07 June 2024

Wifredo "Wifi" Fernández
Head of US & Canada
Global Government Affairs

X Corp.
1355 Market St #900
San Francisco, CA 94103

Dear Chairman Warner,

As the global town square for public conversation, the public debate in the run-up to the elections will unfold in large part on X. We are proud that our platform is a driver of conversations that fuel the democratic debate and life around the world, while simultaneously recognizing our significant responsibility to keep these conversations safe during electoral processes.

Artificial Intelligence

X is also the platform on which the public conversation about the advancements of artificial intelligence is happening. In the last year, there's been over 10 billion impressions worth of conversation about AI on X. On May 15 alone there were 330 million impressions — that's up 323% from the same day last year. Given the volume and diversity of conversation on X around AI, our experimentation with new tools and services, and our users' sharing of content, there is a heightened awareness among our users around AI-generated content and thus more conversation about the provenance of content. To that end, we are working on product features that will allow users to label their content as AI-generated. We are also exploring product interventions that would help us detect AI-generated content.

Platform Manipulation and synthetic media

In times of elections and at all times, we believe that it is critical to maintain the authenticity of the conversation on X. Our Safety teams remain alert to any attempt to manipulate the platform by bad actors and networks. We have a robust policy in place to prevent platform spam and manipulation, and we routinely take down accounts engaged in this type of behavior. We also make sure that we are well equipped to fight against any manipulated media—including the recent trend of ''deepfakes''—and would put visible labels on any such content that has been debunked by a credible source.

Accounts engaged in information manipulation frequently employ sophisticated and constantly evolving platform manipulation tactics. As part of our set of rules on platform integrity and authenticity, X forbids platform manipulation and spam. We define platform manipulation as using X to engage in bulk, aggressive, or deceptive activity that misleads others and/or disrupts their experience. For instance, this can take the form of inauthentic engagements, that attempt to make accounts or content appear more popular or active than they are, or coordinated harmful activity, which attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting.

At the content level, X prohibits the sharing of synthetic, manipulated, or out-of-context media ("SAMM") that may deceive or confuse people and lead to

harm ("misleading media"). This policy prohibits misleading media which has been:

- Significantly and deceptively altered, manipulated, or fabricated, or
- Shared in a deceptive manner or with false context, and
- Likely to result in widespread confusion on public issues, impact public safety, or cause serious harm

To determine the whether a piece of SAMM meets the threshold of being actioned, our enforcement team considers several criteria, including:

1. Whether the content is significantly and deceptively altered, manipulated, or fabricated;
2. Whether the content was shared in a deceptive manner or with false context;
3. Whether the content is likely to result in widespread confusion on public issues, impact public safety, or cause serious harm.

Based on the above questions, content may be deleted, labeled, and the posting account may be locked or suspended.

Under X's Civic Integrity policy, it is prohibited to use X's service with the purpose of manipulating or interfering in elections or other civic processes, such as posting or sharing content that may suppress participation, to mislead people about when, where, or how to participate in a civic process, and to incite 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 visible labels to violative posts informing users that the content violates our Civic Integrity policy and we disable the ability to share, like, re-post, or bookmark the post.

X has clear policies on **misleading and deceptive identities**. Accounts use tactics such as impersonation and fake identities to conduct information manipulation and propagate false or misleading information. X does not allow users to misappropriate the identity of individuals, groups, or organizations or use a fake identity to deceive others. We want X to be a place where people can find authentic voices. While users are not required to display their real name or image on their profile, accounts should not use false profile information to represent itself as a person or entity that is not affiliated with the account owner, such that it may mislead others who use X.

We prohibit the following behaviors under this policy:

- Impersonation: You may not pose as an existing person, group, or organization to mislead others about who you are or who you represent. Accounts violate this policy when they misrepresent their identity by using at least two elements of another identity, such as the name, image, or false claims of affiliation with another individual or organization in their profile or posts.

- Deceptive Identities: You may not pose as someone who doesn't exist to mislead others about who you are or who you represent. This

includes using at least one element of someone else's identity on your profile or posts in a misleading way, such as using another individual's image or making a false statement of affiliation with an existing individual or entity. We also consider accounts to be deceptive if they use a computer generated image of a person to pose as someone who doesn't exist.

If you believe an account is posing as you or your brand, users can file a report via our Help Center. If you believe an account is using a deceptive fake identity or misusing the identity of somebody else, users can flag it as a bystander by reporting directly from the account's profile.

In cases where an account is suspected of misusing a specific individual or entity's identity, we may require more information to determine whether the account is run or authorized by the entity portrayed in the profile.

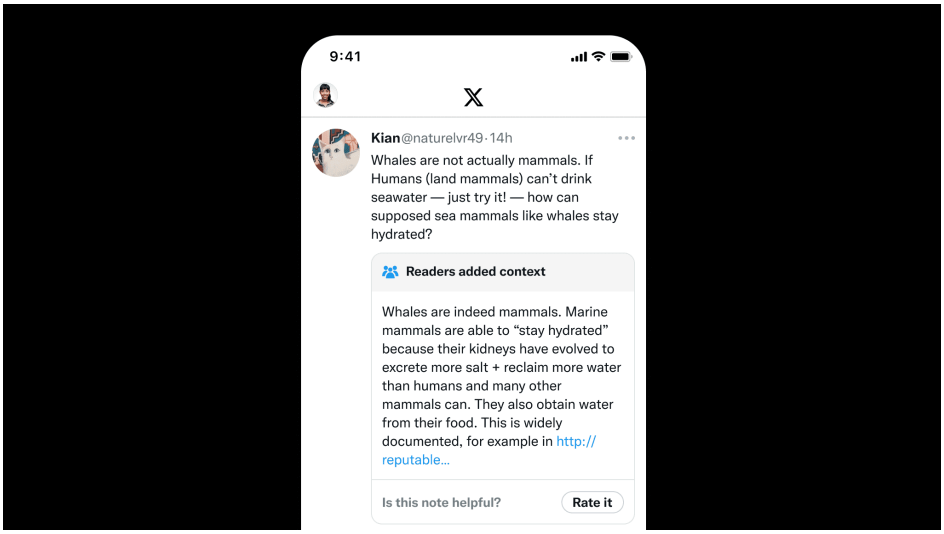
Community Notes

Community Notes aim to create a better informed world by empowering people on X to collaboratively add context to potentially misleading posts and media (including advertisements). By making this feature an integral and highly visible part of X, and by ensuring that the user interface is simple and intuitive, we are investing in a tool that is global in its application. It also reduces our reliance on forms of content moderation that are more centralized, manual and bespoke; or which require intensive and time-consuming interactions with third parties.

Here is how it works:

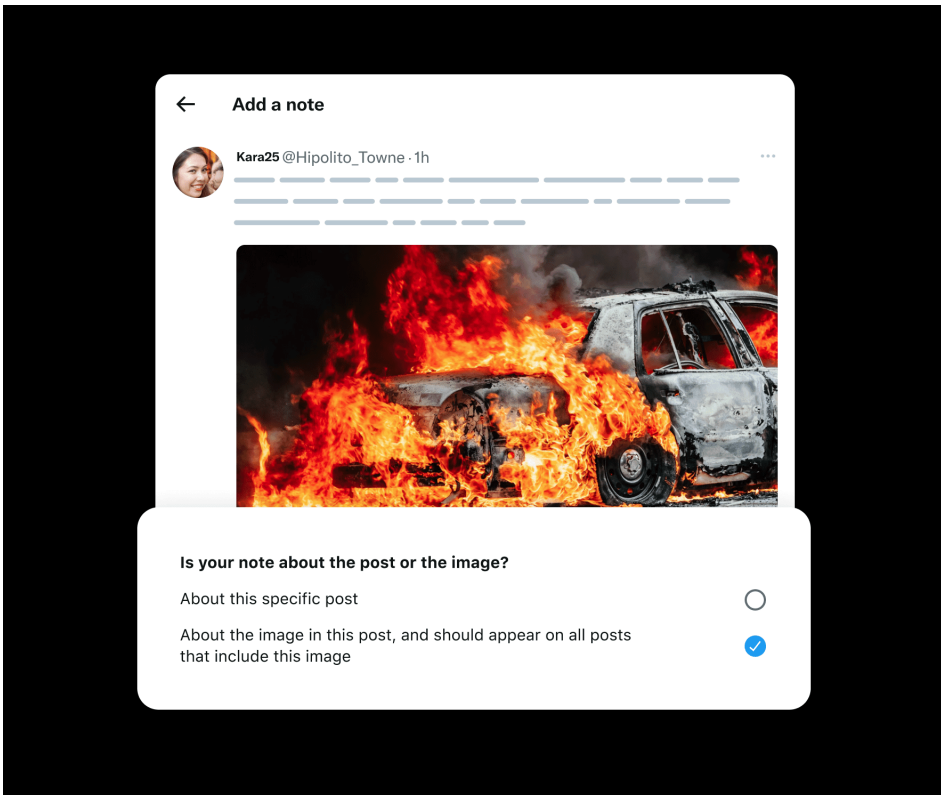
- **Contributors write and rate notes:** Contributors are people on X who sign up to write and rate notes. The more people that participate, the better the program becomes.
- **Only notes rated helpful by people from diverse perspectives appear on posts:** Community Notes do not work by majority rules. 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 ratings. We have published and will continue to learn more about how Community Notes handles diverse perspectives.
- **X does not choose what shows up, the people do:** X does not write, rate, or moderate notes (unless they break the X Rules.) 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 is important for people to understand how Community Notes work 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 anyone. Learn more about how it works through our dedicated Community Notes Guide.

We acknowledge and are keenly aware that a product like this can be subject to attempts of abuse and manipulation, which we proactively measure and mitigate. You can read more here on how we are thinking about quality control, guardrails, circuit breakers, and the various remediations we have in place to challenge bad actors.



Notes on Media

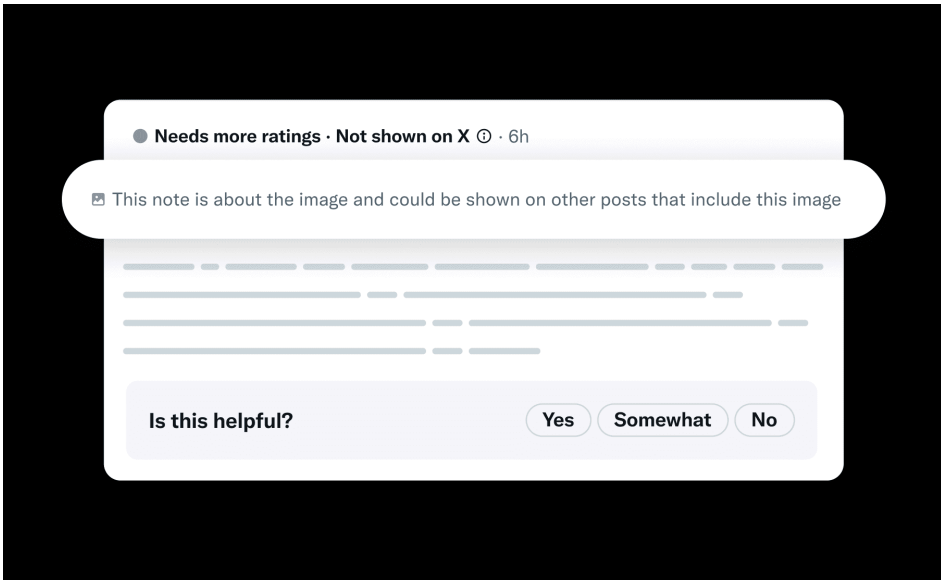
Community Notes are frequently added to posts that feature images or videos. In many cases, these notes can provide valuable context, not just for a single post, but for any post containing the same media. This feature is especially important for addressing the challenges of media produced by generative artificial intelligence tools. Contributors with a Writing Impact score of 10 or above have the option to write notes about the media found within posts, as opposed to focusing on the specific post. Contributors should select this option when they believe the context added would be helpful independently of the post the note is attached to.



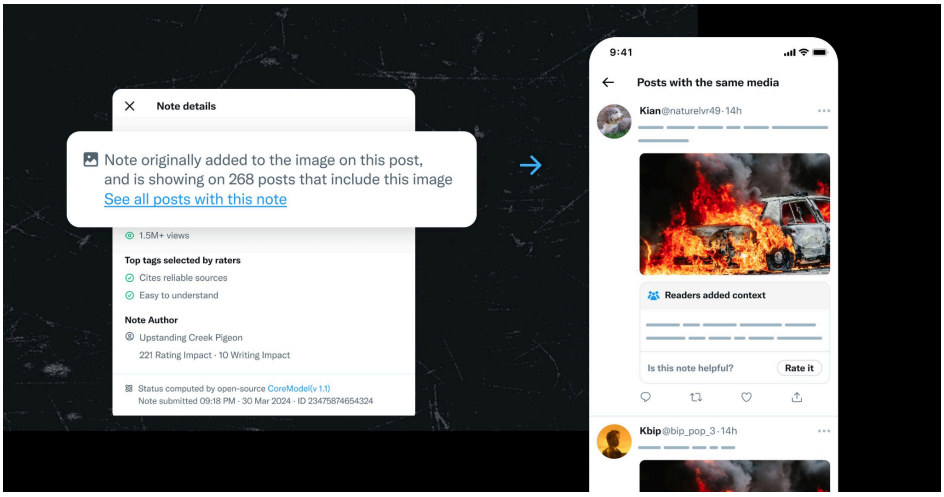
Tagging notes as “about the image” makes them visible on all posts that our system identifies as containing the same image. These notes, when deemed Helpful, accumulate view counts from all the posts they appear in, but only count as one Writing and Rating Impact for the author and raters.

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.

This number grows automatically if the relevant images and video are re-used in new posts.



Now you can see all the posts showing a media note. If a media note is matching to other posts with the same image or video, you can see all the matches in Note Details.



Currently, this feature supports posts with a single image. We are actively working on expanding it to support posts with multiple images, GIFs, and videos.

There are now over 500,000 Community Notes contributors in 70 countries around the world. In 2023, we showed 37,000+ notes well over 14 billion times, and in just the first four months of 2024, we've already shown 29,000+ notes that have been seen over 9 billion times. An amazing pace of growth, covering more topics in more languages every day.

People contribute to Community Notes because they want to create a better informed world and they are having an impact. Thanks to Community Notes' open source data & code, it's possible for the public to study its performance and effectiveness. Four new studies have done just that. They found:

Posts with notes are (organically) reshared less. External researchers found that users repost 61% less often after a post gets a Community Note, while another study 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 research 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 from the University of Giessen in Germany found 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 study published in the Journal of the American Medical Association found 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 faster 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 with a post before it receives a note get a notification about it.

Deepfakes, "shallowfakes," AI-generated photos, out-of-context media and more worry people. This past year, we put a superpower into contributors' hands, allowing them to write notes that are automatically shown on posts with matching media. To give you a sense of the multiplying effect this has on impact, the ~3,500 media notes that have been written are now showing on over 331,000 distinct posts and have been seen over 1.1 billion times. In some cases a single note will match to thousands of posts. Notes have identified everything from reported deepfake audio of military officials and politicians, to fake magazine covers, photoshopped images, out-of-context conflict footage, video game footage masquerading as real video, and more. We've initially given the ability to write media notes to contributors who have earned Top Writer status, and may expand access further.

This work is made possible by each and every contributor, whose work is helping the world stay informed. The public's feedback is also helping shape @CommunityNotes — from recently optimized note writing limits, improvements to the way writing ability is earned and lost, and to the upcoming launch of note translation, your feedback leads to regular updates to Community Notes itself.

Collaboration with State Officials

Here in the US, we are in regular communication with Secretaries of State and State Election Directors, for whom we've hosted multiple webinars and education sessions. These elections regulators will have direct escalation pathways to report potential violations of our Terms for evaluation. We are participating in tabletop exercises with local elections officials in multiple states, onboarding state attorneys general offices to our law enforcement portal to be able to report illegal content, and engaging with NGOs to support

their civic outreach efforts. We are in regular contact with our European regulator the European Commission, and have reached out to all DSCs across Europe to explain our efforts to protect the conversation around EU elections, and provide them with an emergency channel in case of needs. We have also offered our NGO partners in the EU refresher training on our safety tools so they can partner with us in securing the service.

We are constantly iterating on our approach as we learn from every election around the world, and in this special year where so many are engaging in the franchise of voting, we are learning more than ever. We will continue to explore ways to collaborate with fellow signatories of the Tech Accord to share best practices and information that advance our principal goals. We appreciate the opportunity to share our approach to the challenges around elections and novel AI-generated content, and welcome your thoughts and feedback.

Sincerely,

Wifredo “Wifi” Fernández
Head of US & Canada
Global Government Affairs