the discoverability of posts that contain this type of content. One such measure that we have recently implemented has reduced the number of successful searches for known Child Sexual Abuse Material (CSAM) patterns by over 99% since December 2022.

Advanced technology and proactive monitoring

We are investing in products and people to bolster our ability to detect and action more content and accounts, and are actively evaluating advanced technologies from third-party developers that can enhance our capabilities. Some highlights include:

- **Automated NCMEC reporting:** In February 2023, we sent our first ever fully-automated NCMEC CyberTipline report. Historically, every NCMEC report was manually reviewed and created by an agent. Through our media hash matching with Thorn, we now automatically suspend, deactivate, and report to NCMEC in minutes without human involvement. This has allowed us to submit over 50,000 automated NCMEC reports in the past year.
- **Expanded Hash Matching to Videos and GIFs:** For the first time ever, we are

⁶ <https://x.com/Safety/status/1776012391597162574>

⁷ <https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf>

⁸ <https://x.com/Safety/status/1787970979706081621>

⁹ <https://www.thorn.org/solutions/for-platforms>

¹⁰ https://blog.x.com/en_us/topics/company/2023/an-update-on-our-work-to-tackle-child-sexual-exploitation-on-x



evaluating all videos and GIFs posted on X for CSAM. Since launching this new approach in July 2023, we have matched over 70,000 pieces of media.

- **Launched Search Intervention for CSE Keywords:** CSAM impressions occur more on search than on any other product surface. In December 2022, we launched the ability to entirely block search results for certain terms. We have since added more than 2,500 CSE keywords and phrases to this list to prevent users from searching for common CSE terms.

We break down our enforcement actions regarding Child Safety policy over a six-month period from January to June 2024. It is further broken down into the specific actions taken, and if they were performed through automated or human review¹¹.

CHILD SAFETY			
At X, we have zero tolerance for child sexual exploitation and are committed to removing media that depicts physical child abuse. We also suspend users who engage with that content to prevent the normalization of violence against children. Along with taking action under our Rules, we also work closely with the National Center for Missing and Exploited Children (NCMEC).	We send a mix of automated and human reviewed reports to NCMEC. We will never stop using the full extent of our capabilities to eliminate these harmful practices.		
	Read our comprehensive policy that outlines how we make our platform inhospitable for those who seek to exploit minors in any way here .		
	TOTAL ACTIONS	AUTOMATED	HUMAN
NCMEC Reports	370,588	35,176	335,412
Accounts Suspended ¹	2,781,634	2,388,683	392,951
Content Removed	14,571	1,645	12,926

¹ Accounts suspended for engaging with (liking, replying, bookmarking, etc.) Child Sexual Abuse Media (CSAM) are not reportable to NCMEC

Table 1 provides a breakdown of enforcement actions on CSE from January to June 2024

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

With reference to the specific categories of **Outcomes 2-5** of the Code, the following X Rules and policies operate to provide safeguards to reduce the risk of harm from such harmful contents:

¹¹ <https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf>



Sexual Content:

- **Child Safety**¹² updated in May 2024 and as discussed in **Outcome 1**
- **Non-consensual nudity/Private Content**¹³
- **Violent Content**¹⁴ (which covers content including that which depicts sexual violence) updated in May 2024. Healthy conversations can't thrive when Violent Speech is used to deliver a message, and not every participant wishes to be exposed to Violent Media. As a result, we may remove or reduce the visibility of Violent Content to ensure the safety of our users and prevent the normalisation or glorification of violent actions. We also do not allow sharing Violent Content in highly visible places such as profile photos, banners or bio.

Violent Content:

- **Violent Content**
- **Violent & Hateful Entities**¹⁵
- **Perpetrators of Violent Attacks**¹⁶

Cyberbullying Content:

- **Abuse and Harassment**¹⁷ updated in March 2024. To facilitate healthy dialogue on the platform, and empower individuals to express diverse opinions and beliefs, we prohibit behaviour and content that harasses, shames, or degrades others. In addition to posing risks to people's safety, these types of behaviour may also lead to physical and emotional hardship for those affected.
- **Hateful conduct**¹⁸ We are committed to combating abuse motivated by hatred, prejudice, or intolerance. For this reason, we prohibit behaviour that targets individuals or groups with abuse based on their perceived membership in a protected category¹⁹.
- **Private Content**²⁰ updated in March 2024. X's Rules strictly prohibit the publication or posting of other people's private information without their express authorization and permission. This includes contexts suggesting abusive intent, harassment or incitement to harass, as well as the distribution of media that may lead to emotional or physical harm. We further clarified our external policy to explicitly include provisions for user anonymity reinforcing X's commitment to maintaining a safe and secure platform²¹.

Suicide and Self-Harm Content:

- **Suicide and Self-harm**²²

The X platform boldly champions the vital principles of free speech and community safety. In a world where these values are constantly challenged, we proudly support organisations like Netsafe New Zealand, who tirelessly work to protect global freedoms of speech, information, and press, as well as safeguard our digital community. We stand united in our commitment to fostering an environment where expression and safety are not mutually exclusive, but rather, are the cornerstones of a thriving society²³.

Crisis response

¹² <https://help.x.com/en/rules-and-policies/child-safety>

¹³ <https://help.x.com/en/rules-and-policies/personal-information>

¹⁴ <https://help.x.com/en/rules-and-policies/violent-content>

¹⁵ <https://help.x.com/en/rules-and-policies/violent-entities>

¹⁶ <https://help.x.com/en/rules-and-policies/perpetrators-of-violent-attacks>

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

¹⁸ <https://help.x.com/en/rules-and-policies/hateful-conduct-policy>

¹⁹ <https://x.com/Safety/status/1715077166025613463>

²⁰ <https://help.x.com/en/rules-and-policies/personal-information>

²¹ <https://x.com/Safety/status/1770647182279921840>

²² <https://help.x.com/en/rules-and-policies/glorifying-self-harm>

²³ <https://x.com/Safety/status/1734676893348069787>



We may also activate a crisis protocol to address any rapidly evolving crisis or conflict situation with the highest level of priority. That includes the formation of a cross-functional leadership team that ensures our global community has access to real-time information and to safeguard the platform for our users and partners. One example that highlights the work X has done in the past one year is provided here²⁴.

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

We have policies in place that reduce the risk of harm arising from online misinformation and disinformation²⁵. Here is an overview of X's key policies, which was also shared in detail in our 2023 annual report, for easy reference:

- **Misleading and deceptive identity policy**²⁶ People on X may not misappropriate the identity of individuals, groups, or organisations or use a fake identity to deceive others.
- **Platform manipulation and spam policy**²⁷ People on X may not use X's services in a manner intended to artificially amplify or suppress information or engage in behaviour that manipulates or disrupts people's experience or platform manipulation defences on X.
 - In **April 2024**, X kicked off a significant, proactive initiative to eliminate accounts that violate our Rules against platform manipulation and spam. While we aim for accuracy in the accounts we remove, we are casting a wide net to ensure X remains secure and free of bots. As a result, users may have observed changes in follower counts²⁸.
- **Synthetic and manipulated media policy**²⁹ People on X 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 posts containing misleading media to help people understand their authenticity and to provide additional context.
- **Civic integrity policy**³⁰ People on X may not use X's services for the purpose of manipulating or interfering in elections or other civic processes, such as posting or sharing content that may suppress participation, mislead people about when, where, or how to participate in a civic process, or lead to offline violence during an election. Any attempt to undermine the integrity of civic participation undermines our core tenets of freedom of expression and as a result, we will apply labels to violative posts informing users that the content is misleading.

We continue to strengthen our service by building new defences such as improving our auto-detection technology against attempted manipulation, which includes malicious accounts created via automation or mass registration, spam, as well as other activities that violate our Terms of Service³¹.

When we identify suspicious activity, we require an individual using the service to confirm human control of the account or their identity.³² For example, we may require the account holder to verify a phone number or email address, do a password reset or complete a captcha-based test. While these challenges are simple for authentic account owners to solve, they are difficult (or costly) for spammy or malicious accounts to complete. Accounts

²⁴ <https://x.com/Safety/status/1843345200787202137>

²⁵ <https://help.twitter.com/en/rules-and-policies/enforcement-options>

²⁶ <https://help.X.com/en/rules-and-policies/X-impersonation-and-deceptive-identities-policy>

²⁷ <https://help.twitter.com/en/rules-and-policies/platform-manipulation>

²⁸ <https://x.com/Safety/status/1775942160509989256>

²⁹ <https://help.twitter.com/en/rules-and-policies/manipulated-media>

³⁰ <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>

³¹ <https://twitter.com/en/tos>

³² <https://help.twitter.com/en/rules-and-policies/enforcement-options>



which fail to complete a challenge within a specified period of time may be suspended. We have also implemented mandatory email or phone verification for all new accounts.

About profile labels and checkmarks on X

At X, we use visual identity signals to give our community more context about the accounts they see and engage with on our platform. These include checkmarks and labels that distinguish different types of accounts, like the grey checkmark denoting government or multilateral organisations or officials and labels on automated or “bot” accounts³³. These labels support our core value of transparency, giving users valuable information about the source of the content they see on our platform³⁴.

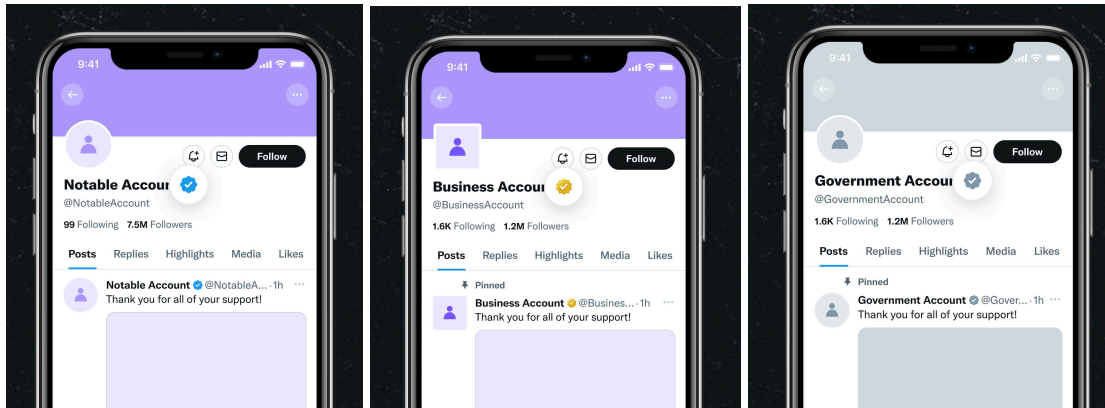


Table 2 provides screenshots for Blue, Gold and Grey checkmarks

X-applied Blue checkmark

X’s [help centre page](#) sets out our current process for applying a blue checkmark as part of a X Premium subscription. The blue checkmark means that an account has an active subscription to X Premium and meets our eligibility requirements. The specific eligibility criteria for individuals to receive a blue checkmark as part of a X Premium subscription are detailed [here](#). These accounts may represent an individual or an organisation.

Verified Organisations (Gold and grey checkmarks)

[Verified Organizations](#) is a subscription for businesses, governments and nonprofits on X which comes with a gold or grey checkmark, affiliation badges, premium support, impersonation defence and more features for organisations.

The gold checkmark indicates that the account is an official business account subscribed to X’s Verified Organizations.

A grey checkmark indicates that an account is an official government account subscribed to X’s Verified Organisations.

In addition to government organisations subscribing through Verified Organizations, X has a complimentary option for select government/multilateral organisations and officials to obtain a grey checkmark. This verification route comes without features and benefits associated with Verified Organizations and is reserved for high level government/multilateral officials and organisations.

³³ <https://help.x.com/en/rules-and-policies/profile-labels>

³⁴ <https://x.com/Safety/status/1715435020926464510>



Eligible government organisations at the national level may include: Main executive office accounts, agency accounts overseeing specific areas of policy, main embassy and consulate accounts, and parliamentary or equivalent institutional and committee accounts.

Eligible government organisations at the state and local level include: Main executive office accounts and main agency accounts overseeing crisis response, public safety, law enforcement, and regulatory issues.

Eligible government individuals may include: Heads of state (presidents, monarchs and prime ministers), deputy heads of state (vice presidents, deputy prime ministers), national-level cabinet members or equivalent, the main official spokesperson for the executive branch or equivalent, and individual members of all chambers of the supranational or national congress, parliament, or equivalent.

Eligible multilateral organisations may include: the main headquarters-level, regional-level, and country-level institutional accounts.

Eligible multilateral individuals include: The head and deputy-head or equivalent of the multilateral organisation.

More information on profile labels is available [here](#).

Any government or multilateral accounts that do not qualify under our current grey checkmark criteria can see if they're eligible under our Verified Organizations feature³⁵.

Affiliation badges

Through Verified Organizations, organisations can affiliate other accounts with their account. Affiliated accounts receive a label with the image from the organisation's profile picture.

Automated account labels

[Automated labels](#) provide transparency by helping you identify if an account is a bot or not. When an account displays the "automated" account label you know the account is generating automated content not produced by a human. Automated account labels — currently in testing — appear on account profiles under profile names and handles.

Self-selected

Professional category labels are selected by people on X when they convert to a Professional Account. X does not control the selection of these labels, and users may change their professional category at any time.

Community Notes

Community Notes is a collaborative way to add helpful context to posts and keep people better informed as it aims to create a better-informed world, by empowering people on X to collaboratively add helpful notes to posts that might be misleading. It is one the most important and scalable ways to address and combat misinformation on X³⁶.

- **Contributors write and rate notes.** Contributors are people on X, who sign up to write and rate notes³⁷. The more people that participate, the better the program becomes.

³⁵ <https://help.x.com/en/using-x/verified-organizations>

³⁶ <https://help.X.com/en/using-X/community-notes>

³⁷ <https://communitynotes.x.com/guide/en/contributing/signing-up>



- **Only notes rated helpful by people from diverse perspectives appear on posts.** Community Notes do not work by majority rules. To identify notes that are helpful to a wide range of people, notes require agreement between contributors who have sometimes disagreed in their past ratings. This helps prevent one-sided ratings³⁸. Learn more about how Community Notes handles diverse perspectives.
- **X doesn't choose what shows up, the people do.** X doesn't write, rate or moderate notes (unless they break the X Rules³⁹.) We believe giving people a voice to make these choices together is a fair and effective way to add information that helps people stay better informed.
- **Open-source and transparent** It's important for people to understand how Community Notes works, and to be able to help shape it. The program is built on transparency: all contributions are published daily, and our ranking algorithm can be inspected by anyone⁴⁰.

Community Notes now empower over 750,000 contributors in 197 countries, including from New Zealand to add helpful context to posts on X, including ads.

Below are a few notable updates about Community Notes we have made since our Annual Report 2023:

- In **September 2024**, X admitted Community Notes contributors in 197 countries and territories around the world. With this expansion, we are now admitting from all countries in which we have supported phone carriers needed for signup⁴¹.
 - At the time of the update and since the launch of Community Notes, X had received over one million note requests⁴².

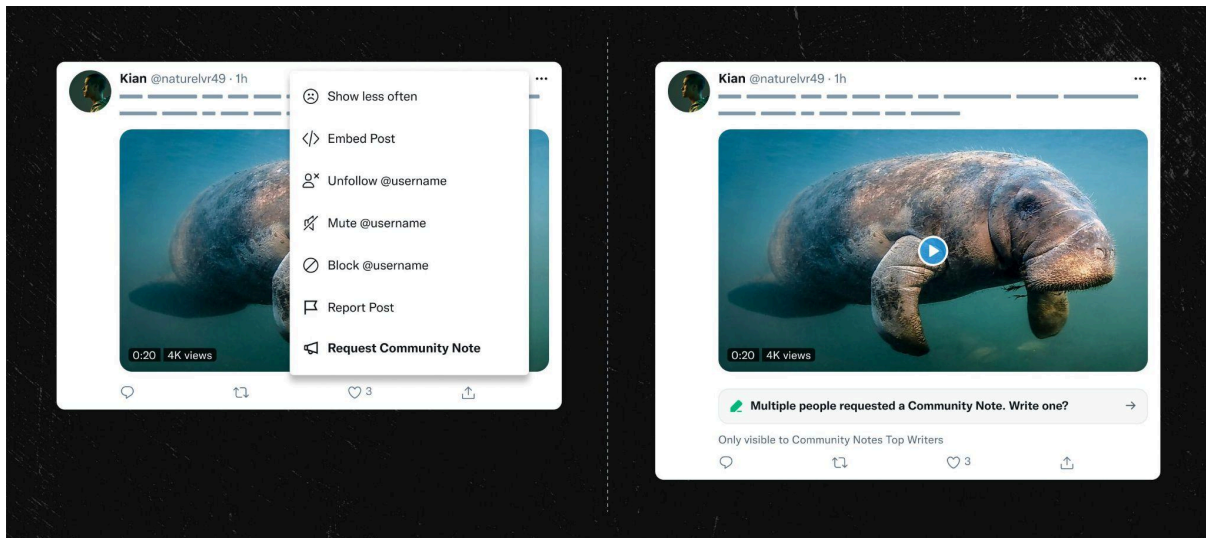


Table 3 provides the illustration of how to request a Community Note

- In **July 2024**, we introduced — by popular demand — Request a Community Note. When there are enough requests, top contributors will be alerted and can propose notes⁴³. For everyone on X, it's a way to help. For contributors, it's a way to see where help is wanted⁴⁴.
- In **June 2024**, users can see all the posts showing a media note. If a media note is

³⁸ <https://communitynotes.x.com/guide/en/contributing/diversity-of-perspectives>

³⁹ <https://help.x.com/en/rules-and-policies/x-rules>

⁴⁰ <https://communitynotes.x.com/guide/en/under-the-hood/download-data>

⁴¹ <https://x.com/CommunityNotes/status/1839035926963695858>

⁴² <https://x.com/CommunityNotes/status/183721415775564355>

⁴³ <https://x.com/CommunityNotes/status/1813980126117609624>

⁴⁴ <https://communitynotes.x.com/guide/en/under-the-hood/note-requests>



matching to other posts with the same image or video, you can see all the matches in Note Details⁴⁵.

- In **May 2024**, we updated that there were over half a million Community Notes contributors in 70 countries around the world⁴⁶ at the time of the update. We also provided updates on few studies on the performance and effectiveness of the feature:
 - **Posts with notes are (organically!) reshared less.** External researchers⁴⁷ found that users repost 61% less often after a post gets a Community Note, while another study⁴⁸ found around a 50% drop in reposts and 80% increase in post deletions after a post received a Community Note. This aligns with our own early research⁴⁹ that found a large causal drop in reposts, quotes, and likes on noted posts in an A/B test. This reduction is entirely due to organic user behavior, since X does not rank posts differently when they are noted. Notes' distributed model of knowledge production leverages the collective intelligence of the public, while also promoting accountability and trustworthiness within the platform.
 - Another recent study found⁵⁰ that, across the political spectrum, **Community Notes were perceived as significantly more trustworthy than traditional, simple misinformation flags.** It also found that Community Notes had a greater effect on improving people's identification of misleading posts. A key driver is believed to be the detailed context that notes provide, right where people can see it. We've heard time and again that people want to be given specific information that they can use to inform their understanding, and this research finding aligns with that sentiment.
 - Evaluating health information, a recent study⁵¹ published in the Journal of the American Medical Association found⁵² that **Community Notes are 97.5% accurate when addressing COVID-19 vaccine topics.** By requiring contributors to substantiate notes with sources they find helpful and showing notes that are found helpful by people who have historically disagreed with one another — aka a “bridging algorithm” — Community Notes highlight information that is found helpful to a broad range of people, even on highly contentious topics.
 - Speed is key to notes' effectiveness — the faster they appear, the more people see them, and the greater effect they have. In the past year, we've seen that notes can respond quickly at critical times. For example, in the first few days of the Israel-Hamas conflict, notes appeared at a median time of just 5 hours after posts were created. (This calculation does not even include notes on images/videos — over 80% of noted posts are showing media notes, which appear ~instantly on new posts that include previously noted media.) It's also common to see Community Notes appearing days faster⁵³ than traditional fact checks — possible because of the collective intelligence of the contributor community.
 - We are working to accelerate notes even further. In the past year, we've shaved 3-5 hours off the typical time it takes for notes to be scored, and we're working on new changes to the scoring system that will further reduce scoring time. On top of this, people who engage

⁴⁵ <https://x.com/CommunityNotes/status/1797693336683458996>

⁴⁶ <https://x.com/CommunityNotes/status/1788617818784792880>

⁴⁷ <https://osf.io/preprints/osf/3a4fe>

⁴⁸ <https://arxiv.org/abs/2404.02803>

⁴⁹ <https://arxiv.org/abs/2210.15723>

⁵⁰ <https://osf.io/preprints/osf/ydc42>

⁵¹ <https://x.com/johnwayersphd/status/1783172847948722512>

⁵² <https://jamanetwork.com/journals/jama/article-abstract/2818054>

⁵³ <https://x.com/kcoleman/status/1727419041038352602>



with a post before it receives a note get a notification about it.

- In **May 2024**, we also updated that improved image matching with notes on ~30% more posts that contain similar or identical images. We just rolled out the update and would be monitoring for any erroneous image matches⁵⁴.

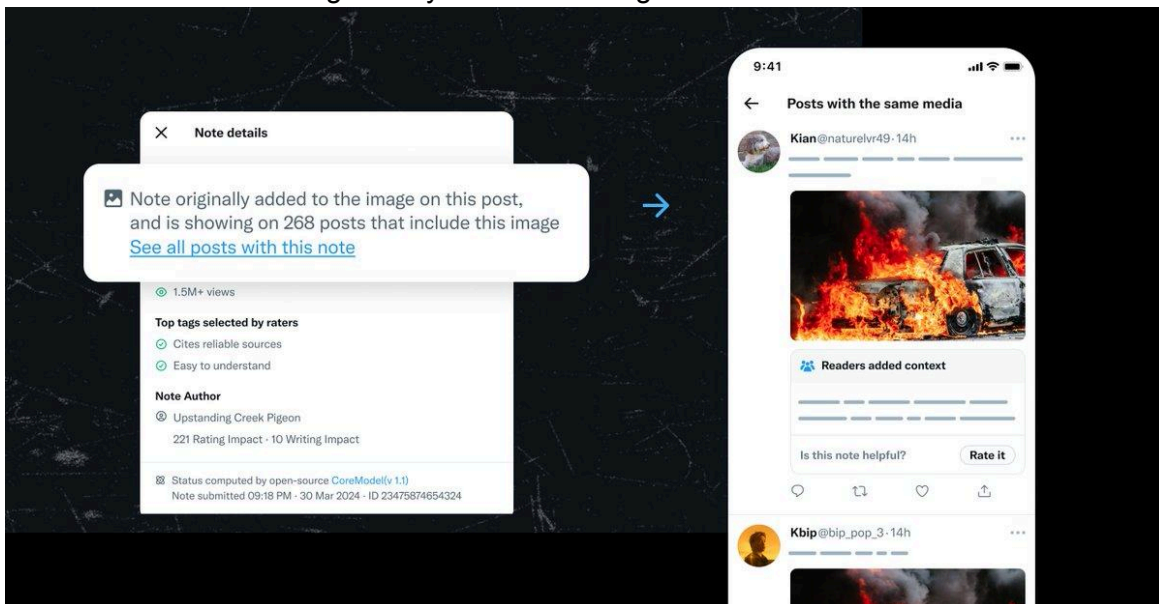


Table 4 provides an example of image matching with notes

- In **April 2024**, we launched an update that improves note quality while simultaneously increasing the number of notes shown on X by about 6%. This update also identifies 6% more Not Helpful notes, which in turn limits writing ability of authors regularly proposing such notes. It works by identifying notes that contributors from a wide range of viewpoints agree are particularly low quality (for example, that they perceive as abuse or harassment) and then reducing Contributor Helpfulness Scores for raters that rate those notes as helpful. As a result, it gives more weight to contributors who consistently help identify notes that people find helpful⁵⁵.
 - We just rolled out an optimization to the scorer that shaves 30mins of the time it takes notes to go live. Under the hood details: we moved some of the computation (e.g. the initial matrix factorization) into a prescoring step, which speeds up the final scoring step and allows it to run more frequently. It also opens up new opportunities, for example to improve prescoring without impacting the speed or frequency with which notes are scored⁵⁶.
- In **March 2024**, we introduced Topic Models, which improve Community Notes' ability to identify notes that are helpful to people from different points of view. Community Notes shows notes that are found helpful by people who've disagreed in their past ratings. Now, it will factor in both whether raters have disagreed in general as well as on the specific topic of the note. This strengthens the algorithm's ability to find notes that bridge across topics. We're initially trialling a small set of Topic Models and will monitor their effectiveness at identifying broadly helpful notes⁵⁷.
 - We updated Topic Models to utilise a larger set of early rating data. This allows them to produce higher confidence scores earlier in the process of ratings arriving⁵⁸.
- In **February 2024**, note writing limits would be considered as hit rate, a writer's ratio of helpful to total notes written. To write many notes in 24h, a writer must have a

⁵⁴ <https://x.com/CommunityNotes/status/1786469512587686213>

⁵⁵ <https://x.com/CommunityNotes/status/1781385223659553064>

⁵⁶ <https://x.com/CommunityNotes/status/1778550347306332191>

⁵⁷ <https://x.com/CommunityNotes/status/1771355926655807972>

⁵⁸ <https://x.com/CommunityNotes/status/1773858820118475025>



history of helpful notes and a solid hit rate. This prevents large volumes of proposed notes that aren't found helpful⁵⁹.


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 behaviour 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

X is a place to share ideas and information, connect with communities, and see the world around us. In order to protect the very best parts of that experience, we provide tools designed to help users control what they see and what others can see about users, so that users can express themselves on X with confidence ([here](#) and [here](#))⁶⁰.

We make it easy for users to take action on a post. Users can tap the icon  at the top of any post, right from their timeline, to quickly access options like unfollow, mute, block, report, and more. To further enhance the X experience, we may also apply controls to an account while we make sure there's a human behind an account.

- [Unfollow](#)
- [Filter notifications](#)
- [Show less often](#)
- [Mute](#)
- [Block](#)
- [Report](#)
- [Control the media you see in posts](#)
- [Protect your posts](#)
- [Photo tagging settings](#)
- [Discoverability settings](#)
- [Sharing your location in posts](#)
- [Sharing precise location through the X app](#)
- [Media settings \(to mark media in users own posts as possibly containing sensitive media\)](#)
- [Tools to authorise and connect to third party application\(s\)](#)
- [Home and latest](#)
- [Autoplay video settings](#)
- [Notifications settings](#)
- [Not interested in Topic suggestions](#)
- [X Search](#)
- [Controlling how your X information appears in Google search](#)

4.3 Enhance transparency of policies, processes and systems

Transparency helps build trust and facilitates accountability. Signatories will provide

⁵⁹ <https://x.com/CommunityNotes/status/1756047573679550659>

⁶⁰ <https://help.x.com/en/safety-and-security/control-your-x-experience>



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

At X, our purpose is to serve the public conversation, ensuring a safe environment where everyone can participate freely and confidently. With this transparency report, we aim to cement X as a truly safe platform for all. We achieve this through transparency about our content moderation systems, policies, and enforcement philosophy.

In **September 2024**, we introduced X's Global Transparency Report⁶¹, covering the period from January to June 2024. This report highlights the extensive efforts by X teams to cultivate a healthy and secure environment, reaffirming that X remains a safe platform for all users. Through this report, we provide full transparency into our content moderation processes, policies, and enforcement practices. Our community can use this comprehensive resource to better understand how we implement our policies, respond to legal requests, and safeguard our community at scale. As we continue our commitment to transparency, we will publish these reports biannually, building on our progress and successes.

It is more important than ever that we shine a light on our own practices, including enforcement of the X Rules and our ongoing work to disrupt global state-backed information operations. The public and policy makers want to be better informed about our actions and we recognize these calls for greater transparency. That is why our transparency reporting has evolved into a more comprehensive X Transparency Center covering a broad array of our transparency efforts.

Our policies and enforcement principles are grounded in human rights, and we have been taking an extensive and holistic approach towards freedom of expression by investing in developing a broader range of remediations, with a particular focus on education, rehabilitation, and deterrence. These beliefs are the foundation of "Freedom of Speech, not Freedom of Reach⁶²" - our enforcement philosophy, which means we restrict the reach of posts, only where appropriate, to make the content less discoverable as an alternative to removal.

We include sections covering information requests, removal requests, copyright notices, trademark notices, email security, X Rules enforcement, platform manipulation, and state-backed information operations. When determining whether to take enforcement action, we may consider a number of factors, including (but not limited to) whether:

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

⁶¹ <https://transparency.x.com>

⁶² https://blog.x.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy



To enforce our Rules, we use a combination of machine learning and human review. These systems either take action automatically, or surface content to human moderators based on user reports and/or proactive detection methods. Our human moderators use important context to make decisions about potential violations. This work is led by an international, cross-functional team with 24-hour coverage and the ability to cover multiple languages. We also have an appeals process for any potential errors that may occur.

We strive to create an environment where users feel empowered to express themselves. When abusive behaviour occurs, we make it easy for users to report violations of X Rules or local laws. Below, you'll find the total number of global user reports from January to June 2024, excluding Platform Manipulation and Spam, which is broken out separately below⁶³.

USER REPORTS		224,129,805
POLICY AREA	TOTAL	% OF TOTAL REPORTS
Abuse & Harassment	81,730,426	36.47%
Child Safety	8,928,019	3.98%
Hateful Conduct	66,898,539	29.85%
Illegal or Regulated Goods and Services	35,695	0.02%
Misleading & Deceptive Identities	5,179,431	2.31%
Non-Consensual Nudity	38,736	0.02%
Private Content	9,902,566	4.42%
Suicide & Self Harm	2,479,097	1.11%
Violent & Hateful Entities	8,932,100	3.99%
Violent Content	40,005,196	17.85%

Table 5 provides the total number of global user reports on X from January to June 2024

In New Zealand, from 1 Oct 2023 to 30 Aug 2024, we suspended 540 accounts and removed 10,608 contents. The majority of which are violations of Violent Content policy⁶⁴.

Transparency on X is of extreme importance. Over the last couple of months, we have started to provide more transparency on labels we may apply to a user's account that impacts how X treats their content. Our goal is to notify a user every time a product label is applied to their account. In **January 2024**, we notified around 16M X users and allowed X Premium accounts to submit feedback on the label that we'll use to improve our product features⁶⁵.

In **March 2024**, we notified and created transparency for over 21 million users when a label is applied to their account that impacts how X treats their content (e.g. accounts that frequently post sensitive media). Users can learn more about why their account has a label, what it means, and request review. Our work to increase transparency on X is ongoing, and

⁶³ <https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf>

⁶⁴ <https://help.x.com/en/rules-and-policies/violent-content>

⁶⁵ <https://x.com/Safety/status/1745144699420356965>



we're excited to continue sharing updates on our progress⁶⁶.

4.4 Support independent research and evaluation

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.

With regard to Measure 44, *“Support or convene at least one event per year to foster multi-stakeholder dialogue, particularly with the research community, regarding one of the key themes of online safety and harmful content.* Please see the list of events that include broader regional or global events undertaken by X which involve Aotearoa New Zealand.

- **On 24 June 2024**, X supported and attended Global Internet Forum to Counter Terrorism (GIFCT) Global Multistakeholder Forum 2024 that brought together industry, experts, civil society, government officials, and practitioners to discuss emerging security and tech trends and dynamics that shape terrorist and violent extremist threats online, and the solutions GIFCT is developing to address them.⁶⁷
- **On 10 December 2023**, our head of Global Government affairs attended the G7 Interior and Security Ministers' Communiqué in Japan discussing the global challenges to safety and security, and reiterated our commitment to fight against online harm⁶⁸.
- **During 10-11 November 2023**, X is an official sponsor of the Paris Peace Forum 2023 in promoting the Forum with 75,000USD worth of advertising credits that aim to “Seek Common Ground in a World of Rivalry”, against a backdrop of global polarization (most notably between China and the United States) jeopardizing international cooperation on issues vital to humanity – all in an environment marked by the eruption of conflict in the Middle East⁶⁹. Our team also participated in a dialogue and contributed to the Paris Peace Forum⁷⁰, bringing together key players in global governance at the Palais Brongniart.
- At the Paris Peace Forum, X team supported and attended Christchurch Call Leaders' Summit 2023 in Paris, a forum for upholding the commitment by online service providers, governments, and civil society to promote and respect human

⁶⁶ <https://x.com/Safety/status/1768319335997608028>

⁶⁷ <https://gifct.org/events/gifct-2024-global-multistakeholder-forum/>

⁶⁸ https://www.npa.go.jp/bureau/soumu/kokusai/20231210_G7ISMM_communique_principal.pdf

⁶⁹ <https://parispeaceforum.org/the-2023-forum/>

⁷⁰ <https://x.com/GlobalAffairs/status/1722958371563729336>



rights, and a free, open, secure internet, which was established in the aftermath of the 2019 terrorist attack in Christchurch, New Zealand⁷¹.

- **On 5 October 2023**, we attended the International Network Against Cyber Hate (INACH) “Cyber Hate Summit – Connecting to Build Bridges” in Malaga, Spain⁷².

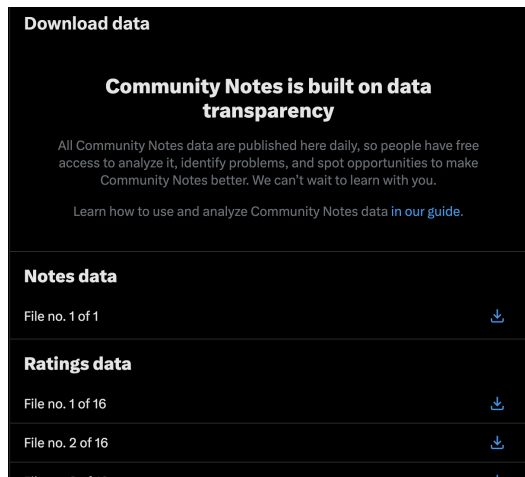
Independent research

On GitHub⁷³, you’ll find repositories (main repo⁷⁴, ml repo⁷⁵) containing the source code for many parts of X, including our recommendations algorithm, which controls the posts you see on the *For You* timeline. We’ve also shared more information on our recommendation algorithm in this post⁷⁶ on our Engineering Blog. 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.

In **May 2024**, we announced that we are transforming our API platform to new modern systems and are in the process of migrating all developers to our new X API v2 by June 30th, 2024⁷⁷. Any developer with an emergency or public utility app, will also need to move to our new v2 API by this timeline. We would continue enabling Verified government or publicly owned services who post weather alerts, transport updates and emergency notifications without cost⁷⁸.

Open-source code

Regarding Community Notes, researchers and users can find links to code⁷⁹ that reproduces the note scoring/ranking code that X runs in production. If users download the data files made available on the Data Download page⁸⁰ and put them in the same directory as the following code files, they can then run `python main.py` to produce a `scoredNotes.tsv` file that contains note scores, statuses, and explanation tags that will match what’s running in production (as of the time the data was from)⁸¹.



Downloading data

We can't wait to learn with you!

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.

If you have questions or feedback about the Community Notes public data or would like to share your analyses of this data with us, please DM us at [@CommunityNotes](#).

Working with the Community Notes data

Data snapshots

The [Community Notes data](#) is released as four separate files:

- **Notes:** Contains a table representing all notes
- **Ratings:** Contains a table representing all ratings
- **Note Status History:** Contains a table with metadata about notes including what statuses they received and when.
- **User Enrollment:** Contains a table with metadata about each user's enrollment state.

These tables can be joined on the `noteid` field to create a combined dataset with information about users, notes, and their ratings. The data is released in separate tables/files to reduce the dataset size by avoiding data duplication (this is known as a normalized data model).

⁷¹ https://x.com/GIFCT_official/status/1719747331556548845

⁷² <https://www.inach.net/inachs-2023-annual-conference-cyber-hate-summit-connecting-to-build-bridges/>

⁷³ <https://github.com/twitter/>

⁷⁴ <https://github.com/twitter/the-algorithm/>

⁷⁵ <https://github.com/twitter/the-algorithm-ml>

⁷⁶ https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

⁷⁷ <https://x.com/XDevelopers/status/1785779279134925003>

⁷⁸ <https://devcommunity.x.com/t/x-api-v2-migration/203391>

⁷⁹ <https://github.com/twitter/communitynotes/tree/main/sourcecode>

⁸⁰ <https://x.com/i/communitynotes/download-data>

⁸¹ <https://communitynotes.x.com/guide/en/under-the-hood/note-ranking-code>



Table 6 provides how to download data for all Community Notes contributions

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 better⁸².

Note that the algorithm is split into two main binaries: prescoring and final scoring. In production at X, each binary is run separately as often as possible, each always reading the most recent input data available. One subtle implication of this is that in order to exactly reproduce the scoring results as they are run in production at X, the prescorer should be run on input data that's 1 hour older than the final scorer (although this makes very little difference in practice)⁸³. Community Notes' open-source nature has facilitated independent research and exchanging ideas, including the following:

- **Posts with notes are (organically!) reshared less.** External researchers⁸⁴ found that users repost 61% less often after a post gets a Community Note, while another study⁸⁵ found around a 50% drop in reposts and 80% increase in post deletions after a post received a Community Note. This aligns with our own early research⁸⁶ that found a large causal drop in reposts, quotes, and likes on noted posts in an A/B test. This reduction is entirely due to organic user behavior, since X does not rank posts differently when they are noted. Notes' distributed model of knowledge production leverages the collective intelligence of the public, while also promoting accountability and trustworthiness within the platform.
- Another recent study found⁸⁷ that, across the political spectrum, **Community Notes were perceived as significantly more trustworthy than traditional, simple misinformation flags.** It also found that Community Notes had a greater effect on improving people's identification of misleading posts. A key driver is believed to be the detailed context that notes provide, right where people can see it. We've heard time and again that people want to be given specific information that they can use to inform their understanding, and this research finding aligns with that sentiment.
- Evaluating health information, a recent study⁸⁸ published in the Journal of the American Medical Association found⁸⁹ that **Community Notes are 97.5% accurate when addressing COVID-19 vaccine topics.** By requiring contributors to substantiate notes with sources they find helpful and showing notes that are found helpful by people who have historically disagreed with one another — aka a "bridging algorithm" — Community Notes highlight information that is found helpful to a broad range of people, even on highly contentious topics.

⁸² <https://communitynotes.x.com/guide/en/under-the-hood/download-data>

⁸³ *Id.*

⁸⁴ <https://osf.io/preprints/osf/3a4fe>

⁸⁵ <https://arxiv.org/abs/2404.02803>

⁸⁶ <https://arxiv.org/abs/2210.15723>

⁸⁷ <https://osf.io/preprints/osf/ydc42>

⁸⁸ <https://x.com/johnwayersphd/status/1783172847948722512>

⁸⁹ <https://jamanetwork.com/journals/jama/article-abstract/2818054>