23 February 2024

Ofcom
Riverside House
2a Southwark Bridge Road
London
SE1 9HA

Via Email

Re: Illegal harms consultation

Dear Ofcom,

X believes that freedom of speech and platform safety can and must coexist. X is reflective of real conversations happening in the world, and that sometimes includes perspectives that may be offensive or controversial to others. While we welcome everyone to express themselves on X, we have clear rules in place to protect the safety of the service and the people using it.

Our Terms of Service and Rules - which are regularly reviewed by X's safety specialists and are informed by feedback from the people who use X - help ensure that everyone using X feels safe expressing themselves. In the UK, X is committed to complying with the Online Safety Act.

Causes and impacts of online harms

As the global town square, X promotes global standards to combat online harms, particularly those impacting children. Last year, X suspended 12.4 million accounts for violating our Child Sexual Exploitation (CSE) policies. While our efforts to combat CSE continue, global enforcement was vital to our success. We therefore encourage Ofcom to continue collaborating with global regulators and industry to encourage multilateral standards that combat online harms effectively.

X welcomes Ofcom's efforts to provide thorough guidance that helps X combat online harms. Below, we make some additional suggestions.

- **Combating unconscious bias**: X strives to implement content moderation policies that actively prevent unconscious bias, fostering fairness and respect in the online community. To achieve this, X's internal content review policies set objective criteria against which agents can judge potentially illegal content. X welcomes the detailed nature of Volume 5 Illegal Content Judgements Guidance ("Volume 5") and Annex 10 Guidance on Judgement for Illegal Content ("Annex 10"), but due to the broad interpretation of the 'state of mind' or 'mental element', may lead to subjective decisions. X recommends updating Ofcom's Annex 10 to include objective criteria for agents

conducting mens rea analysis (e.g., questions that agents can answer themselves to point them to an objective decision, clear examples of offences). This will mitigate the risk of unconscious bias in decision-making processes.

- [✂]

- **Hateful conduct:** We appreciate Ofcom's research on hateful conduct and radicalization. Ofcom's guidance suggests that recommender systems geared toward maximizing user engagement could exacerbate hate crime and direct susceptible users to less-regulated social media platforms where hateful ideologies are more prevalent. X is taking steps towards combating content that does not violate UK law, but does contravene X's terms of service through its Freedom of Speech, Not Reach policy.[4] By applying a "restricted reach" label to content that may potentially violate X policies, X transparently reduces the discoverability of content. A restricted reach label removes users' ability to engage with the content and the content's reach is restricted to views occurring directly on the author's profile. Restricted reach labels are not in use for all policies, and is part of a broader enforcement toolkit to ensure we're applying

[1] https://help.twitter.com/en/rules-and-policies/glorifying-self-harm
[2] https://help.twitter.com/en/using-x#blocking-and-muting
[3]https://help.twitter.com/en/using-x/x-lists#:~:text=X%20Lists%20allow%20you%20to,by%20group%2C%20topic%20or%20interest.
[4]https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy

appropriate, proportionate and timely intervention strategies. We continue to remove posts that target specific individuals with hateful content and we directly hear from the target, and also suspend accounts that repeatedly harass individuals. Our restricted reach labels were initially only applied to Hateful Conduct, but we have since expanded application to our Abuse & Harassment, Civic Integrity, and Violent Speech Policies. For the avoidance of doubt, our Freedom of Speech, Not Reach policy is an enforcement action against our own policies, and does not replace our legal obligations to withhold access to illegal content. We would welcome opportunities to discuss our Freedom of Speech, Not Reach policy in more detail with Ofcom.

Governance

We believe that governance is key to platform safety. However, governance must be sufficiently flexible to accommodate different business types and recognise the diverse size and nature of technology companies. Many platforms that Ofcom will be regulating, including X, operate cross-border and therefore, we encourage Ofcom to collaborate with global regulators to encourage global governance standards that combat user harm effectively and consistently.

Services' risk assessment, auditing and record-keeping

X advocates for enhanced collaboration between industry stakeholders and global regulators to develop effective risk assessments and auditing. Given that online harms experienced by users transcend geographical boundaries, a global approach to risk assessments ensures a responsive strategy to emerging threats, particularly those posed by AI, while promoting efficiency across platforms. Effective risk assessments require input from senior managers with practical experience in addressing online harms, necessitating a balance between their time spent combating illegal activities and conducting risk assessments. Mutual recognition of safety risk assessments by regulators, including Ofcom, is essential to prevent excessive duplication of tasks associated with risk assessments and record keeping, allowing for greater focus on combating user-facing harms.

We would encourage Ofcom to consider the Risk Assessment requirements contained in the Digital Services Act and to look to align with these so far as reasonably possible.

Illegal Content Codes of Practice

X welcomes Ofcom's efforts in providing certainty surrounding platforms' obligations to review content for potential illegality. However, the guidance applies the same standard of review for non legally binding requests to remove content (e.g. from law enforcement, regulators) with legally enforceable court orders to remove content. For added legal certainty, the guidance would benefit from explaining the factors that platforms should take into account to discharge their freedom of expression obligations when considering non-legally binding requests to remove content (e.g. requests from law enforcement authorities and regulators). Platforms are not required to perform the same type of freedom of expression analysis where they are legally compelled to remove content by a court order.

We appreciate Ofcom's thorough work in outlining all relevant offences in Annex 10. However, we anticipate that deciding illegality in relation to some offences will be challenging for agents given the highly nuanced nature of some of these offences, and only able to see one platform's piece of a potentially much larger pattern of offender behaviour , potentially leading to overenforcement against content and stymying freedom of expression. At a headline level, we suggest: consolidating Volume 5 with Annex 10 to drive more consistency in decision making; adding examples of known offences to Annex 10 to encourage greater consistency in decision-making; and inserting objective criteria to combat subjective decision making and unconscious bias. We also suggest a graduated enforcement process as Ofcom iterates on the Annex 10 guidance available to platforms, thereby encouraging Ofcom's stated aim of voluntary compliance. Some examples to further illustrate our recommendation include:

- [✂]
- **Threatening/abusive behaviour which is likely to cause fear, alarm and abusive behaviour**: Assessing content's potential to cause fear or alarm based on a "reasonable person" standard, as outlined in Annex 10, can be highly subjective and prone to over-enforcement. To mitigate against biassed decision-making due to challenges in predicting psychological or emotional impacts, X urges Ofcom to establish clearer thresholds for determining illegal content and behaviour.
- **False Communications** is an offence that could apply to many different types of communication on platforms. We would encourage Ofcom guidance to be updated to provide examples of clear False Communications offences.

Automated Moderation

We welcome Ofcom's efforts in supporting automated moderation. We're 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. Global multi-stakeholder programs and initiatives have successfully supplemented our automated systems combatting CSE and terrorist content; we are hopeful that similar, multilateral collaboration can be deployed to combat other types of harmful content. We urge Ofcom to collaborate with industry stakeholders and regulators around the world to create common automated moderation standards that are globally scalable to protect global users consistently.

Implementation timeline

We urge Ofcom to grant platforms adequate time for implementing their guidance. We believe that ensuring platforms have ample time is crucial for fostering Ofcom's enforcement strategy of voluntary compliance. While we appreciate Ofcom's comprehensive guidance and engagement, it requires thorough analysis post-finalization. We understand that Ofcom's implementation roadmap may change, especially due to the upcoming general election. Presently, the roadmap indicates platforms must implement guidance upon parliamentary approval and we suggest that Ofcom should review these timelines in light of the length and complexity of the guidance.

We remain committed to participating in Ofcom's future consultations and any Ofcom initiatives that help support user safety.

Sincerely,

[✂]

[✂]

[✂], X