Basic Online Safety Expectations

Summary of response from X Corp. (Twitter) to eSafety's transparency notice on online hate

January 2024

Australian Government | eSafetyCommissioner

eSafety.gov.au

Contents

1. Executive Summary 2

2. Glossary 4

3. Key Notice Findings 6

4. Information about the Notice 15

4.1 The Basic Online Safety Expectations 15

4.2 Online hate 15

4.3 Deciding to give the Notice 16

4.4 What questions did eSafety ask? 17

4.5 What was the Notice process? 17

4.6 What information has been published, and what has been excluded? 18

4.7 What happens next? 18

5. Non-compliance with the Notice and action taken by eSafety 19

5.1 Finding of non-compliance 19

6. Transparency: X Corp.'s response 20

6.1 Questions about resources and expertise to ensure safety, including enforcing Twitter's hateful conduct policy. 20

6.2 Questions about ensuring enforcement of terms of service and policies reflect harms faced by vulnerable groups 23

6.3 Questions about steps Twitter has taken to detect online hate on its service 30

6.4 Questions about steps Twitter has taken to respond to user reports of online hate 37

6.5 Questions about steps Twitter has taken to enforce its terms of use and policies to prevent hateful conduct, including in relation to Twitter Blue accounts 38

6.6 Questions about steps Twitter has taken to manage the risk of reinstating previously banned or suspended accounts for breaches of terms of use and policies to prevent online hate 47

6.7 Questions about preventing the amplification of hateful conduct 50

1. Executive Summary

On 21 June 2023, the eSafety Commissioner (**eSafety**) issued a non-periodic reporting notice (**the Notice**) to X Corp. in relation to the Twitter service (subsequently renamed as 'X'). The Notice) required X Corp. to report on Twitter's compliance with applicable Basic Online Safety Expectations (**the Expectations**) with respect to online hate. eSafety received a response from X Corp. on 17 August 2023, after granting two extensions to the due date at the provider's request.

The Notice was given in accordance with section 56(2) of the Online Safety Act 2021 (**the Act**). Under the Act, eSafety can publish summaries of the information received through the Notice. eSafety's exercises these statutory powers in order to provide transparency and increase accountability from industry.

The aim of this report is to improve transparency and accountability by providing better information about what X Corp. is actually doing – or not doing – to keep Australians safe on the Twitter social media service, and to incentivise improvements to safety standards.

Amongst other factors, eSafety gave this notice to X Corp. as eSafety received more complaints about online hate on Twitter than any other service in the 12 month period of May 2022 to May 2023, with an increase in the number of complaints about hate on Twitter since its acquisition in October 2022.[1]

X Corp. was asked specific questions about the tools, policies and processes Twitter has to prevent online hate on its platform, how it enforces its hateful conduct policy, and the resources it has in place to ensure safety.

[1] eSafety administers complaint-based schemes under the Act to address harmful material online. This includes the Adult Cyber Abuse Scheme which operates as a safety net to help remove cyber-abuse material targeting an Australian adult. Australian adults targeted by online hate can report the online hate to eSafety under the Adult Cyber Abuse Scheme. This legislative scheme does not specifically cover online hate but some forms of online hate may also amount to adult cyber abuse and fall within eSafety's remit. When a complaint is received under the scheme, eSafety will assess the material against the criteria of cyber-abuse material under the Act before considering what options are available to help targeted adult.

The Commissioner noted when the Notice was given that:

> "There is nothing new about hate speech or the misguided perceptions of superiority that drive it. It has been used for centuries to devastating effect, attacking people on the basis of race, belief or identity.
>
> What is new are the channels by which it can now be disseminated to millions in an instant via social media, often compared to the 'digital town square.'
>
> I am deeply concerned about the growing prevalence of online hate. eSafety's own research shows that nearly 1 in 5 Australians have experienced some form of online hate. This level of online abuse is already inexcusably high, but if you're a First Nations Australian, disabled or identify as LGBTIQ+ you experience online hate at double the rate of the rest of the population.
>
> Without transparency about how Twitter's own rules to prevent online hate are set and enforced, there is a risk that bad actors will continue to run rampant on the platform, leaving already marginalised voices abused, sidelined and ultimately silenced."

eSafety closely considered X Corp.'s response to the Notice and found that X Corp. did not comply with the Notice by the deadline including extensions granted. X Corp. provided answers in some instances that were incorrect, significantly incomplete or irrelevant. However, X Corp. provided relevant information after the Notice deadline, and this was taken into account in deciding upon the appropriate enforcement action.

eSafety has given X Corp. a service provider notification, confirming its non-compliance. Details are outlined in chapter 5. eSafety will continue to use the full range of powers available to eSafety to ensure transparency and deter providers from not complying with a statutory notice.

2. Glossary

- **The Act**: The Online Safety Act (Cth) 2021.
- **Artificial intelligence (AI)**: Generally considered to be digital technology which has the capability to exhibit human-like behaviour when faced with a specific task. AI can rapidly process data and spot patterns enabling these tools, in the context of this report, to support the moderation of online content.
- **Automated tools**: Technology used to sort data into categories automatically. In the context of this report, these tools are used to support content moderation actions and decisions.
- **Downranking**: Content shared by an account is given lesser priority by ranking algorithms, so it is less visible to other users.
- **Median time to respond**: The time it takes for a service to respond to a user report, measured from the point at which a report is made to the service to the time that the service takes a content moderation action, such as downranking content or requiring its removal.
- **Online hate**: eSafety defines online hate as any kind of online communication that attacks, discriminated, insults or uses hateful language against a person or group based on their race, religion, ethnicity, sexual orientation, disability or gender. Online hate can be 'intersectional' – meaning a person may be attached for more than one of these characteristics.
- **Hateful Conduct**: In the context of this report, hateful conduct refers to Twitter's hateful conduct policy. This states '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... 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.'
- **Reinstating an account**: Re-activating accounts that had previously been suspended or banned from use or put in 'read only' mode as a result of breaching terms of use or conduct policies.
- **Strikes**: A record on an account that it has previously breached a service's rules. A certain number of strikes may lead to an intervention by the service, such as an account ban, or downranking of content shared by the account.
- **The Voice referendum**: A referendum was held in October 2023 about whether to change the Australian Constitution to recognise the First Peoples of Australia by establishing an Aboriginal and Torres Strait Islander Voice to Parliament.

- **Trusted Flagger**: A user or organisation which meet certain criteria defined by the service. Trusted flagger status means that the user/organisations reports about breaches of a service's terms of service are generally prioritised.
- **Tweet**: A Tweet refers to a public message posted on the Twitter service, which appears on the sender's profile page and timeline as well as in the timeline of anyone who is following the sender, and may be promoted to other users. This post may contain text, photos, a GIF, and/or a video.
- **Twitter's account amnesty:** On 25 November 2022, former CEO of Twitter Mr. Elon Musk, announced a general amnesty to reinstate the accounts of users that had previously been banned for breaches of Twitter's terms of use – excepting those that had 'broken the law or engaged in egregious spam.'[2]
- **Volumetric attacks**: High volume, sometimes cross-platform, attacks on an individual, sometimes coordinated; also known as 'pile-on attacks'.
- **Websites dedicated to hateful conduct**: Websites with the primary purpose to disseminate material that would breach Twitter's hateful conduct policy.

[2] Elon Musk [@elonmusk], 'Twitter will be forming a content moderation council with widely diverse viewpoints. No major content decisions or account reinstatements will happen before that council convenes', 29 October 2022, accessed 7 June 2023, URL: https://twitter.com/elonmusk/status/1586059953311137792

6.6.4 Conducting safety risk assessments before reinstating accounts previously banned and/or suspended that have breached Twitter's rules and policies

X Corp. was asked whether any safety risk assessments were conducted before reinstating accounts previously banned and/or suspended for breaches of Twitter's rules and policies in order to understand and mitigate risks from 28 October 2022 to 31 May 2023[53].

X Corp. responded 'yes' to this question.

X Corp. was asked what safety risks were identified in the risk assessment. X Corp. stated:

> Any accounts that were withheld based on legal requests were not subject to or available for amnesty and an assessment was conducted to identify any accounts that were withheld based on a legal request. Further, before we rolled out the amnesty program, the relevant teams conducted a harm analysis of the violation types and high severity violations were not included in the program. These included policies where there was a risk of physical harm (CSAM, NCN, violent speech, privacy violations etc.).
>
> Our efforts on amnesty were based on the principle of rehabilitation and providing these accounts opportunities to engage in public conversations. Note that these accounts were never exempt from any future violations. Reinstatement was part of the broader effort to promote the right to free expression especially where we believed our approach to content moderation (which historically has had more binary decisions of "take it down" or "leave it up" which led to suspensions after three strikes) was disproportionate.

In response to questions about its engagement with any external organisations or individuals when drafting the safety risk and impact assessment, X Corp. responded that Twitter had not engaged external organisations for this purpose.

6.7 Questions about preventing the amplification of hateful conduct

6.7.1 Reducing the amplification of Tweets/Posts that violate Twitter rules and policies

In the Notice eSafety referred to the fact that, in a 17 April 2023 blogpost, Twitter Safety stated that Twitter would be 'restricting the reach of Tweets[54] that violate our policies by making the content less discoverable.[55]'

---

[53] This question sought information on whether any safety risk assessments were conducted before reinstating accounts previously banned and/or suspended for breaches of Twitter's rules and policies in order to understand and mitigate risks since Twitter's acquisition on 27 October 2022.

[54] A Tweet refers to a public message posted on the Twitter service, which appears on the sender's profile page and timeline as well as in the timeline of anyone who is following the sender. This message may contain text, photos, a GIF, and/or a video.

[55] Twitter, 'Freedom of speech, not reach: An update on our enforcement philosophy', 17 April 2023, accessed 13 June 2023, URL: https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy

considered automated (bots), Twitter stated that it does not specifically isolate automated accounts for action.

X Corp. was asked what metrics it used to assess whether it was successfully reducing the number of bot accounts involved in breaches of Twitter's Hateful Conduct policy; what proportion of the total number of bot accounts removed were found to have breached Twitter's Hateful Conduct policy; and what proportion of bot accounts removed were found to have breached broader Twitter rules and policies on safety.

X Corp. responded that bots fall under Twitter's Platform Manipulation and Spam policy and that it does not cross assess these accounts for hateful conduct because if an account is malicious, it is permanently suspended from the platform.

6.7.5 Toxic tweet impressions

In the Notice, eSafety referred to the fact that Twitter had published analysis with third party service Sprinklr that reported that 'toxic tweets received 3 times fewer impressions'[62] than non-toxic tweets. eSafety noted that Sprinklr and Twitter had stated that the analysis drew upon 300 English language 'slur words' provided by Twitter, which were 'designed to capture hateful slurs and language that targets marginalized and minority voices.'[63]

X Corp. was asked what the 300 words were that Twitter provided to Sprinklr to assess whether hateful slurs and language received fewer impressions on Twitter than other content.

X Corp. stated that it was following up with its partner, and that it would provide an update in due course.

Following a subsequent question from eSafety, providing X Corp. with a further opportunity to provide the information required by the Notice, X Corp. provided the full list of 302 words it had sent to Sprinklr to conduct the analysis.

eSafety has reviewed the list and can confirm that it is comprised of terms that would generally be regarded to be hateful slurs and language. eSafety has not audited Sprinklr's analysis.

---

[62] Twitter Safety, 'Our focal metric is hate speech impressions, not the number of Tweets containing slurs. Most slur usage is not hate speech, but when it is, we work to reduce its reach. Sprinklr's analysis found that hate speech receives 67% fewer impressions per Tweet than non-toxic slur Tweets', 22 March 2023, accessed 15 June 2023, URL: https://twitter.com/TwitterSafety/status/1638262034864263188?lang=en

[63] Sprinklr, 'How Sprinklr helps identify and measure toxic content with AI', 21 March 2023, accessed 15 June 2023, URL: https://www.sprinklr.com/blog/identify-toxic-content-with-leading-analytical-ai/; Twitter Safety, 'We recently partnered with '@Sprinklr for an independent assessment of hate speech on Twitter, which we've been sharing data on publicly for several months. Sprinklr's AI-powered model found that the reach of hate speech on Twitter is even lower than our own model quantified', 22 March 2023, accessed 15 June 2023, URL: https://twitter.com/TwitterSafety/status/1638255718540165121?s=20; Yoel Roth [@yoyoel], 'Over the last 48 hours, we've seen a small number of accounts post a ton of Tweets that include slurs and other derogatory terms. To give you a sense of scale: More than 50,000 Tweets repeatedly using a particular slur came from just 300 accounts.', 30 October 2022, accessed 13 June 2023, URL: https://twitter.com/yoyoel/status/1586542286342475776?s=20

eSafety has decided not to publish the list of words to avoid the information being misused.

6.7.6 Terms used to calculate 'hate speech impressions' metric

In the Notice, eSafety referred to the fact that former CEO of Twitter, Mr. Elon Musk, has referred to reviewing a list of terms used to calculate a 'hate speech impressions' metric[64].

X Corp. was asked to provide a list of the terms used to calculate this metric.

In its response to the Notice, X Corp. did not provide a list of terms used to calculate its hate impressions metric. X Corp. stated that keywords and hashtags are part of the challenges Twitter sought to address through its policies and enforcement efforts.

Following a subsequent question from eSafety, providing X Corp. with a further opportunity to provide the information required by the Notice, X Corp. stated that the list of terms used to calculate 'hate speech impressions' was the same list of words provided to Sprinklr to conduct its analysis.

[64] Elon Musk [@elonmusk], 'Yeah, these are umm ... bad words. I read through the list last week & have to say I learned a few things', 24 November 2023, accessed 13 June 2023, URL: https://twitter.com/elonmusk/status/1595635085172162565?s=20

eSafety.gov.au