**Wifredo "Wifi" Fernández**
Global Government Affairs

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

24 April 2026

*Via Email*

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

**Re: Follow-ups from March Committee hearing**

**Q1: All witnesses to write with figures and analysis of if there is more or less misinformation on your platform – according to your definition – compared with a year ago.**

In serving the public conversation and as the internet's global town square, X focuses on helping people find reliable and credible information. We seek to ensure the authenticity of activity on X, and empower users to engage critically with the information that they encounter. This is achieved principally through the relevant X Rules policies, detection and enforcement systems, account badges, and Community Notes.

***X Rules***

The X Rules describe which content and conduct is permitted on X, and how violative content and conduct may be enforced. This includes the 'Authenticity' policy, which stipulates that it is prohibited to engage in inauthentic activity that undermines the integrity of X. Prohibited activities include:

- Operating automated or scripted accounts that do not comply with X's Developer Policy (i.e. 'Unauthorised automation');
- Manufacturing identities to engage in disruptive or deceptive behaviour (i.e. 'Fake personas');
- Impersonating other identifies to deceive others (i.e. 'Impersonation');
- Coordinated inauthentic activity that artificially influences conversations or disrupts X (i.e. 'Multiple Accounts and Coordination');
- Circumventing X enforcement actions (i.e. 'Ban evasion');
- Unauthorised access or modification of someone's X account (i.e. 'Account compromise');
- Sharing or posting content in a bulk, duplicative, irrelevant or unsolicited manner that disrupts people's experience (i.e. 'Content Spam');
- Inauthentic use of X engagement features to artificially impact traffic or disrupt people's experience (i.e. 'Engagement Spam');
- Engaging in scam tactics to obtain money, property or private information (i.e. 'Scams');
- Sharing inauthentic media, including manipulated or out-of-context media, that may result in widespread confusion on public issues, impact public safety, or cause serious harm (i.e. 'Synthetic and Manipulated Media').

The consequences for violating the Authenticity policy depend on the severity of the violation, as well as any previous history of violations. Potential enforcement actions include anti-spam challenges, restricting reach, temporary loss of access to X features or products, profile modifications, and/or suspension. Learn more here.

Additionally, the 'Civic Integrity' policy addresses misleading content that could affect participation in elections or other civic processes. This includes content that may confuse people about how, when, or where to participate, or that aims to discourage participation. The policy addresses three categories of misleading behaviour and content, namely:

- Misleading information about how to participate (i.e. advancing verifiably false or misleading information about how to participate in an election or other civic process);
- Suppression (i.e. advancing verifiably false or misleading information about the circumstances surrounding a civic process intended to intimidate or dissuade people from participating in an election or other civic process); and
- Intimidation (i.e. engaging in or promoting behaviours that may coerce others to refrain from participating in a civic process).

Users can report content that violates the 'Civic Integrity' policy through the in-app reporting flow, located alongside individual items of content. These reports are routed to a request for a Community Note, and Community Notes contributors can then add helpful context to ensure that people are well informed. See further detail below on Community Notes.

In certain cases, X may apply a Civic Integrity label to posts that violate the Civic Integrity policy. This label is reserved for critical escalations. Posts that receive this label will have their reach restricted by X, and display a label informing both the author and viewers that visibility has been restricted. Authors can appeal the label if they believe that their post was restricted in error.

***Premium Organisations and the Grey Checkmark***

We add visual identity signals such as labels and checkmarks to account profiles to provide more context about, and to help distinguish, different types of accounts. With X Premium Organisations, we are creating the most trusted place on the internet for government entities to verify their affiliations and engage with citizens. The Premium Organisations subscription enables government entities of all sizes – from local municipalities to national agencies and multilateral organisations – to manage their certification, and affiliate and verify any related account. Any eligible government entity or multilateral organisation that purchases a subscription will receive a grey checkmark and square avatar, denoting that the account represents an official government or multilateral organisation or individual. Affiliated accounts receive an affiliate badge; i.e. a small image of their patent organisation's profile picture, displayed on their profile.

***Community Notes***

Community Notes is a community-driven system that adds helpful, informative context to posts that may contain misleading information. Notes are written and rated by contributors, not by X. Notes are not used to enforce the X Rules, do not represent X's views and cannot be edited or modified by X's teams. When

enough contributors with diverse perspectives rate a note as helpful, it appears directly on the post.

***Please see the following data:***

Accounts suspended (Global)

- 1 October, 2023 - 30 September, 2024
    - ‘Misleading and Deceptive Identities’: 1.3M.
    - ‘Platform Manipulation and Spam’: 465.9M.
- 1 October, 2024 - 30 September, 2025
    - ‘Impersonation’: 1.3M.
    - ‘Multiple Accounts and Coordination’: 282.6M.

Community Notes – Number of posts with a CN

- Year 2024: 574,303 posts
- Year 2025: 705,555 posts

Community Notes – Number of contributors in the UK

- Year 2024: Over 76,000
- Year 2025: Over 71,000

**Q2: All witnesses to write with improvements that have been made to internal governance systems.**

***Compliance governance***

X maintains a multi-layered governance framework to ensure that online safety decisions are made with appropriate oversight, accountability, and transparency. The X Internet Unlimited Company (***“XIUC”***) Board is the ultimate body responsible for owning, assessing and providing management direction on matters related to ensuring compliance with UK legal requirements. The Board is advised by members of the Safety, Privacy and Legal functions, the Independent Compliance Function (***“ICF”***), and other advisors from X business teams as appropriate.

***Day-to-day content moderation***

X also maintains relevant governance systems in relation to its day-to-day content moderation operations. Our 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.

We engage in content moderation through a combination of automated detection (using machine learning models, natural language processing, image processing, and heuristics) and/or human review to proactively identify and act on violations of the X Rules, and reactively act on violations of the X Rules following user reports. Further, we have enabled UK users and affected persons to report on UK illegal content and content harmful to children, and take action on such reports in accordance with UK law.

Enforcement actions include content removal, account suspensions, visibility restrictions, and soft measures like labels or warnings. Our moderation

workflows follow an ""information-first"" approach, which helps to reduce bias and increase consistency by requiring moderators to reach an enforcement decision by answering defined questions. Appeals systems are commonly available where affected users would like to context content moderation actions.

*Automated content moderation systems*

X employs automated means through a hybrid system combining machine learning models, heuristics and large language models (***"LLM"***) to detect and moderate content violating the X Rules and policies. These models output confidence scores or violation probabilities, with parameters set to balance precision and recall—typically high-confidence thresholds trigger automatic actions, while lower-confidence flags route to human review.

Prior to launch, automations undergo statistical sampling and item-by-item human review to confirm acceptable precision levels; post-launch, parameters are dynamically monitored for anomalies (e.g., spikes in volume, changes in appeal/overturn rates), with adjustments by engineering/data science teams. For high-priority harms (e.g., CSE/terrorism), parameters prioritise proactive, near-immediate action using tools such as hash-matching and keyword blocking to minimise exposure.

*Ensuring accuracy*

We use a variety of performance metrics to monitor performance in detecting violative content, which may include – but is not limited to – volume, false positive rate, appeal rate, appeal overturn rate, user reports and more. For example, we have feedback loops that monitor the performance of our automated detection systems, including the rate at which human moderators agree with automated decisions.

Reviewers have expertise in the applicable policies and are trained by our policy specialists to ensure the reliability of their decisions. Human review helps us to confirm that these automations achieve a level of precision, and sizing helps us understand what to expect once the automations are launched.

Thresholds for automated action are set according to the severity of the violation and remediation. We set precision targets that balance the rate of potential false positives with the ease of overturning the enforcement decision as well as the severity of missing violative content. This strategy holds both before we launch a new control, and when we reassess the effectiveness of an existing measure.

Given the inherent risks of false positives and false negatives, our automation systems are monitored dynamically for ongoing performance and health. If we detect anomalies in performance (for instance, significant spikes or dips against the volume we established during sizing, or significant changes in user complaint/overturn rates), our Engineering teams - with support from other functions - revisit the automation to diagnose any potential problems and adjust the automations as appropriate.

**Q3: X to follow up with what guardrails are in place to prevent Grok being used to facilitate violence against women and girls.**

**Content moderation and enforcement**

*X Rules*

Under the X Rules, it is prohibited to engage in behaviours or share content which may constitute violence against women and girls. This includes engaging in child sexual exploitation, sharing non-consensual nudity, sharing unwanted sexual content or engaging in unwanted sexual objectification, sharing, or threatening to share, private information without permission, making violent threats, wishes of harm, or inciting violence, directly attacking other people with hateful conduct on the basis of protected characteristics (including gender), and targeting others with abuse or harassment or encouraging other people to do so. Any content shared on X, whether organic or AI-generated (including content generated by Grok and shared by users on X, or shared by xAI's @Grok account) must comply with - and is subject to enforcement under - the above rules. This may include account suspension and/or content removal.

Users can report content and accounts which they believe to violate the X Rules, and X also variously deploys automated systems to proactively detect and enforce violations.

*UK illegal content*

X has implemented measures under the UK OSA to enable users to report UK illegal content, including CSAM, intimate image abuse, and controlling or coercive behaviours. X takes action to remove UK illegal content from the platform once identified.

**User controls**

X provides users with privacy and interaction controls that enable them to take steps to protect themselves from potential harmful interactions on the platform. This includes:

- *Protected posts*: When a user switches on 'Protected Posts' (which is automatically switched on for U18 accounts), their posts and replies will only be visible to their followers. They will receive a request when a new user wants to follow them, which they are able to approve or deny. Learn more.
- *Block*: Users can restrict specific accounts from following, direct messaging, or otherwise engaging with them. Learn more.
- *Controlling replies*: Users can choose who is able to reply to their posts. The default position for public accounts is that anyone can reply, but options are available to turn off all replies or only allow accounts mentioned in the post to reply. A user can also change who can reply to their posts, or turn off replies, after they have published a post. Learn more.
- *Discoverability settings*: In order to help users make connections with people on X that they already know, X may use the email address and phone number associated with their account to make their account discoverable to others (e.g. if someone allows X to suggest accounts to follow based on contacts held on their device, X may suggest accounts that have a phone number or email address from that person's contacts). However, users can control whether their account will be discoverable to others based on their phone number or email address by adjusting their discoverability settings, with this setting switched *off*

by default. Even if the discoverability setting is switched on, a user's email address and phone number are not publicly displayed on X. Learn more.

**xAI technical safeguards – Grok models**

Further, X understands that xAI has designed and deployed comprehensive, multi-layered safeguards to mitigate potential misuse of its models. This includes xAI's overarching Frontier Artificial Intelligence Framework (***"FAIF"***), which serves as the core governance structure for identifying and mitigating risks across model development, deployment, and release. Risks are evaluated based on model behaviours, including abuse potential, and tested and measured according to relevant benchmarks, such as the NIST's AI Risk Management Framework, ISO/IEC 420001, and Frontier Model Forum best practices. Relevant safeguards applied to manage risks include:

- *Safety training:* Models are trained by xAI to recognise and decline harmful requests.
- *System prompts:* Models are provided with high-priority instructions to enforce a basic refusal policy (i.e. a policy which sets out when to decline requests that indicate an intent to cause serious harm).
- *Input and output filters:* xAI applies classifiers to user inputs or model outputs to ensure safety when a model is queried regarding certain risk categories.

**Q4: X to follow up on how much revenue is made from users under the age of 18.**

This is business sensitive information.

**Q5: X API Access for Researchers**

From social science to computer science, UK researchers may advance nearly any research objective on topics as diverse as the global conversations happening on X. Researchers may also use X data to conduct scientific studies that solve problems to impact the mission of a non-profit organization or lab.

The X API gives you programmatic access to X's public conversation. The following endpoints are available on pay-per-use plans: Posts, Users, Direct Messages, Spaces, Lists, Likes, Trends, Media, Community Notes, News, and Compliance. We've recently overhauled the legacy system and pricing structure which historically made it difficult for researchers to conduct smaller studies or projects that required more flexible usage. Costs scale with your actual usage, with no monthly caps and less restrictive rate limits.

More information can be found at: https://developer.x.com/

Thank you for the opportunity to provide evidence and look forward to an ongoing dialogue on these important matters.

Sincerely,

**Wifredo "Wifi" Fernández**
Director
Global Government Affairs
X Corp.