Wifredo "Wifi" Fernández
Global Government Affairs

X Corp.
865 FM 1209 Building 2
Bastrop, TX 78602

10 April 2025

Via Email

Chair Chi Onwurah MP
Science, Innovation, and Technology Committee, House of Commons

Re: Follow-ups from 25 February oral evidence session

Chair Onwurah:

I write in response to your letter of March 20, 2025 regarding your follow-up questions to the oral evidence session held in the Science, Innovation, and Technology Committee on February 25, 2025.

Q1) Details on the resources that X dedicated to responding to the unrest in the UK last summer, broken down by number of personnel, location, and type of role; and stating the period they were deployed, and the type of work they were undertaking.

As external extraordinary events continue to disrupt our business and harm our users on the platform, mitigating crisis situations remains one of the critical work streams for our Safety team. This means, not only do we need a cohesive, consistent process that enables us to make risk-informed decisions, allocate resources and apply timely and appropriate remediation measures, but our mission to #MakeCrisesBoring stays. Our end state should be to proactively prevent risks from each crisis, thereby increasing our overall crisis preparedness. Our crisis response efforts are made possible by a cross-functional team that provides 24/7 diligence and vigilance as the events unfold.

The manner in which we deploy individuals or teams to address crises or specific types of content, and the detail of our related enforcement processes and policies, is sensitive. Bad actors seek to exploit publicly available information and develop new tactics and techniques to circumvent platform rules and policies. As a consequence, we cannot disclose the specific details the Committee requests.

However, as of 30 September 2024, we had 1275 people working in content moderation across the world (and are not specifically designated to work on UK matters or specific responses to world events). English is the first language of 1117 of these individuals and the majority have a minimum of two years' experience in content moderation. The Committee may find additional details of X's content moderation activities and philosophy in its Global and DSA Transparency Reports, published on the X website.

In crisis situations, X's unique purpose to serve the public conversation becomes all the more critical; users use our platform to access and share information, raise awareness about the situations they experience and witness on the ground, and openly and freely exchange ideas on a wide array of topics. The following principles are intended to guide our actions and decisions during crises:

Harm prevention and prioritisation
In times of crisis, the propensity for harm associated with online content and behaviour increases significantly. Where such content and behaviour threaten the safety or security of our users, the public, and vulnerable populations, our goal is to mitigate the risk of harmful consequences. As such, we may intervene more rapidly and broadly than normal. While we strive to balance the rights to free expression, civic participation and privacy, we recognise that these could conflict with one another. If there is an instance of acute potential for harm, we may need to prioritise protection of physical safety over other considerations.

Impartiality and unbiased decision-making
We strive to objectively enforce the X Rules and maintain safeguards in an effort to avoid unintended discrimination and biases in our policies and processes. Our decisions are made based on evidence, rigorous assessment frameworks, relevant international legal standards and commitments, and local contexts, such as linguistic, social and cultural nuance. Our remediations are applied as necessary, proportionate, legitimate, and in a nondiscriminatory manner.

Proportionality, wherever possible
Each of our policies is meant to prevent or mitigate a known, quantifiable harm. The higher the risk, the more policy enforcement guidance and processes should be focussed on taking action to prevent the potential harm. We use a continuum of interventions to calibrate necessary enforcement based on the propensity for harm. While we may err on the side of over-enforcing during situations of heightened risk, having a robust appeals mechanism should help mitigate enforcement overreach. Additionally, we seek to regularly reevaluate the situation and conditions to ensure that circumstances continue to warrant our interventions.

Transparency
We strive to ensure that the actions we undertake in crisis situations are externally transparent and clearly communicated to the people we serve. This includes being transparent and clear about when we've taken a remediation action and why, when we modify our policies or product interventions, when we deviate from standard enforcement processes to prioritise harm prevention, and sharing information on user appeals and enabling the processes to correct our mistakes.

Q2) Detail on the advertisers that paused activity on X during the summer unrest last year, including numbers of advertisers, the period for which they paused this activity, the volume of adverts and sums of money involved, and X's response to this.

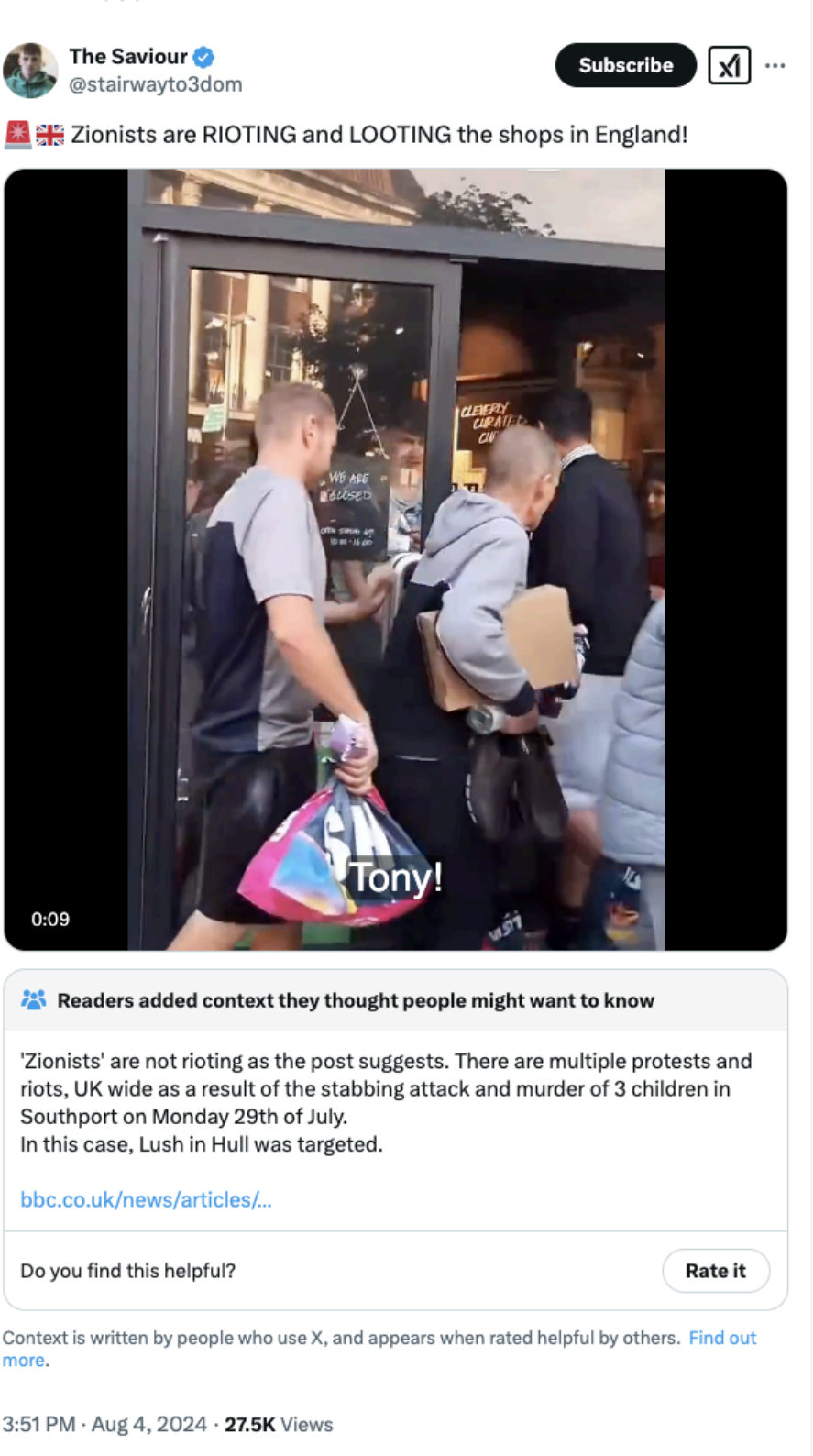
The Committee seeks information regarding advertiser activity during the summer unrest. Details regarding X's customers and clients are sensitive business information and in the interest of our customer's data and business privacy, we cannot disclose any further details.

Q3) Details of examples that you mentioned in the session of posts by figures such as Andrew Tate and Stephen Yaxley-Lennon (aka Tommy Robinson) that received 'Community Notes' in the wake of the Southport attack.

- Detail on how many posts with content related to the Southport attack received 'Community Notes' in the days after the attack, how quickly these notes were attached to the posts, and your evaluation of the efficacy of this system in that context, including how effective these Notes were in preventing further sharing of misleading or harmful content.

The following sample of five (5) Posts were retrieved on February 24, 2025 from the Community Notes public database via Community Notes Leaderboard

tool and referenced during the oral evidence session. This tool, created by an independent community member, is made possible by the open-source code powering the Community Notes system and the publicly available Community Notes data which can be found at: https://communitynotes.x.com/guide/en/under-the-hood/download-data



https://x.com/stairwayto3dom/status/1820110292094705802

Alex Jones ✓
@RealAlexJones

Subscribe

Islamic attack forces are above the law in the UK.

> **Tommy Robinson** 🇬🇧 ✓ @TRobinsonNewEra · Aug 4, 2024
> Holiday Inn express in Tamworth is set on fire.
>
> Police stand aside as it's done.
>
> Very sus.

Readers added context they thought people might want to know

This isn't a video of 'Islamic forces'. These are British protesters attacking a hotel that houses asylum seekers.

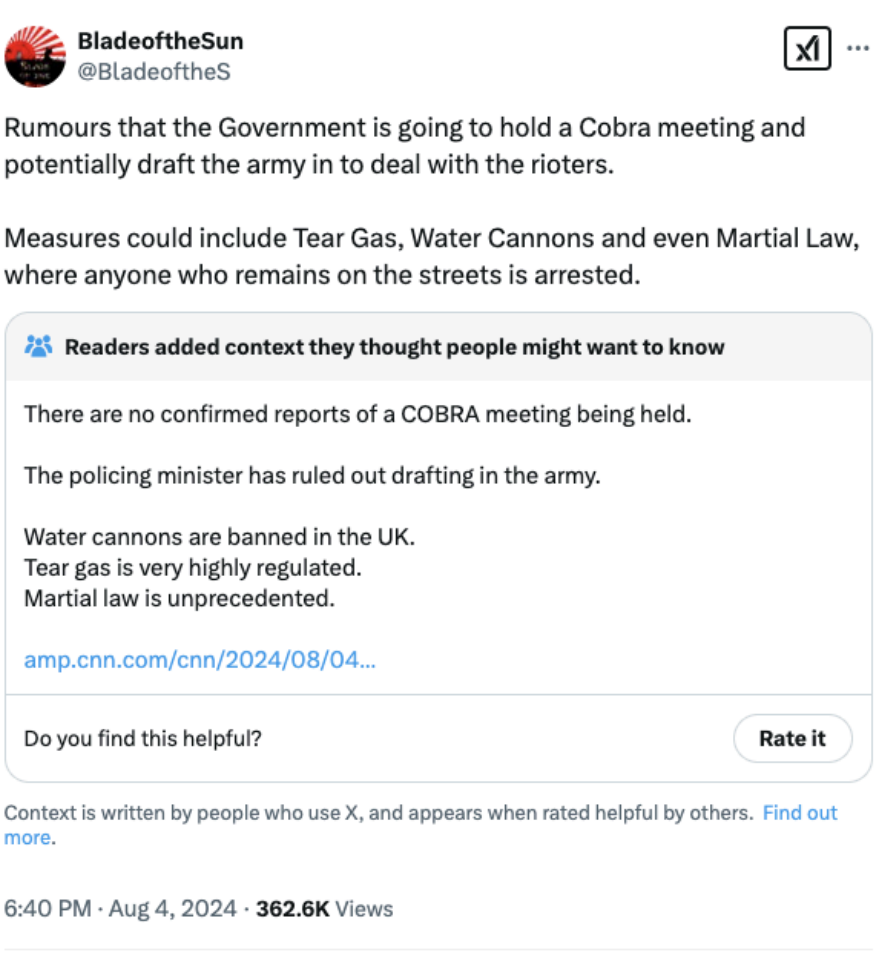
news.sky.com/video/uk-riots...

Do you find this helpful? **Rate it**

Context is written by people who use X, and appears when rated helpful by others. Find out more.

11:15 PM · Aug 4, 2024 · **188.4K** Views

https://x.com/RealAlexJones/status/1820221929921417330



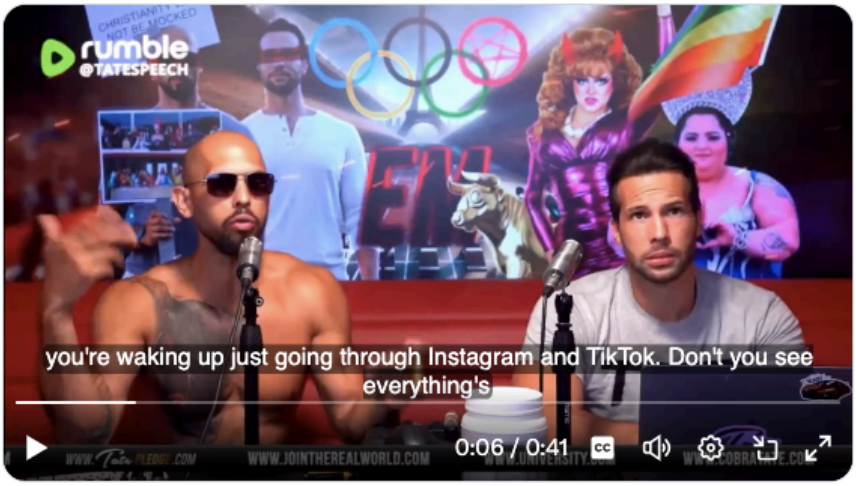
https://x.com/BladeoftheS/status/1820152850715988008

← **Post**

Tate News
@TateNews_
Parody account

Subscribe

Andrew Tate speaks on the asylum seeker stabbing little girls 💔



Readers added context they thought people might want to know

The suspect was born in Cardiff. This has been confirmed by the Police and several media outlets

amp.theguardian.com/uk-news/live/2...

independent.co.uk/news/uk/crime/...

Do you find this helpful? **Rate it**

Context is written by people who use X, and appears when rated helpful by others. Find out more.

11:21 AM · Jul 31, 2024 · **169.7K** Views

https://x.com/TateNews_/status/1818592879083651515

https://x.com/TiceRichard/status/1820082619582664754

We want X to be the most accurate source of information on the internet. That is why we have deeply invested in the development and expansion of Community Notes, which now empowers nearly 1,000,000 contributors, including over 71,000 in the UK, to add helpful context to posts, including ads. Community Notes contributor enrollment is open in 197 countries and territories, and it covers all languages in which X is available.

Like with our recommendation algorithm, the bridge-ranking algorithm and all the code that powers Community Notes is also open-source. Because Community Notes data since its inception is publicly available, it serves as a rich source for researchers studying information integrity. A recent study found that, across the political spectrum, Community Notes were perceived as

significantly more trustworthy than traditional, simple misinformation flags.[1] It also found that Community Notes had a greater effect on improving people's identification of misleading posts. A separate independent study shows that posts with notes are shared 50-61% less[2], and deleted 80% more.[3]

X remains committed to investing in and enhancing Community Notes. Deepfakes, shallowfakes, AI-generated photos, out-of-context media, and similar content are a source of public concern. This past year, we put a new superpower into contributors' hands, allowing them to write notes that are automatically shown on posts with matching media.

We have also introduced the ability for anyone to request a Community Note. With enough requests, top contributors will be alerted and can propose notes. For everyone on X, it is a way to help. For contributors, it is a way to see where help is needed.

X has also introduced Lightning Notes, which builds upon the Community Notes system and is designed to enhance community fact-checking speed by dramatically reducing the time between post publication and the publication of applicable contextual notes.[4] Notes can go live in as little as 14 minutes after being written.

By empowering users and contributors worldwide and accelerating the delivery of contextual notes, Community Notes enfranchises civil society in content moderation and fact-checking. It offers a decentralized, transparent approach to harness collective insight to combat misinformation, fostering a more informed and engaged online community.

Q4) Further detail on any other action that was taken on posts potentially spreading related misleading or harmful content in the period, such as algorithmic restriction, post takedowns, banning of accounts, and any other measures, and how quickly these actions were taken.

We were deeply disturbed by the brutal attack that took place in Southport on Monday, July 29, 2024. Our Safety team were vigilant and steadfast in response, and in our commitment to keeping the platform safe following the initial attack and in relation to subsequent incidents that occurred across the UK.

X has internal incident response protocols in place when a high-visibility event occurs and virality triggers rapid and widespread proliferation of various content types on the platform. X has a dedicated team monitoring the platform conversation and on the ground developments. This team proactively reviews narratives on the platform against various policies, including Violent Speech, Hateful Conduct, and Synthetic and Manipulated Media.

When violative trends are found, this team works with relevant stakeholders to update platform enforcement guidelines, informs operational teams, and proactively escalates violative posts for swift actions.

In response to the brutal attack, we were vigilant and our Safety team conducted active monitoring of the platform to surface content in violation of the X Terms of Service. As a result, proactive action was taken on tens of thousands of pieces of content under our rules. Actions included:

- Account suspension

[1] https://academic.oup.com/pnasnexus/article/3/7/pgae217/7686087
[2] https://osf.io/preprints/osf/3a4fe_v1
[3] https://arxiv.org/abs/2404.02803
[4] https://web.swipeinsight.app/posts/x-launches-lightning-notes-with-enhanced-speed-for-community-fact-checking-12199

- Content removal
- Actioning content for violations of the X Hateful Conduct and Violent
- Speech policies
- Labelling content as graphic, violent or sensitive media
- Labelling content as manipulated media
- Labelling content as out of context

Q5) During the evidence session, a committee member highlighted examples of X accounts posting hateful and threatening content, some of them with 'X Premium' blue tick status, which seemingly had not been subject to enforcement action. In response, you stated that your teams would review these posts under X's terms of service. Please provide the Committee with an update on the status of these posts and accounts, including the dates and description of any reports made by other users, and of actions taken by X.

The Committee member has yet to follow up directly with X with specific citations and links to the posts referenced in the hearing.

Q6) Details of X's Community Notes system, including the algorithm and 'random number generator' that onboards X users to become Community Notes contributors, and any data allowing for a comparison between the effectiveness of Community Notes-based moderation and content-moderator based Moderation.

Community Notes is built on transparency as X believes that it is important for its users and the public to understand how it works. X aims to publish all contributions to Community Notes daily.[5] X's Community Notes algorithm and program details are also open sourced on GitHub. This allows Community Notes to be academically studied and inspected by the public, who can then identify any bugs, biases and opportunities for improvement.

We know there are many challenges involved in building an open, participatory system like Community Notes — from making it resistant to manipulation attempts, to ensuring it isn't dominated by a simple majority or biased because of the distribution of contributors.

Here are a handful of particular challenges we are aware of as well as steps we are taking to address them:

Preventing Coordinated Manipulation Attempts
Attempts at coordinated manipulation represent a crucial risk for open rating systems. We expect such attempts to occur, and for Community Notes to be effective, it needs to be resistant to them.
The program currently takes multiple steps to reduce the potential for this type of manipulation:

- First, all X accounts must meet the eligibility criteria to become a Community Notes contributor. For example, having a unique, verified phone number. These criteria are designed to help prevent the creation of large numbers of fake or sock puppet contributor accounts that could be used for inauthentic rating.
- Second, Community Notes doesn't work like many engagement-based ranking systems, where popular content gains the most visibility and people can coordinate to mass upvote or downvote content they don't like or agree with. Instead, Community Notes uses a bridging algorithm — for a note to be shown on a post, it needs to be found helpful by people who have tended to disagree in their past ratings.
  Academic research indicates that bridging-based ranking can help

[5] https://communitynotes.x.com/guide/lt/under-the-hood/download-data

to identify content that is healthier and higher quality, and reduce the risk of elevating polarizing content.

- In addition to requiring ratings from a diversity of contributors, Community Notes has a reputation system in which contributors earn helpfulness scores for contributions that people from a wide range of perspectives find helpful.
  Helpfulness scores give more influence to people with a track record of making high-quality contributions to Community Notes, and lower influence to new accounts that have yet to demonstrate a track record of helpful ratings and contributions.
- Lastly, Community Notes tracks metrics that alert the team if suspicious activity is detected, and has a set of guardrails and procedures to identify if contribution quality falls below set thresholds. This helps Community Notes to proactively detect potential coordination attempts and impacts to note quality.

Reflecting Diverse Perspectives, Avoiding Biased Outcomes

Community Notes will be most effective if the context it produces can be found to be helpful by people of multiple points of view–not just people from one group or another. To work towards this goal, Community Notes currently takes the following steps:

- First, as described above, Community Notes uses a bridging based algorithm to identify notes that are likely to be helpful to people from many points of view. This helps to prevent one-sided ratings and to prevent a single group from being able to engage in mass voting to determine what notes are shown.
- Second, Community Notes can proactively seek ratings from contributors who are likely to provide a different perspective based on their rating history. This is currently done in the Needs Your Help tab, and we are exploring new ways to quickly collect ratings on notes from a wide range of contributors.
- Third, to help ensure that people of diverse backgrounds and viewpoints feel safe and empowered to participate, Community Notes has implemented program aliases that aren't publicly associated with contributors' X accounts. This can help prevent one-sided-ness by providing more diverse contributors with a voice in the system.
- Finally, we regularly survey representative samples of customers who are not Community Notes contributors to assess whether a broad range of people on X are likely to find the context in Community Notes to be helpful, and whether the notes can be informative to people of different points of view.
  This is one indicator of Community Notes' ability to be of value to people from a wide range of perspectives vs. to be biased towards one group or viewpoint. Customers who aren't enrolled Community Notes contributors can also provide rating feedback on notes they see on X. This provides an additional indicator of note helpfulness observed over time.

Avoiding Harassment

It's crucial that people feel safe contributing to Community Notes and aren't harassed for their contributions. It's also important that Community Notes itself does not become a vector for harassment. Here are measures Community Notes takes to keep everyone safe:

- First, as described above, all contributors get a new, auto-generated display name (or alias) when they join Community Notes. These aliases are not publicly associated with contributors' X accounts, so everyone can write and rate notes privately. This inhibits public identification and harassment of contributors.
- Second, contributors have an open communication line with the Community Notes team to report any issues they experience (they

can reach us by DM @CommunityNotes). Community Notes has a dedicated community manager who gathers and responds to contributors' feedback or concerns via this handle. This provides a way for contributors to flag potential issues to the Community Notes team.

- Finally, all Community Notes contributions are subject to X's Rules, Terms of Service and Privacy Policy. If you think note might not abide by the rules, you can report it by clicking or tapping the ••• menu on a note, and then selecting "Report", or by using the Report a Community Note form. This provides a mechanism to address violating content in notes.

Reducing the impacts of low quality contributions on the system

As Community Notes expands and grows, there's an increasing need to mitigate contributor rater burden and ranking system impacts that can be caused by high volumes of low quality notes or ratings. This could occur as more well-intentioned new contributors onboard to Community Notes and go through a period of learning how to write and identify helpful notes, or through bad faith attempts to overwhelm Community Notes with spammy, low quality notes or ratings.

Community Notes currently takes the following steps to reduce impact of low quality contributions:

- First, Community Notes has implemented an onboarding system, which requires new contributors to begin by rating notes before they unlock the ability to write new notes. This helps to ensure new contributors have ample opportunity to understand what makes notes Helpful and Not Helpful before they begin writing.
- Second, as a part of the system described above, contributors can also have their note writing ability temporarily locked if they write too many notes that reach a status of Not Helpful after being rated by other Community Notes contributors.
  Contributors receive warnings when they are at risk of having note writing ability locked, as well as after it is locked, to help them understand how to improve. To unlock the ability to write again, they must repeat the process of successfully rating and identifying helpful notes. This can help well-intentioned contributors learn to improve their note writing, and prevents bad actors from continuing to write high volumes of Not Helpful notes over time.
- Finally, to reduce impacts of spammy or low quality ratings, Community Notes has a note rating reputation system (described above in preventing coordinated manipulation). This gives lower influence to new accounts that have yet to demonstrate a track record of helpful ratings and contributions, making it more difficult for an account to join and spam the system with low quality ratings.

To become a Community Notes contributor, accounts must have: (1) no recent notice of violations of X's Rules; (2) joined X at least 6 months ago; and (3) a verified phone number. That phone number must be: (1) from a trusted phone carrier and (2) not associated with other Community Notes accounts. These criteria make it more likely that contributors are real people instead of bots or adversarial actors.

Our goal is to admit applicants on a periodic basis. We will admit all contributors who meet the required criteria, but if we have more applicants than slots, we will randomly admit accounts — so as to reduce the likelihood that contributors would be predominantly from one ideology, background, or interest space.

Applicants from countries with larger user bases may experience longer waiting times between signup and admission, as random admissions are drawn from country-specific waitlists.

To promote transparency, all contributions to Community Notes are anonymized and publicly visible on the Community Notes site, even if an account's posts are protected. Contributions are subject to X's Rules, Terms of Service and Privacy Policy. Failure to abide by the rules can result in removal from the Community Notes program, and/or other remediations.

Note-ranking
Community Notes are submitted and rated by contributors. Ratings are used to determine note statuses ("Helpful", "Not Helpful", or "Needs More Ratings"). Note statuses determine which notes are displayed on each of the Community Notes Site's timelines, and which notes are displayed on posts.

All Community Notes start with the Needs More Ratings status until they receive at least 5 total ratings. Notes with 5 or more ratings may be assigned a status of Helpful or Not Helpful according to the algorithm described below. If a note is deleted, the algorithm will still score it (using all non-deleted ratings of that note) and the note will receive a status if it's been rated more than 5 times, although since it is deleted it will not be shown on X even if its status is Helpful.
Notes marking posts as "potentially misleading" with a Note Helpfulness Score of 0.40 and above earn the status of Helpful. At this time, only notes that indicate a post is "potentially misleading" and earn the status of Helpful are eligible to be displayed on posts. Notes with a Note Helpfulness Score less than -0.05 -0.8 * abs(noteFactorScore) are assigned Not Helpful, where noteFactorScore is described in Matrix Factorization.

Additionally, notes with an upper confidence bound estimate of their Note Helpfulness Score (as computed via pseudo-raters) less than -0.04 are assigned Not Helpful, as described in Modeling Uncertainty. Notes with scores in between remain with a status of Needs more Ratings.

Identifying notes as Not Helpful improves contributor helpfulness scoring and reduces the time contributors spend reviewing low quality notes. We plan to enable Helpful statuses for notes marking posts as "not misleading" as we continue to evaluate ranking quality and utility to users.

When a note reaches a status of Helpful / Not Helpful, they're shown alongside the two most commonly chosen explanation tags which describe the reason the note was rated helpful or unhelpful. Notes with the status Needs More Ratings remain sorted by recency (newest first), and notes with a Helpful or Not Helpful status are sorted by their Helpfulness Score.

This ranking mechanism is subject to continuous development and improvement with the aim that Community Notes consistently identifies notes that are found helpful to people from a wide variety of perspectives.
Rating statuses are computed at periodic intervals, so there is a time delay from when a note meets the Helpful / Not Helpful criteria and when that designation appears on the Community Notes site. This delay allows Community Notes to collect a set of independent ratings from people who haven't yet been influenced by seeing status annotations on certain notes.

Helpful Rating Mapping

When rating notes, contributors answer the question "Is this note helpful?" Answers to that question are then used to rank notes. When Community Notes (formerly called Birdwatch) launched in January 2021, people could answer

“yes” or “no” to that question. An update on June 30, 2021 allows people to choose between “yes,” “somewhat” and “no.” We map these responses to continuous values from 0.0 to 1.0, hereafter referred to as “helpful scores”:

- Yes maps to 1.0
- Somewhat maps to 0.5.
- No maps to 0.0.
- Yes from the original 2-option version of the rating form maps to 1.0.
- No from the original 2-option version of the rating form maps to 0.0.

Specific values in the mapping may change in the future, and will be updated here.

Matrix Factorization

The main technique we use to determine which notes are helpful or unhelpful is matrix factorization on the note-rater matrix, a sparse matrix that encodes, for each note and rater, whether that rater found the note to be helpful or unhelpful. This approach, originally made famous by Funk in the 2006 Netflix prize recommender system competition, seeks a latent representation (embedding) of users and items which can explain the affinity of certain users for certain items [1] [2]. In our application, this representation space identifies whether notes may appeal to raters with specific viewpoints, and as a result we are able to identify notes with broad appeal across viewpoints.

One challenge is that not all raters evaluate all notes - in fact most raters do not rate most notes - and this sparsity leads to outliers and noise in the data. Regularization techniques are a common solution to these issues; a key distinction in our approach is that we use much higher regularization on the intercept terms, which capture the helpfulness of a note or rater that is not explained by viewpoint agreement, relative to the embedding factors. This encourages a representation that uses user and note embeddings to explain as much variation in the ratings as possible before fitting additional note- and user-specific intercepts. As a result, for a note to achieve a high intercept term (which is the note’s helpfulness score), it must be rated helpful by raters with a diversity of viewpoints (factor embeddings). Notes are given a single global score-- their intercept term-- rather than using this algorithm in the traditional way to personalize content as in a recommender system.

Matrix factorization identifies notes that are liked by people who normally disagree by assigning high intercepts to notes when users with different factors rate the note Helpful. In cases where there are few ratings available from a given part of the factor space, matrix factorization is forced to estimate the note intercept based on the limited data available. Unfortunately, acting based on limited rating signal can result in volatility in the note intercept as new ratings arrive, leading to note status changes.

To reduce the likelihood that notes are shown to users as Helpful and then change status, we require notes to have at least 5 ratings from positive factor users and 5 ratings from negative factor users before a model can assign Helpful status to a note. Notes that otherwise meet the Helpful standard but do not have enough ratings are highlighted to contributors with a special note preview UI treatment to encourage more ratings. Once the note has enough ratings and meets all other Helpful status criteria, the model will rate the note as Helpful.

Modeling Uncertainty

While the matrix factorization approach above has many nice properties, it doesn't give us a natural built-in way to estimate the uncertainty of its parameters. We take two approaches to model uncertainty:

Pseudo-rating sensitivity analysis

While the matrix factorization approach above has many nice properties, it doesn't give us a natural built-in way to estimate the uncertainty of its parameters. One approach that we use to help quantify the uncertainty in our parameter estimates is by adding in "extreme" ratings from "pseudo-raters", and measuring the maximum and minimum possible values that each note's intercept and factor parameters take on after all possible pseudo-ratings are adding. We add both helpful and not-helpful ratings, from pseudo-raters with the max and min possible rater intercepts, and with the max and min possible factors (as well as 0, since 0-factor raters can often have outsized impact on note intercepts). This approach is similar in spirit to the idea of pseudocounts in Bayesian modeling, or to Shapley values.

We currently assign notes a "Not Helpful" status if the max (upper confidence bound) of their intercept is less than -0.04, in addition to the rules on the raw intercept values defined in the previous section.

Supervised confidence modeling

We also employ a supervised model to detect low confidence matrix factorization results. If the model predicts that a note will lose Helpful status, then the note will remain in Needs More Ratings status for up to an additional 180 minutes or until the supervised model predicts the note will remain rated Helpful. The additional time in Needs More Ratings status allows the note to gather a larger set of ratings. If after that time the note still meets Helpful standards based on the matrix factorization scoring, the note will be rated Helpful and shown on X. In all cases, the final status of the note is determined by matrix factorization. The maximum effect of the supervised model is no more than a 180 minute delay, and all notes will receive a minimum 30 minute delay to gather additional ratings. This helps reduce instances of notes briefly showing and then returning to Needs More Ratings status.

The training data for the supervised confidence model includes all notes that meet the criteria for Helpful status at some point in time. Notes that ultimately lose Helpful status are treated as positives, and notes that retain Helpful status are treated as negatives.

Since ratings are often a reflection of both a note and the associated post, we also include ratings from other notes on the same post. We refer to other notes on the same post as peer notes. For example, if a note already has many peer notes that have been frequently assigned the Note Not Needed tag, the tag may reflect the associated post and all proposed notes may experience similar rating outcomes. During feature extraction, we vectorize features for peer notes suggesting the post is misleading and not-misleading separately since the meaning of the rating changes depending on agreement between the note and the post.

The model uses logistic regression to predict note status outcomes. The feature vectorization process involves both discretization and feature crosses, yielding a sparse representation and allowing the model to learn non-linear relationships. The model is calibrated to delay Helpful status for no more than 60% of notes that ultimately stabilize to Helpful status.

Complete Algorithm Steps:

Prescoring

1. Pre-filter the data: to address sparsity issues, only raters with at least 10 ratings and notes with at least 5 ratings are included (although we don’t recursively filter until convergence). Also, coalesce ratings made by raters with high post-selection-similarity.

2. For each scorer (Core, CoreWithTopics, Expansion, ExpansionPlus, and multiple Group and Topic scorers):
    a. Fit matrix factorization model, then assign intermediate note status labels for notes whose intercept terms (scores) are above or below thresholds.
    b. Compute Author and Rater Helpfulness Scores based on the results of the first matrix factorization, then filter out raters with low helpfulness scores from the ratings data as described in Filtering Ratings Based on Helpfulness Scores.
    c. Fit the harassment-abuse tag-consensus matrix factorization model on the helpfulness-score filtered ratings, then update Author and Rater Helpfulness scores using the output of the tag-consensus model.

Scoring

1. Load the output of step 2 above from prescoring, but re-run step 1 on the newest available notes and ratings data.
2. For each scorer (Core, CoreWithTopics, Expansion, ExpansionPlus, and multiple Group and Topic scorers):
    a. Re-fit the matrix factorization model on ratings data that's been filtered by user helpfulness scores in step 3.
    b. Fit the note diligence matrix factorization model.
    c. Compute upper and lower confidence bounds on each note's intercept by adding pseudo-ratings and re-fitting the model with them.
3. Reconcile scoring results from each scorer to generate final status for each note.
4. Update status labels for any notes written within the last two weeks based on the intercept terms (scores) and rating tags. Stabilize helpfulness status for any notes older than two weeks.
5. Assign the top two explanation tags that match the note's final status label as in Determining Note Status Explanation Tags, or if two such tags don't exist, then revert the note status label to "Needs More Ratings".

A complete detailing of the algorithm and all aspects of the Community Notes system, including all the publicly available data can be found at: https://communitynotes.x.com/guide/en/about/introduction

Q7) Clarity on whether videos of a Nazi salute are permissible on X under current Rules.

While we cannot evaluate a generalized hypothetical, under the X Rules, you may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.

X's mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right – we believe that everyone has a voice, and the right to use it. Our role is to serve the public conversation, which requires representation of a diverse range of perspectives.

We recognize that if people experience abuse on X, it can jeopardize their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature, and more harmful.

We are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized. For this reason, we prohibit behavior that targets individuals or groups with abuse based on their perceived membership in a protected category.

We prohibit targeting individuals or groups with content that references forms of violence or violent events where a protected category was the primary target or victims, where the intent is to harass. This includes, but is not limited to media or text that refers to or depicts: genocides, (e.g., the Holocaust); or lynchings. We prohibit inciting behavior that targets individuals or groups of people belonging to protected categories.

We prohibit targeting others with repeated slurs, tropes or other content that intends to degrade or reinforce negative or harmful stereotypes about a protected category. In some cases, such as (but not limited to) severe, repetitive usage of slurs, or racist/sexist tropes where the context is to harass or intimidate others, we may require Post removal. In other cases, such as (but not limited to) moderate, isolated usage where the context is to harass or intimidate others, we may limit Post visibility as further described below.

We prohibit the dehumanization of a group of people based on their religion, caste, age, disability, serious disease, national origin, race, ethnicity, gender, gender identity, or sexual orientation.

We consider hateful imagery to be logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, gender identity or ethnicity/national origin. Some examples of hateful imagery include, but are not limited to: symbols historically associated with hate groups; images depicting others as less than human, or altered to include hateful symbols, e.g., altering images of individuals to include animalistic features; or images altered to include hateful symbols or references to a mass murder that targeted a protected category, e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust.

Media depicting hateful imagery is not permitted within live video, account bio, profile or header images. All other instances must be marked as sensitive media. Additionally, sending an individual unsolicited hateful imagery is a violation of this policy.

You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

Q8) A visual representation of the algorithm that X uses to recommend content to users on the ‘For You’ pages. The committee would like to know if there have been any updates to the following aspects since X published parts of its code in 2023:

- How the graph (network) of people and content is formulated, and which algorithms are used to construct it.
- How communities are formed within the network, what attributes are considered, and how.
- How the content is sourced for ranking for each individual’s feed.
- What kind of ranking algorithms are used, what attributes they use for ranking, and the relative weights given to each of these attributes.
- The content recommendation measures used.
- How X filters for malicious content.

X aims to deliver you the best of what's happening in the world right now. 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 your device's For You timeline.

Our recommendation system is composed of many interconnected services and jobs, which we detail below. While there are many areas of the app where posts are recommended—Search, Explore, Ads—we will focus on the home timeline's For You feed as requested.

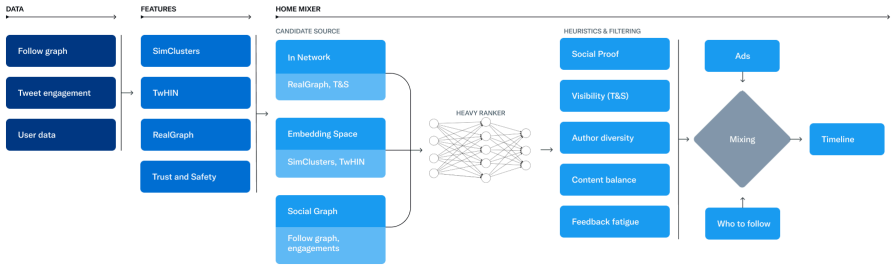
The foundation of X's recommendations is a set of core models and features that extract latent information from Post, user, and engagement data. These models aim to answer important questions about the X network, such as, *"What is the probability you will interact with another user in the future?"* or, *"What are the communities on X and what are trending posts within them?"* Answering these questions accurately enables X to deliver more relevant recommendations.

The recommendation pipeline is made up of three main stages that consume these features:

1. Fetch the best posts from different recommendation sources in a process called candidate sourcing.
2. Rank each Post using a machine learning model.
3. Apply heuristics and filters, such as filtering out posts from users you've blocked, NSFW (or 'not safe for work') content, and posts you've already seen.

The service that is responsible for constructing and serving the For You timeline is called Home Mixer. Home Mixer is built on Product Mixer, our custom Scala framework that facilitates building feeds of content. This service acts as the software backbone that connects different candidate sources, scoring functions, heuristics, and filters.

This diagram below illustrates the major components used to construct a timeline:



The below paragraphs explore the key parts of this system, roughly in the order they'd be called during a single timeline request, starting with retrieving candidates from Candidate Sources.

Candidate Sources

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 you follow (In-Network) and from people you don't follow (Out-of-Network). Today, 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.

In-Network Source

The In-Network source is the largest candidate source and aims to deliver the most relevant, recent posts from users you follow. It efficiently ranks posts of

those you follow based on their relevance using a logistic regression model. The top posts are then sent to the next stage.

The most important component in ranking In-Network posts is Real Graph. Real Graph is a model which predicts the likelihood of engagement between two users. The higher the Real Graph score between you and the author of the Post, the more of their posts we'll include.

Out-of-Network Sources
Finding relevant posts outside of a user's network is a trickier problem: *How can we tell if a certain Post will be relevant to you if you don't follow the author?* X takes two approaches to addressing this.

Social Graph
Our first approach is to estimate what you would find relevant by analyzing the engagements of people you follow or those with similar interests.

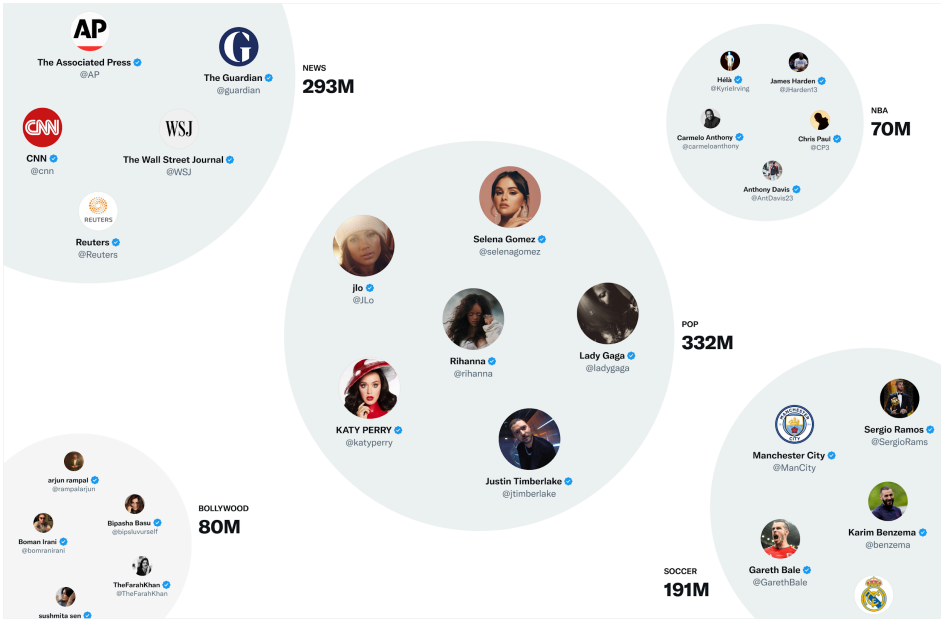
We traverse the graph of engagements and follows to answer the following questions:

> *What posts did the people I follow recently engage with?*
> *Who likes similar posts to me, and what else have they recently liked?*

We generate candidate posts based on the answers to these questions and rank the resulting posts using a logistic regression model. Graph traversals of this type are essential to our Out-of-Network recommendations; we developed GraphJet, a graph processing engine that maintains a real-time interaction graph between users and posts, to execute these traversals. While such heuristics for searching the X engagement and follow network have proven useful (these currently serve about 15% of Home Timeline posts), embedding space approaches have become the larger source of Out-of-Network posts.

Embedding Spaces
Embedding space approaches aim to answer a more general question about content similarity: *What posts and Users are similar to my interests?* Embeddings work by generating numerical representations of users' interests and posts' content. We can then calculate the similarity between any two users, posts or user-Post pairs in this embedding space. Provided we generate accurate embeddings, we can use this similarity as a stand-in for relevance.

One of X's most useful embedding spaces is SimClusters. SimClusters discover communities anchored by a cluster of influential users using a custom matrix factorization algorithm. There are 145k communities, which are updated every three weeks. Users and posts are represented in the space of communities, and can belong to multiple communities. Communities range in size from a few thousand users for individual friend groups, to hundreds of millions of users for news or pop culture. These are some of the biggest communities:



We can embed posts into these communities by looking at the current popularity of a Post in each community. The more that users from a community like a Post, the more that Post will be associated with that community.

Ranking

The goal of the For You timeline is to serve you 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 your 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.

Heuristics, Filters, and Product Features

After the Ranking stage, we apply heuristics and filters to implement various product features. These features work together to create a balanced and diverse feed. Some examples include:

- Visibility Filtering: Filter out posts based on their content and your preferences. For instance, remove posts from accounts you block or mute.
- Author Diversity: Avoid too many consecutive posts from a single author.
- Content Balance: Ensure we are delivering a fair balance of In-Network and Out-of-Network posts.
- Feedback-based Fatigue: Lower the score of certain posts if the viewer has provided negative feedback around it.
- Social Proof: Exclude Out-of-Network posts without a second degree connection to the Post as a quality safeguard. In other words, ensure someone you follow has engaged with the Post or follows the Post's author.
- Conversations: Provide more context to a Reply by threading it together with the original Post.
- Edited posts: Determine if the posts currently on a device are stale, and send instructions to replace them with the edited versions.

Mixing and Serving
At this point, Home Mixer has a set of posts ready to send to your 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 your 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 you perceive on the app.

The goal of our open source endeavor is to provide full transparency to our users about how our systems work. As the Committee has noted, we released the code powering our recommendations, which you can view here (and here) to understand our algorithm in greater detail.

X is the center of conversations around the world. Every day, we serve over 150 billion posts to people's devices. Ensuring that we're delivering the best content possible to our users is both a challenging and an exciting problem. We continue to work on new opportunities to improve our user experience, including by refining our recommendation systems and our content moderation practices, in order to better serve the public conversation and build the digital town square of the future.

Thank you again for the opportunity to provide evidence before the Committee.

Sincerely,

Wifredo "Wifi" Fernández
Global Government Affairs
X Corp.