Basic Online Safety Expectations

Summary of response to non-periodic notice issued to X Corp. (Twitter) in June 2023

Key findings

Focus: Online hate

Overview

Focus: Online hate

On 21 June 2023 eSafety issued a legal notice (**the Notice**) under the Online Safety Act 2021 (the **Act**) to X Corp. requiring it to detail the steps it was taking to address online hate on the Twitter service (rebranded as 'X' on 23 July 2023), and enforce its hateful conduct policy. The Act empowers eSafety to require online service providers to provide specific information as to what steps they are taking to meet Australia's Basic Online Safety Expectations. X Corp. was required to provide responses to questions for the period 24 January 2022 – 31 May 2023 (the **Report Period**). This document highlights some of the key findings from X Corp.'s response to the Notice.

eSafety assessed X Corp.'s response and determined that it did not comply with the Notice. Subsequent information was provided by X Corp. after the Notice deadline that did seek to rectify earlier omissions of information provided. eSafety took this into account in deciding upon the appropriate enforcement action. X Corp. was given a Service Provider Notification confirming its non-compliance.

The full report contains additional information and context for these key findings and sets out eSafety's decisions in relation to X Corp's non-compliance with the Notice. It is available at eSafety.gov.au.

BOSE Summary of response to non-periodic notice issued to X Corp. (Twitter) in June 2023

Key findings - Online hate

Twitter's policy on hateful conduct

eSafety asked questions about Twitter's hateful conduct policy, which 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.'*

Twitter's hateful conduct policy states that policy violations can result in a range of enforcement actions such as making content less visible; excluding Tweets/posts or accounts in email or in-product recommendations; requiring a Tweet/post to be removed; and suspending accounts that violate the policy.**

Key findings from X Corp.'s response to the Notice

Twitter staffing and safety resources

Adequate resourcing of trust and safety functions is important to ensure online safety. Based on eSafety's observations over the past eight years of online safety regulation, companies with low numbers of trust and safety personnel may have reduced capacity to respond to online hate, as well as other online harms. The result is that the burden for safety tends to fall on the user or group experiencing the abuse, rather than the platform taking responsibility for harmful content and conduct on their service.

X Corp. provided the following information in response to questions about Twitter's safety staffing, before and after its acquisition in October 2022:

Category of staff asked about in the Notice	Number of staff 27 October 2022 (day before Twitter's acquisition)	31 May 2023 (end of the notice reporting period)	Staffing change between 27 October 2022 and 31 May 2023 by percentage
Engineers focussed on trust and safety issues globally	279	55	**-80%**
Trust and safety staff dedicated to hateful conduct issues globally	0	0	**0%**
Trust and safety staff globally (employees and contractors)	4062	2849	**-30%**
Trust and safety staff in the APAC region (employees and contractors)	111	61	**-45%**
Trust and safety staff in Australia (employees and contractors)	1	0	**-100%**
Content moderators globally	FTE: 107 Contractors: 2613	FTE: 51 Contractors: 2305	**FTE: -52% Contractors: -12%**
Content moderators in the APAC region	FTE: 39 Contractors: 2014	FTE: 27 Contractors: 1868	**FTE: -31% Contractors: -7%**
Public policy staff globally	68	15	**-78%**
Public policy staff in the APAC region	15	4	**-73%**
Public policy staff in Australia	3	0	**-100%**

In answer to a question about whether Twitter had staff dedicated to hateful conduct issues, **X Corp. stated there were no full time staff that are specifically and singularly dedicated to hateful conduct issues globally, and no specific team for this policy. It said that instead, a broader cross-functional team has this in scope and collaborates on a set of policies that are related to toxicity more broadly.**

Issues impacting Australian communities

A recent survey[a] by eSafety of Australian adults found that 18% of Australians aged over 18 said they had experienced online hate speech at least once in the past 12 months, including:

- 36% of people identifying as LGBTIQ+
- 35% of Aboriginal and Torres Strait Islander people
- 25% of people who speak a language other than English at home
- 25% of people with a disability.

Hate speech can be highly contextual. X Corp. was asked if Twitter received formal or informal advice or insights from organisations or individuals representing communities targeted by hateful conduct about the ways in which these harms were perpetuated online. **X Corp. stated that it did not maintain a record, but said it participated in 'conversations on this topic on a global stage', including in relation to the European Union's Code of Conduct on countering illegal hate speech online, ARCOM's Online Hate Observatory and the Singapore Government's IMDA/MCI consultations regarding online safety.**

Online hate targeting First Nations peoples in Australia has increased on a range of services. A recent eSafety study found that First Nations youth are three times more likely to experience hate speech online than their non-indigenous counterparts.[b]

'during the specific notice period there was no formal engagement with any First Nations organisations'

X Corp. added that Twitter had previously had engagement with a wide range of First Nations organisations and individuals over many years.

eSafety referred in the Notice to the fact that, on 28 October 2022, Elon Musk tweeted that Twitter would form a content moderation council with 'widely diverse viewpoints'. He noted that 'no major content decisions or account reinstatements' would happen before the council convenes.'[c]

In its response to a question in the Notice, **X Corp. confirmed that Twitter's Trust and Safety Council was disbanded in December 2022, and that the company had not replaced the Trust and Safety Council with another advisory body for taking advice from external experts from diverse backgrounds on matters relating to the safety of users, including hateful conduct.**

Detecting material or activity covered by Twitter's Hateful Conduct policy

Under the Basic Online Safety Expectations, providers are expected to take reasonable steps to proactively minimise the extent of unlawful or harmful material and activity on a service.

X Corp. provided the following information about whether it used any automated tools[d] to detect material or activity covered by Twitter's policy on hateful conduct:

Automated tools used to detect hateful conduct in **Tweets**:	**Yes**
Automated tools used to detect hateful conduct in **direct messages**:	**No**

18

The total number of direct messages that Twitter identified involving hateful conduct during the Report Period. These were all identified from user reports.

BOSE Summary of response to non-periodic notice issued to X Corp. (Twitter) in June 2023
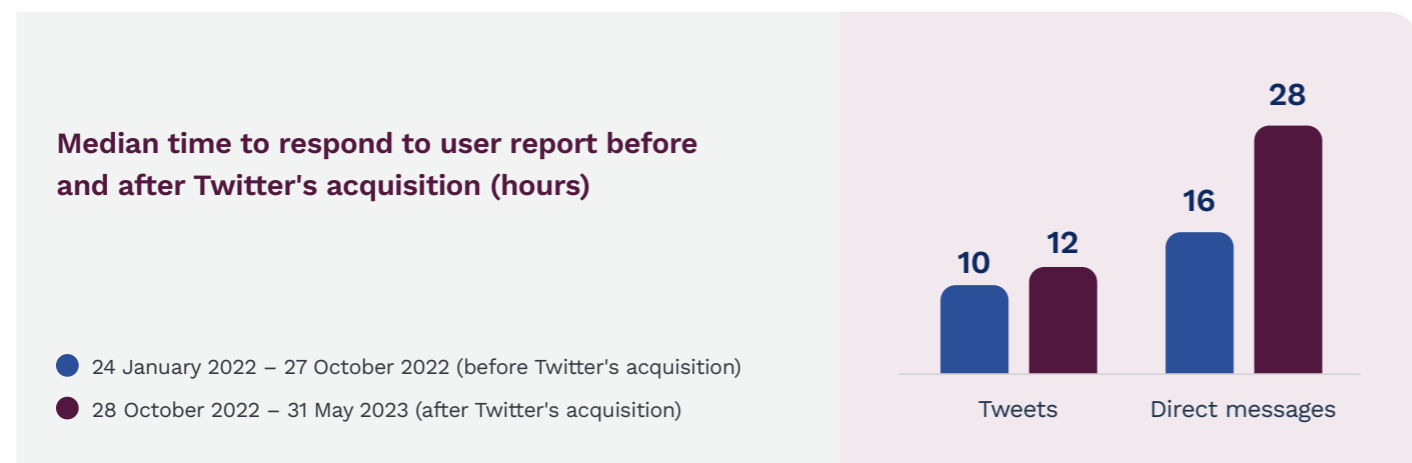
Key findings - Online hate

In addition to relying on user reports and automated tools, many online services use 'Trusted Flagger' programs to prioritise reports of terms of use or policy breaches from verified external experts, and/or prevent access to certain sites that breach a service's terms of service in response to reports from such experts.

X Corp. provided the following information regarding Twitter:

Uses a Trusted Flagger program for prioritising reports of hateful conduct:	**No**
Block links (URLS) that direct to websites dedicated to harmful content:	**No**

Median time to respond to user reports

X Corp. was asked to provide Twitter's median time to respond to user reports of hateful conduct.

Median time to respond to user report before and after Twitter's acquisition (hours)

● 24 January 2022 – 27 October 2022 (before Twitter's acquisition)
● 28 October 2022 – 31 May 2023 (after Twitter's acquisition)

The response indicates that since Twitter's acquisition there has been a 20% increase in median time to respond to user reports about Tweets and a 75% increase in the median time to respond to direct messages. eSafety notes that prompt action on user reports is particularly important given that Twitter solely relies on user reports to identify hateful conduct in direct messages.

eSafety also asked about the proportion of user reports of hateful conduct that were determined by Twitter to be hateful conduct, and whether this changed following Twitter's acquisition. X Corp. provided the following information regarding reports in Australia:

	Total number of reports of hateful conduct	Number of reports of hateful conduct that Twitter determined had breached its hateful conduct policy	Proportion of reports that Twitter determined had breached its hateful conduct policy of all hateful conduct reports received
January 2022 – 27 October 2022			
Reports from all users	~ 865,000	~ 7,400	0.86%
28 October 2022 – May 2023			
Reports from all users	~ 830,000	~ 6,200	0.75%

eSafety understands from this information that Twitter does not consider that the vast majority of user reports of hateful conduct breach its terms and policies. From X Corp.'s response, there was no significant difference in how reports were treated before and after its acquisition.

Twitter Blue and amplification

> The Centre for Countering Digital Hate has reported that Twitter failed to act on 99% of content involving 'Twitter Blue' accounts that CCDH considered to be hate, and reported to the service.[e]

eSafety sought to ascertain whether Twitter's enforcement of its terms of service differed depending on the user. **X Corp. stated that its enforcement policy applied to Twitter Blue subscribers in the same way as other accounts for breaches of Twitter's hateful conduct policy.**

In response to a question regarding how tweets from Twitter Blue accounts are treated, **X Corp. stated that no accounts are artificially or manually amplified**.

eSafety asked whether Twitter conducted tests to its recommender systems with the goal of reducing the risk that hateful conduct is amplified. eSafety provided examples such as internal audits, external audits, risk and impact assessments, and a/b testing.

Tests to recommender systems conducted to reduce risk of amplification of hateful conduct	**No**

[e]Center for Countering Digital Hate, 'Twitter fails to act on 99% of Twitter Blue accounts tweeting hate', 1 June 2023, URL: https://counterhate.com/research/twitter-fails-to-act-on-twitter-blue-accounts-tweeting-hate/#about

Prevalence and reach of hateful conduct

The Institute for Strategic Dialogue (ISD) reported in March 2023 that it had found:

‘In total, analysts detected 325,739 English-language antisemitic Tweets in the 9 months from June 2022 to February 2023, with the weekly average number of antisemitic Tweets increasing by 106% (from 6,204 to 12,762), when comparing the period before and after acquisition’[f]

In June 2023, researchers from the University of Southern California presented a paper to the 2023 International AAAI Conference on Web and Social Media which reportedly quantified an increase in the volume of online hate in the months following Twitter's acquisition. Using machine learning analysis, researchers reportedly extracted samples of users that posted online hate and found that **‘the proportion of hate words in hateful tweets increased’** after the acquisition.[g] The researchers reportedly found that the **‘average daily online hate of hateful users nearly doubled.’**[h]

Twitter's enforcement policy includes the ‘Freedom of Speech not Reach’[#] approach, which it states involves labelling and making less visible violative content that does not meet the threshold it sets for content removal. In response to a question from eSafety to describe changes to Twitter's policy on hateful conduct and how data captured was used to inform the changes, X Corp. stated the following:

‘In July 2023, we were able to report that labels have been applied to more than 700,000 violative posts that fall under our Hateful Conduct policy. Compared to a healthy post, posts with these labels – or restricted posts – receive 81% less reach or impressions and we proactively prevent ads from appearing adjacent to content that we label. More than one third of authors proactively choose to delete the Tweet after they are informed that its reach has been restricted. On average, only 4 percent of authors have appealed labels.’

Twitter previously published information which it said demonstrated that toxic tweets[i] had 3 times fewer impressions than non-toxic tweets[j] eSafety asked for the terms that were used to calculate this metric.

Twitter provided eSafety with the 300 terms that it used to calculate toxic tweet impressions. eSafety has confirmed that these terms are broadly associated with hateful conduct.

[f]Institute for Strategic Dialogue, ‘Antisemitism on Twitter Before and After Elon Musk's Acquisition’, 15 March 2023, accessed 15 November 2023, URL: https://www.isdglobal.org/isd-publications/antisemitism-on-twitter-before-and-after-elon-musks-acquisition/ [g]USC Viterbi, ‘New Twitter, Now With More Hate’, 20 April 2023, accessed 15 November 2023, URL: https://viterbischool.usc.edu/news/2023/04/new-twitter-now-with-more-hate/ [h]USC Viterbi, ‘New Twitter, Now With More Hate’, 20 April 2023, accessed 15 November 2023, URL: https://viterbischool.usc.edu/news/2023/04/new-twitter-now-with-more-hate/

[#]Title corrected April 2024. [i]Sprinklr's toxicity model categorises content as toxic if it is used to demean an individual, attack a protected category, or dehumanise marginalised groups. 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/#toc-1 [j]Twitter/X Safety [@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 received 67% fewer impressions per Tweet than non-toxic slur Tweets.’, 22 March 2023, accessed 15 June 2023, URL: https://twitter.com/Safety/status/1638262034864263188

BOSE Summary of response to non-periodic notice issued to X Corp. (Twitter) in June 2023

Key findings - Online hate

Detection of volumetric attacks

eSafety has previously raised concerns with Twitter about volumetric attacks on its service. These are high volume, sometimes cross-platform, attacks on an individual that are sometimes coordinated. They are also known as 'pile-on attacks'.

X Corp. responded as follows to questions relating to automated tools specifically designed to identify volumetric attacks in breach of Twitter's targeted harassment policy:

Automated tools used that are specifically designed to detect volumetric attacks in breach of Twitter's targeted harassment policy:	No

Reinstating accounts previously banned and/or suspended

On 25 November 2022, Twitter announced that it had reinstated accounts that had previously been banned and/or suspended for breaching its rules and policies.[k] Media reports suggested that through what Twitter referred to as the 'general amnesty', Twitter reinstated 62,000 banned accounts, with 75 of those accounts having over 1 million subscribers.[l] In response to the Notice question about reinstated accounts, **X Corp. stated that Twitter performed a risk assessment on all reinstated accounts.**

X Corp. provided the following information on the number of reinstated accounts:

6,103

previously banned accounts were reinstated by Twitter. eSafety understands this to relate to accounts in Australia.

– Of these, 194 accounts were reinstated that were previously suspended for hateful conduct violations.

X Corp. was asked if any of those accounts were put under additional scrutiny given their history of breaching Twitter's rules and policies during the period 25 November 2022 to 31 May 2023, for example if 'fewer strikes' were required before a reinstated account was downranked, put in 'read only' mode, or suspended.[m]

X Corp. responded that Twitter did not place reinstated accounts under additional scrutiny.

[k]Twitter Safety [@TwitterSafety], 'As we shared earlier, we have been proactively reinstating previously suspended accounts. Starting February 1, anyone can appeal an account suspension and be evaluated under our new criteria for reinstatement.', 28 January 2023, accessed 8 June 2023, URL: https://twitter.com/TwitterSafety/status/1619125112716005376?s=20 [l]Platformer, 'Why some tech CEO's are rooting for Musk', November 2022, URL: https://www.platformer.news/p/why-some-tech-ceos-are-rooting-for.

[m]This question sought information about whether accounts suspended and/or banned prior to Twitter's acquisition on 27 October 2022 for breaches of Twitter's policy on hateful conduct were put under additional scrutiny at the time of Twitter's account amnesty, announced on 25 November 2022.

eSafety.gov.au