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

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

3. Key Notice Findings

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:

	Number of staff		
Category of staff asked about in the Notice	**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 focused 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.

[a] The survey explored the attitudes and experiences of over 5,300 Australian adults aged 18 years and over. The survey covered the 12-month period to November-December 2022. eSafety commenced publishing the results of the survey in 2023. Further reports, including one on hate speech, will be published in 2024.

[b] Safety Commissioner, 'Cool, beautiful, strange and scary: The online experiences of Aboriginal and Torres Strait Islander children and their parents and caregivers', March 2023, URL: https://www.esafety.gov.au/research/online-experiences-aboriginal-torres-strait-islander-children-parents-caregivers

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.

[c] 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

[d] Software for detecting potential abuse (such as keyword filters, rules engines, hash matching, as well as Machine Learning or Artificial Intelligence systems) ensuring content and activity is flagged for acting upon

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:

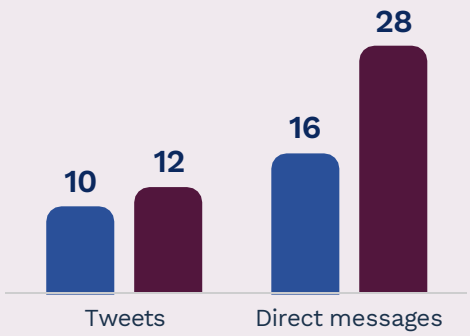
Uses a Trusted Flagger program for prioritising reports of hateful conduct:	**No**
Block links (URLS) that direct to websites dedicated to harmful content:	**No**

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 - 31 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]

[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/

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.

#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

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.**

[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.

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.

[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.

4. Information about the Notice

4.1 The Basic Online Safety Expectations

The Basic Online Safety Expectations Determination 2022 sets out the Australian Government's expectations that social media, messaging, gaming, dating, file sharing services and other apps and websites will take reasonable steps to keep Australians safe. The current determination was registered on 23 January 2022.

Compliance with the Expectations is not mandatory, but eSafety may exercise powers under the Act to require providers to prepare a report on the steps they are taking to meet the Expectations, in the manner and form specified by eSafety and to give that report to eSafety. There are financial penalties as well as non-financial sanctions associated with non-compliance with a Notice.

Further information on the Expectations and associated powers can be found in eSafety's Regulatory Guidance.

4.2 Online hate

eSafety considers online hate to be any kind of online communication that attacks, discriminates, insults or uses hateful language against a person or group based on their race, religion, ethnicity, gender, sexual orientation or disability.

Online hate is a growing issue which can negatively impact a person's mental health, general wellbeing and online engagement. In extreme cases, online hate can lead to harassment and violence offline.

Some online hate cases may be so serious and threatening to an individual that it may meet the criteria of adult cyber abuse as defined in the Act.

For the purposes of the Notice, X Corp. was asked questions about online hate in the context of its own 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[3].

[3] Twitter, 'Hateful Conduct', April 2023, accessed and quote obtained on 8 June 2023, URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

4.3 Deciding to give the Notice

In deciding to give this notice to X Corp. eSafety was required to consider several criteria specified in section 56(5) the Act, including:

- the number of occasions during the previous 12 months on which complaints about material provided on the service were made to eSafety under the Act
- whether the provider had previously contravened a civil penalty provision related to reporting on the Expectations
- whether there are deficiencies in the provider's practices, so far as those practices relate to the capacity of end-users to use the service in a safe manner
- whether there are deficiencies in the service's terms of use, so far as they relate to the capacity of end-users to use the service in a safe manner
- whether the provider had agreed to give the Secretary regular reports relating to the capacity of end-users to use the service in a safe manner
- any other matters the Commissioner considers relevant.

Examples of other matters that eSafety may take into account, as identified in the Regulatory Guidance[4] include:

- a service's reach and the profile of its users, including whether it is used by children
- the measures the provider currently has in place to protect users from harm
- the information already published by a provider and any absence of information regarding a service's safety policies, processes and tools, or limited information about the impact or effectiveness of these interventions
- aggregated evidence from eSafety's other regulatory schemes, such as types of complaints, a provider's responsiveness to removal requests/notices, or other investigative insights regarding service safety issues
- evidence of systemic harm, or evidence of key safety risks, relative to the Expectations, including from victims, charities, media, academics or other experts.

While the criteria above indicates that complaints are relevant to determining the provider(s) to which a notice is given, giving a notice to a particular provider does not, in itself, indicate a level of concern with that provider's compliance with the Expectations. There may be providers with content accessible in Australia which are more, or less, compliant with the Expectations than providers who receive notices.

[4] eSafety Commissioner, 'Basic Online Safety Expectations Regulatory Guidance', updated September 2023, URL: https://www.esafety.gov.au/sites/default/files/2023-09/Basic-Online-Safety-Expectations-Regulatory-Guidance-updated-September-2023_0.pdf

4.4 What questions did eSafety ask?

The Notice required X Corp. to provide information in response to a set of specific questions, using a template provided by eSafety. The questions were a mix of yes and no questions, and questions allowing free text answers or seeking specific data. As noted in the Regulatory Guidance on the Expectations, targeted questions assist both the provider and eSafety. It ensures the provision of meaningful information and minimises the regulatory burden on respondents. eSafety did not seek information that was readily accessible in the public domain.

Through answering the questions, X Corp. provided a report on the specific steps Twitter took to meet the relevant Expectations in relation to online hate.

4.5 What was the Notice process?

X Corp. was given the notice on 21 June 2023, after a short consultation on the draft notice.

X Corp. was invited to discuss with eSafety any questions it had about the Notice, how to respond, or about the scope of the questions.

X Corp. responded to the Notice on 17 August 2023, after eSafety granted two extensions at the provider's request.

Assessment and Follow-up Questions

On receipt of its response to the Notice, eSafety assessed if X Corp. had answered the questions required. Where X Corp. did not provide a response (for example, providing an overarching comment instead of responding to the question), did not provide a full response (for example, not answering each part of the question required by the Notice), or did not respond using the template provided in the Notice, eSafety followed up with X Corp. to:

- Give a further opportunity to provide the information sought by the Notice;
- Understand the reasons why X Corp. was not able to provide the information in answer to a question in the Notice;
- Help eSafety assess X Corp.'s compliance with the Notice.

Where X Corp.'s response was not clear, eSafety followed up to seek clarification of the response and any further information the provider opted to give to provide context.

X Corp. was invited to discuss with eSafety any questions it might have.

Draft Summary Report

X Corp. was given a draft of the summary report relating to its service prior to publication. X Corp. was given the opportunity to make submissions about the information included in the summary. X Corp. was also invited to discuss with eSafety the proposed publication, any

concerns it might have, and any submissions it wished to make. eSafety considered all submissions received from X Corp. to draft this final summary report.

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

This report provides a summary of the information that eSafety received from X Corp. about the Twitter service. It does not reflect X Corp.'s response to the Notice in its entirety. In line with eSafety's Regulatory Guidance, certain information has been withheld where eSafety has considered it was not appropriate to disclose, for example because it was commercial-in-confidence or because publication of the information would not serve the public interest.

In particular, eSafety has determined that it is not in the public interest to publish specific indicators and signals that Twitter uses to detect users seeking to commit crimes and cause harm and to prevent recidivism. eSafety engaged with experts to seek views on what kind of information would not be in the public interest to publish.

The following points should also be noted:

- The information provided in response to the Notice has not been verified by eSafety, although X Corp. was required to respond truthfully and accurately. Information is published in the interests of transparency and accountability.
- The information summarised in this report is based on the response eSafety received, which reflects the practices X Corp. reported that Twitter employed in the period 24 January 2022 to 31 May 2023 inclusive, or other periods within this timeframe as specified. Twitter may have implemented changes to policies and processes since this information was provided.
- Terms used are defined in the glossary to this report, unless otherwise stated.

4.7 What happens next?

- eSafety will raise any specific gaps and vulnerabilities with X Corp. to understand any safety deficiencies and obstacles to compliance with the Expectations.
- eSafety intends that the information in this report is used by researchers, academics, the media and public to improve accountability.
- eSafety's Regulatory Guidance, first published in July 2022 and updated in September 2023, sets out eSafety's planned approach to the exercise of its powers in respect of the Expectations. In the coming months eSafety will:
    - Expand use of non-periodic notices to other acute harms. eSafety welcomes input from all stakeholders on the areas where greater transparency is needed.
    - Issue the first periodic notices to begin tracking key safety issues and progress against them. These are likely to focus on acute harms in the first instance, and potential deficiencies in providers' safety processes.
    - Consider whether to issue statements of compliance and non-compliance with the Expectations now that further regulatory guidance has been published.

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

5.1 Finding of non-compliance

Information is sought through reporting notices to improve transparency and accountability, incentivise improvements in safety standards, and to assist eSafety to determine whether a provider is compliant with the Expectations.

A non-periodic reporting notice requires the provider to prepare a report about compliance with one or more of the Expectations, prepare the report in the manner and form set out in the notice, and to provide it to eSafety.[5] Providers are required to comply with a notice to the extent they are capable of doing so.[6]

When eSafety issued the Notice to X Corp. in June 2023 eSafety provided a response template as part of the Notice, with specific questions tailored to the provider and service related to specific Expectations. The Notice required responses to these questions. eSafety informed X Corp. that it should contact eSafety if it had any questions about the Notice, the information being sought, or how to respond.

X Corp. was required to respond by the deadline set by the Notice, including any extensions granted. eSafety also informed X Corp. that it could request an extension of time to enable it to comply with the Notice. Two extensions were granted to respond at X Corp.'s request.

eSafety found that X Corp. did not comply with the Notice by providing responses to 14 questions in the Notice that were incorrect, significantly incomplete or irrelevant.

Subsequent information was provided by X Corp. after the Notice by the deadline that did seek to rectify earlier omissions of information provided. eSafety took this into consideration into account in deciding upon the appropriate enforcement action.

eSafety has given a service provider notification to X Corp., to confirm its non-compliance by the Notice deadline and to deter it from future non-compliance.

[5] Section 49(2)(b) and (c) and Section 56(2)(b) and (c) of the Act

[6] Section 50 and Section 57 of the Act

6. Transparency: X Corp.'s response

X Corp. was asked questions about its Twitter service. eSafety notes that Twitter was rebranded to 'X' on 31 July 2023. Given the reporting period covered by the Notice predates this rebranding, eSafety refers to the service as Twitter throughout this transparency summary report with the exception of direct quotes where the service provider referred to the service as 'X'.

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

6.1.1 Twitter's Trust and Safety resourcing

X Corp. was asked to provide a breakdown of the people working on trust and safety for Twitter, and the locations they are based in, for the period 1 May 2023 to 31 May 2023.[7]

X Corp. did not provide a breakdown of staff categories or their locations in its response. Instead, it stated that it has approximately 3,000 people working on trust and safety and content moderation issues and that this includes full time employees, contractors and third-party services. X Corp. stated that its trust and safety experts are situated in Twitter office locations and time zones around the world that overlap with Australian time zones.

X Corp. stated:

> X Corp.'s expertise includes Australian and foreign nationals who have the ability to travel to liaise in person, as well as online, with e-Safety, other government and law enforcement agencies, and other organizations. During this period of transition, as it works through various organizational changes, X Corp. has adjusted levels of engagement, but maintained its levels of commitment to ensuring Australians are safe online, and this remains the case.
>
> Its teams continue to consult and learn from key organizations, individuals and information available to help inform and improve its service.
>
> For clarity, Twitter shared that it continues to have employees based in Australia, who are mostly focused on Sales and Marketing duties.

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 following information. eSafety has not published the specific countries that staff are located in, but has provided

[7] This question sought the most up to date information about the number of people working on trust and safety for Twitter, which for the Notice Period, related to May 2023, and the locations they were based in.

aggregate numbers for geographic areas. This provides transparency regarding the extent of Twitter's global coverage, including different time zones.

1 May 2023 – 31 May 2023	**Total Number of employees/contractors/third parties services.**	**Breakdown of employees/contracts/third parties services by location**
Full time employees working on trust and safety at Twitter	229 (as of 31 May 2023)	Australia: 0 APAC: 54 Rest of world: 175; constituting: • North America: 126 • Europe: 49
Twitter contractors working on content moderation.	2430	Australia: 0 APAC: 2,009 Rest of world: 421; constituting: • North America: 101 • Europe: 286 • South America: 34
Twitter third party services working on trust and safety to supplement Twitter's internal trust and safety staff and functions.	59	Australia: 0 APAC: 41 Rest of world: 18; constituting: • North America: 5 • Europe: 13

6.1.2 Number of Twitter staff

X Corp. was also asked to provide the numbers of staff employed or contracted by Twitter for the time periods 24 January 2022 – 27 October 2022 and 28 October 2022 – 31 May 2023, for a number of specific categories.[8] X Corp. did not provide a response to this question, but pointed to the information at section 6.1.1, regarding the number of staff working on trust and safety.

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 following data.

[8] This question sought information on numbers of staff employed or contracted by Twitter before and after its acquisition on 27 October 2022. The date 24 January 2022 corresponds with the start of the Report Period.

Category of staff	**Number of Staff employed or contracted**		
	24 January 2022	**27 October 2022**	**31 May 2023**
Engineers focussed on trust and safety issues globally	277	279	55
Trust and safety staff dedicated to hateful conduct issues globally[9]	0	0	0
Trust and Safety staff globally (employees and contractors)	3317	4062	2849
Trust and Safety staff in the APAC region (employees and contractors)	101	111	61
Trust and Safety staff in Australia (employees and contractors)	1	1	0
Content moderators globally	FTE: 121 Contractors: 2514	FTE: 107 Contractors: 2613	FTE: 51 Contractors: 2305
Content moderators in the APAC region	FTE: 44 Contractors: 1915	FTE: 39 Contractors: 2014	FTE: 27 Contractors: 1868
Public policy staff globally	68	68	15
Public policy staff in the APAC region	13	15	4
Public policy staff in Australia	2	3	0

[9] X Corp. stated it had 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 considered to be related to toxicity more broadly. Similarly its contractors were not dedicated to this specific policy. However, X Corp. provided data on the number of contractors whose workflows 'could have' covered hateful conduct issues. In January 2022 this was 1008; 27 October 2022 it was 949; 28 October 2022 it was 949; and 31 May 2023 it was 544.

6.1.3 Twitter's surge capacity for responding to the risk of hateful conduct around specific events

X Corp. was asked whether it had the ability to move additional internal resources as a surge capacity for responding to the risk of hateful conduct around specific events (e.g. a political event, cultural event, crisis event, or to support specific groups facing a surge in abuse) for the period 28 October 2022 – 31 May 2023[10].

X Corp. responded that it did have this ability and that it had dedicated teams that handle crisis events, ensuring X Corp. was taking necessary actions to mitigate against risks that can occur.

X Corp. also stated:

> Once we become aware of an event, we assess how harmful and imminent the event is and determine the level of our response to ensure the health of the public conversation.

X Corp. stated that its expert teams who focus on threat disruption mainly work proactively, though they also have teams of moderators that work reactively in response to user reports. In addition, X Corp. stated that its Strategic Response team develops crisis or event specific policy and enforcement guidance for content reviewers. The team also proactively searches and detects for violating content to ensure necessary and proportionate measures are taken.

In response to a question about whether these additional internal resources were used to respond to specific events in Australia, X Corp. responded that they were used to respond to escalations and reports in Australia.

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

6.2.1 Using data on groups frequently targeted by hateful conduct to inform policies

X Corp. was asked about its hateful conduct policy[11] and whether it collected or held data on groups frequently targeted by hateful conduct. X Corp. responded that it did not.

X Corp. was also asked whether such data, if any was collected, was used to inform any changes to Twitter's policy on hateful conduct at any stage in the Report Period.

---

[10] This question sought information about whether Twitter had the ability to move additional internal resources as a surge capacity for responding to the risk of hateful conduct around specific events since its acquisition on 27 October 2022.

[11] Twitter, 'Hateful Conduct', April 2023, accessed 8 June 2023, URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

X Corp. stated:

> For clarity, our policies did not significantly change overall during the Reporting Period. There have been several updates. We do not collect data on our users that would allow identification of any groups they belong to.

X Corp. identified that the single largest change it had made during the Report Period was to implement the 'Freedom of Speech Not Reach (FOSNR)' enforcement policy and approach, which includes 'visibility filtering and labelling', from 17 April 2023.

X Corp. further stated:

> Hate speech is often set in social and political context and creates an environment of fear and intimidation experienced by groups of people that are often historically marginalised. Given that X is inherently a public platform, we are sensitive to the risks that hate speech can pose not just at an individual level but at a societal level. At the same time, we are also aware of the risks of censorship and putting undue and unnecessary restrictions on freedom of expression as we build policies and enforcement protocols to address hate speech. Striking the right balance between safety and freedom of expression is not always easy. We believe that any restriction on freedom of expression has to be reasonable, necessary and proportionate. For this reason, the company made a principled decision to move away from its binary take-down/leave-up enforcement approach and invest in visibility filtering as part of the moderation toolkit. We continue to prohibit posts that target specific individuals with hate, abuse and violence, but adopt a more proportionate remediation for posts or content that does not target specific individuals by restricting the reach of such content.
>
> [Freedom of Speech Not Reach] means that users have the right to express their opinions and ideas without fear of censorship, and that we also believe it is our responsibility to keep users on our platform safe from content violating our Rules. The new enforcement philosophy means, where appropriate, restricting the reach of posts that meet the threshold for enforcement according to our terms of service by making the content less discoverable. Restricting this reach of posts – visibility filtering - is one of our existing enforcement actions that allows us to move beyond the binary 'leave up versus take down' approach to content moderation.

X Corp. stated that its enforcement efforts and associated documentation were 'constantly living documents' and were updated to reflect changing behaviours online.

X Corp. provided data on the application of its policies:

> 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. 90 percent of all appeals receive a response within 30 minutes.

X Corp. was asked about what alternative steps it took during the Report Period to ensure that any changes to its policy on hateful conduct provided for the safety of groups targeted by hateful conduct.

X Corp. responded that while it does not gather information on groups frequently targeted by hateful conduct, that it does consider the impact of its policies on the most vulnerable populations. It added that the company's approach to policy and enforcement factors in potential impacts on all human rights, with negative impacts to physical safety, privacy, and freedom of expression being the most significant and most important it seeks to prevent and mitigate.

X Corp. also stated that it has an internal glossary for slurs and tropes that includes all the slurs that target people based on their protected characteristics and that the glossary includes terms that are often used to target marginalised groups. It added that these lists of slurs and tropes are based on both the use of the words themselves as well as the surrounding context. As an example of the ongoing efforts to address hate speech, X Corp. stated that it had made recent additions to the existing antisemitic slurs glossary that it identified in the last few months that were added to both its proactive and reactive enforcement.

X Corp. stated that its abuse and harassment policy prohibits behaviours that encourages others to harass or target specific individuals or groups of people with abusive or harassment online and as well as behaviours that urges offline action such as physical harassment. In addition, its violent speech policy prohibits behaviour that threatens, incites, glorifies, or expresses desire for violence or harm.

6.2.2 Engaging with third-party experts to inform changes in hateful conduct policy

eSafety referred in the Notice to the fact that Twitter's policy on hateful conduct has been amended several times in 2023, and most recently in April 2023[12]. eSafety noted that, in July 2019, when Twitter expanded its policy on hateful conduct to include the dehumanisation of

[12] Twitter, 'Hateful Conduct', April 2023, accessed 8 June 2023, URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

people, it stated that it had 'collaborated with civil society, academics, and third-party experts.[13]'

X Corp. was asked questions about whether it engaged with civil society, academics, and third-party experts, including those targeted by hateful conduct, to inform Twitter's policy on hateful conduct during the period 28 October 2022 to 31 May 2023.[14]

X Corp. responded that it had not consulted with third parties with respect to 2023 policy updates and that it does not necessarily consult with third parties for all policy reviews.

It noted that:

> We clarify, in any case, that a range of organizations such as Human Rights Watch, Article19, and eSafety itself have provided (and continue to provide) reports or advice, insights and information via a full range of means. We have the ability to engage and receive feedback about our policies and enforcement.

It added that information obtained from previous consultations with organisations and individuals on the design and details of policies over the years, including with representatives of marginalised groups in Australia, continues to inform Twitter's policy updates.

X Corp. was asked about what steps it had taken to ensure its policy on hateful conduct protects the safety of end-users targeted by hateful conduct. X Corp. responded that its policies did not change significantly between October 2022 and April 2023[15], other than regarding enforcement actions for hateful conduct violations to ensure proportionality in enforcement.

X Corp. provided an example of this change regarding enforcement action:

> We have created a dedicated Harassment policy whereby we prohibit targeting specific individuals with hateful and abusive content. Such posts will be bounced and accounts may be suspended after their second violation. For accounts dedicated to harassing individuals (i.e. persistent harassment), we will immediately suspend such accounts.

X Corp. added that it had created a standalone Twitter policy on violent speech and that its enforcement is much stricter than it was previously.

6.2.3 Twitter's Trust and Safety Council

In the Notice eSafety referred to the fact that on 8 October 2022, former CEO of Twitter, Mr. Elon Musk, tweeted that Twitter 'will be forming a content moderation council with widely

---

[13] Twitter, 'Updating our rules against hateful conduct'. Accessed 13 June 2023, URL: https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate

[14] This question sought information about Twitter's engagement with civil society, academics, and third party experts, including those targeted by hateful conduct, to inform Twitter's policy on hateful conduct since its acquisition on 27 October 2022 to the end of the Report Period.

[15] The date of changes to Twitter's hateful conduct policy

diverse viewpoints. No major content decisions or account reinstatements will happen before that council convenes'.[16] In the Notice, eSafety also referred to the fact that on 12 December 2022, media reported that Twitter had disbanded its Trust and Safety Council, and noted that Twitter had reportedly stated in an email to former Council members that it had determined that the Council was 'not the best structure' to 'bring external insights into our product and policy development work.[17]'

In response to a question from eSafety seeking confirmation regarding whether Twitter disbanded its Trust and Safety Council, X Corp. confirmed that this body was disbanded in December 2022.

In response to a question about whether Twitter had 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, X Corp. confirmed that it did not have an advisory body in place.

X Corp. pointed to the information provided in sections 6.2.1 and 6.2.2, regarding the safety measures in place. It reiterated that

> [its] policies remain as robust as ever and we have not diluted how we define hateful content, in fact certain definitions have been expanded, and our policies (as currently defined) were originally informed by engagement with groups [mentioned in section 6.2.1].

6.2.4 Engaging with organisations or individuals representing communities targeted by hateful conduct

In response to questions about whether 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 from 13 December 2022 to 31 May 2023[18], X Corp. stated that Twitter did receive such advice.

X Corp. was asked to name the organisations and individuals who provided advice or insights to Twitter between 13 December 2022 and 31 May 2023. X Corp. stated that Twitter participated in 'global conversations' such as in relation to European Union (EU) Code of Conduct on countering illegal hate speech since its inception, and ARCOM[19]'s Online Hate Observatory. It also stated

---

[16] 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 9 June 2023, URL: https://twitter.com/elonmusk/status/1586059953311137792

[17] CNN, 'Twitter disbands its 'Trust and Safety Council' that tackled harassment and child exploitation', 13 December 2022, accessed 9 June 2023, URL: https://edition.cnn.com/2022/12/12/tech/twitter-disbands-trust-and-safety-council/index.html

[18] This question sought information about Twitter's engagements with organisations or individuals representing communities targeted by hateful conduct since its Trust and Safety Council was disbanded.

[19] France's Audiovisual and Digital Communication Regulatory Authority (*Autorité de régulation de la communication audiovisuelle et numérique*)

that Twitter engaged with the Singaporean government's IMDA/MCI consultations regarding online safety.

Following a subsequent question from eSafety, which provided X Corp. with a further opportunity to provide the information required by the Notice, X Corp. stated that there was no formal consultation during the relevant period.

X Corp. stated that it did not maintain a record and stated that it was not reasonably possible for it to collect information listing informal conversations that took place between 13 December 2022 and 31 May 2023. X Corp. stated that it also uses open source material to inform its work but that it does not maintain a list of sources so is unable to provide a list of organisations with which it formally consulted.

6.2.5 Insights into the specific issues faced by Australian communities

X Corp. was asked questions about the steps it took to ensure that Twitter's policy on hateful conduct and its implementation are informed by an understanding of the specific issues faced by Australian communities (including but not limited to First Nations people who face disproportionate abuse online[20]), following the withdrawal of public policy staff from Australia.

X Corp. pointed to Twitter's hateful conduct policy and emphasised that the policy covers content that

> dehumanises, incites fear, incites harassment, incites discrimination, statements of exclusion based on race or ethnicity, use of slurs, tropes and harmful stereotypes, violent event denial (including holocaust denial), content targeted PC [protected categories] with hateful references.

X Corp. also pointed to its Abuse and Harassment policy.

X Corp. added that Twitter has glossaries that include hateful terms from around the world – including Australia, which include slurs and tropes. It said these lists are regularly updated based on trends on the platform and may also be informed based on engagement with external stakeholders.

X Corp. stated:

> These policies and glossaries are informed in a number of ways by an understanding of the specific issues faced by Australian communities e.g. First Nations Australians as diverse, historically disadvantaged, marginalized and underrepresented groups. While their participation on the platform has also been low overall, we have sought over a period of years to ensure our policies are informed and with a broader understanding of the unique risks they may face for example specific contextualized racial abuse.

[20] 'eSafety Commissioner, 'Cool, beautiful, strange and scary: The online experiences of Aboriginal and Torres Strait Islander children and their parents and caregivers', March 2023, accessed 13 June 2023, URL: https://www.esafety.gov.au/research/online-experiences-aboriginal-torres-strait-islander-children-parents-caregivers

In relation to its engagement with First Nations organisations and/or individuals to ensure its policy on hateful conduct and its implementation are informed by an understanding of the specific issues faced by First Nations people, X Corp. stated that it had engaged with a range of stakeholders including Indigenous X, Reconciliation Australia, Black Rainbow and leaders of the Go Foundation, but qualified that this took place 'over the years', which eSafety understands to mean before Twitter's acquisition.

Following a subsequent question from eSafety, providing X Corp. with a further opportunity to provide the information required by the Notice for the period in question, X Corp. stated that it had not formally engaged with any First Nations organisations between when it ceased having public policy/trust and safety staff in Australia (December 2022) to 31 May 2023 but that it had previously had engagement with a wide range of First Nations organisations and individuals over many years.

In relation to additional measures (i.e. beyond the measures normally in place by Twitter in Australia) that it put in place, if any, to reduce the risk of hateful conduct directed at First Nations people during the lead up to the Voice referendum, X Corp. stated that Twitter planned to 'continue key engagements' with the aforementioned stakeholders to ensure that 'information is up to date in the run up to the event'.

X Corp. did not provide a response to questions about precisely when Twitter ceased having public policy and/or trust and safety staff in Australia in its response to the Notice. X Corp. stated that Twitter's content moderation functions were based in Singapore and that some commercial and marketing teams are based in Australia.

X Corp. did not provide the required information in response to a question about when Twitter ceased having public policy and/or trust and safety staff in Australia.

Following a subsequent question from eSafety X Corp. stated that it ceased having public policy staff in Australia on 22 December 2022, when it said all relevant staff no longer had access to internal Twitter systems.

X Corp. was asked about alternative steps it took to ensure Twitter's policy on hateful conduct and its implementation were informed by an understanding of the specific issues faced by Australian communities.

X Corp. stated:

> Broadly speaking, hateful rhetoric and content is often set in social, historical and political context and creates an environment of fear and intimidation experienced by groups of people that are often historically marginalized. Wherever possible, teams at X apply the relevant context before making content moderation decisions. While it is ideal to make contextual assessments, it is also important to highlight that context and intent may not always be obvious to our frontline moderators or our proactive enforcement tools. We strive to get it right, but it can sometimes be hard in the absence of relevant information and context, and we're also aware of the risk of over enforcement or disproportionate enforcement if we allow subjective decision-making in the process.

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

6.3.1 The use of automated tools to detect material or activity covered by Twitter's Hateful Conduct policy

X Corp. was asked whether Twitter used any automated tools[21] to detect material or activity covered by its policy on hateful conduct on the following parts of the service from 1 May 2023 to 31 May 2023[22]. X Corp. provided the following information:

Part of service	Automated tools used to detect material or activity covered by Twitter's Hateful Conduct policy in tweets, but not in direct messages.
Tweets	Yes
Direct messages	No

X Corp. was then asked to name all the tools used, including any proprietary tools and those provided by third parties. In its response to the Notice, X Corp. did not provide a list of tools it used for this purpose.

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 Twitter used tools named Botmaker, Smyte, Bulk Media Enforcement, Proactive ToS Models and Reactive ToS Models to surface violative content. X Corp. also provided the following context.

> We proactively detect content using a combination of models and heuristics, as well as media matching (i.e. comparison of hashes extracted from media for the purpose of comparison with other media uploaded to the platform). When our systems detect content that may violate our terms of service, we either take an action on the content, or send it for human review. In the instance of a 100% match, the uploaded media may be automatically deleted. Ideally we would catch all violative content proactively with automation. However, context matters, and user reports help us identify violations that our systems are unable to detect.
>
> There are a few ways in which we proactively detect content:
>
> - One is by collecting a large volume of violating and non-violating content (which human reviewers have reviewed to be violations based on user or system generated reports), and training a model to identify similar violation content types.

[21] Software for detecting potential abuse (such as keyword filters, rules engines, hash matching, as well as Machine Learning or Artificial Intelligence systems) ensuring content and activity is flagged for acting upon.

[22] This question sought the most up to date information on whether Twitter used any automated tools to detect material or activity covered by its policy on hateful conduct.

- Second is by defining rules/heuristics, such as specific keywords/terms, for proactive detection of content that may contain such words and terms.
- Third, and this is primarily for media detection, is to add representations of media (hashes, embeddings, etc.) to tools and models for proactive detection. This is done primarily This is done primarily for violations such as child sexual exploitation material, non-consensual nudity and intimate imagery, etc.

To mitigate risks associated with our automated detections, they go through a multi-stage approval process before launch to assess their purpose relative to the proposed remediation, guardrails, and checks to both the logic and sample for precision.

X Corp. also pointed to information Twitter had published in its Digital Services Act Transparency Report: transparency.twitter.com/dsa-transparency-report.html

X Corp. was asked whether the automated tools used on Tweets took into account whether the **account that posted the tweet previously engaged in hateful conduct** in violation of Twitter's policies when assessing whether content of the tweet was likely to involve hateful conduct.

X Corp. responded that the automated tool does not consider previous violations or account activity and that the tool applies the appropriate remediation on the post itself.

X Corp. was then asked whether the automated tools used on Tweets took into account whether the account that the material is **directed at has previously been the target of hateful conduct** in assessing whether a tweet is likely to involve hateful conduct.

X Corp. responded that its tools and models are trained to catch violating content irrespective of whether it is targeted at one individual or more.

As this response was not clear, eSafety sought clarification on whether automated tools took into account whether the account that the material is directed at has previously been the target of hateful conduct, X Corp. stated that the automated tools are not specifically designed to take into account whether the account that the material is directed at has previously been the target of hateful conduct.

X Corp. was also asked what other steps were taken to detect material and/or activity covered by Twitter's Hateful Conduct policy in relation to parts of the service where automated tools are not used. Given X Corp.'s answer regarding where automated tools are used, this question related to direct messages.

X Corp. stated that users can report accounts that persistently harass individuals with abuse and hate through its updated harassment policy, which its moderators enforce reactively based on user report (but not proactively). It reported that the Harassment policy was applied at an account level and accounts were able to be immediately suspended if there is evidence of persistent harassment.

eSafety understands from X Corp.'s response that Twitter did not take any other proactive steps to detect material and/or activity covered by its Hateful Conduct policy in direct messages.

X Corp. subsequently added that:

> this is on account of privacy and freedom of speech concerns directly connected to the methodologies that would need to be employed in order to detect such material or activity within direct messages. Users can report posts, profiles, lists, spaces, and Direct Messages for containing hateful conduct, harassment, violent speech etc. When users report direct messages that violate our hateful conduct policy, Twitter responds by stopping the violator from sending messages to the account that reported them. The conversation will also be removed from the reporter's inbox.

6.3.2 Indicators used by automated tools to identify hateful conduct

X Corp. was asked if specific terms (including for example racist slurs and tropes), hateful imagery (for example, the 'Nazi swastika' is cited in Twitter's policy on hateful conduct[23]), or other indicators (for example account behaviour) associated with hateful conduct, were used by automated tools to identify hateful conduct[24].

X Corp. responded that such terms, imagery and indicators were used to identify hateful conduct.

X Corp. was then asked to list all the organisations that these terms, hateful imagery or other indicators were sourced from (including from civil society organisations, governments, law enforcement and others).

In its response to the Notice, X Corp. stated that Twitter's keyword-based detection was sourced from its policy team, which uses diverse sources and empirical examples from the platform to make a judgement about what should be included. X Corp. also stated that while Twitter's sources are from a range of organisations, there were no specific organisations' keyword lists that it adopted directly. X Corp. stated, as an example, that Twitter has drawn hateful symbols, logos and symbols of designated violent/terrorist organisations from the Anti-Defamation League.

Following a subsequent question from eSafety, providing X Corp. with a further opportunity to provide a list of all organisations that were used as a source, X Corp. responded that Twitter does not have an exhaustive list of organisations from which these terms, hateful imagery or other indicators are/were sourced. X Corp. stated that the majority of Twitter's work in this area

[23] Twitter, 'Hateful Conduct', April 2023, accessed 8 June 2023, URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[24] This question sought the most up to date information about whether specific terms (including for example racist slurs and tropes), hateful imagery (for example, the 'Nazi swastika' is cited in Twitter's Hateful Conduct policy, or other indicators (for example account behaviour) associated with hateful conduct, were used by automated tools to identify hateful conduct.

is based on 'internal research and investigations, as well as assessments of any new trends' that it encounters on the platform. In addition, X Corp. stated that Twitter's database of hateful keywords and slurs is based on its in-house policy and enforcement expertise, and feedback from external partners, and has been built over an extended period of time. X Corp. also responded that it may use public databases and research to corroborate its findings and that it has engaged with hundreds of organisations globally and in the region.

X Corp. was asked if these terms, hateful imagery, or other indicators were used by tools on both tweets and direct messages. X Corp. responded with the following information:

Part of service	Terms, hateful imagery, or other indicators used by automated tools to identify hateful conduct
Tweets	Yes
Direct messages	No

X Corp. stated that they proactively send potential hateful symbols in profile banners and avatars to Twitter moderators to review.

X Corp. was asked if any of the terms, hateful imagery, or indicators were blocked before they are shared on the service. X Corp. responded

> We do not block posts before they are shared on service, even if they contain content that falls under hateful conduct. Once shared, if our models catch them, when multiple signals are simultaneously present and above defined precision thresholds, then we would take the appropriate remediation action.

X Corp. was asked to provide information on any other indicators or signals (for example, third party reports; account behavioural signals) was used to identify hateful conduct on Twitter, in addition to terms and hateful imagery.

X Corp. responded that it used a range of account behavioural signals to inform enforcement on Twitter. X Corp. stated that at the content level, Twitter also utilised user reports to identify and remediate hateful content on the platform and that accounts can also be identified by user reports for containing majority hateful content. It stated that accounts would be removed from the platform.

6.3.3 URL blocking

X Corp. was asked if it blocked any uniform resource locators (URLs) that direct to websites dedicated to hateful conduct. X Corp. responded that it did not and provided the following statement:

> X's enforcement philosophy of 'Freedom of Speech Not Reach' permits users to share links to a wide range of websites, while at the same time, also preventing amplification and distribution of those links that X may consider hateful conduct across the platform.

6.3.4 Trusted Flagger program

X Corp. was asked questions about whether Twitter had a Trusted Flagger program from 1 May 2023 to 31 May 2023[25] to prioritise reports of hateful conduct by relevant experts. X Corp. provided the following information:

1 May 2023 – 31 May 2023	Yes/No
Did Twitter have a Trusted Flagger program for prioritising reports of hateful conduct by relevant experts	No

X Corp. also stated that while Twitter currently does not have a dedicated Trusted Flagger program, it was in the process of evaluating and reviewing options for a 'globally scalable framework'.

In response to questions about what other steps it had taken to ensure that reports by organisations with specific expertise are prioritised, X Corp. stated that Twitter has direct relationships with expert organisations around the world and a dedicated alias for non-governmental organisations to raise concerns. X Corp. further stated that this allows Twitter to receive reports from these expert groups which it escalates for review.

6.3.5 Mechanisms used to identify hateful conduct

X Corp. was asked to provide the proportion of hateful conduct determined[26] by Twitter that was identified[27] through the following sources for the period 28 October 2022 to 31 May 2023[28].

[25] This question sought the most up to date information on whether Twitter had a Trusted Flagger program to prioritise reports of hateful conduct by relevant experts.

[26] i.e. Twitter made a decision that the conduct breached its hateful conduct policy

[27] The content was flagged to Twitter from a user report, an automated tool, proactive searching by Twitter moderators, trusted flaggers, or via other means of flagging potentially violative conduct.

[28] This question sought information on the proportion of hateful conduct that Twitter determined was identified by the listed sources on tweets and direct messages since its acquisition on 27 October 2022.

X Corp. provided the following information for Tweets:

Tweets	**%** of total tweets determined by Twitter to be hateful conduct, according to source	**Number** of tweets determined by Twitter to be hateful conduct, according to source	**Total number** of tweets determined by Twitter to be hateful conduct
User reports	25.65%	366.3K	1.4M
Automated tools	71.45%	1.0M	
Twitter moderators proactively searching	2.85%	40.7K	
Other	0.05%	0.7K	

X Corp. was also asked for the proportion of hateful conduct Twitter had identified on direct messages for the period 28 October 2022 to 31 May 2023. In its response to the Notice, X Corp. stated that this data was not available.

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 following data:

Direct messages	**%** of total direct messages determined by Twitter to be hateful conduct according to **identification tool**	**Number** of direct messages determined by Twitter to be hateful conduct according to identification tool	**Total number** of direct messages determined by Twitter to be hateful conduct
User reports	100%	18	18
Automated tools	0%	0	0
Other	0%	0	0

X Corp. clarified that as automated tools are not used on direct messages to identify hateful conduct, there was only information related to user reports. It added that:

> When users report direct messages that violate our hateful conduct policy, we respond by stopping the violator from sending messages to the account that reported them. The conversation will also be removed from the reporter's inbox.

6.3.6 Proportion of reports for breaching Twitter's policy on hateful conduct

X Corp. was asked to provide the proportion of reports of hateful conduct it had determined breached Twitter's policy on hateful conduct in the periods 23 January 2022 – 27 October 2022

and 28 October 2022 – 31 May 2023[29], corresponding with the periods before and after the service's acquisition.

In its response to the Notice, X Corp. provided the total number of user reports, globally. X Corp. stated this was 56.8 million reports.

X Corp. also stated that under its 'freedom of speech not reach' policy it labelled approximately 335,000 tweets for hateful conduct between October 28 2022 - May 31 2023. It stated that this was identified by automated tools.

X Corp. added that visibility filtering labels and enforcement were focused on 'tweet-level behaviour and action'. The total number for the report period, of approximately 324,000 reflects testing and launch phases which it said were ongoing. It quoted the following categories of 'hateful conduct':

Hateful conduct: 324,000
Hateful conduct low severity: 11,000
Grand total: 335,000

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 following data for Australia.

	% of reports made to Twitter that were determined to have breached Twitter's Hateful Conduct policy	**Total number of reports** of hateful conduct	**Number of reports** of hateful conduct that were determined to have breached Twitter's Hateful Conduct policy
23 January 2022 – 27 October 2022			
Reports from all users	0.86%	~865,000	~7,400
28 October 2022 – 31 May 2023			
Reports from all users	0.75%	~830,000	~6,200

6.3.7 Detecting volumetric attacks

In the Notice, eSafety noted that it had previously raised concerns with Twitter about volumetric attacks[30] on the service. X Corp. was asked questions about automated tools specifically designed to identify volumetric attacks in breach of Twitter's Targeted Harassment policy for the period 1 May 2023 to 31 May 2023[31]. X Corp. responded with the following information:

[29] This question sought information on the proportion of reports of hateful conduct Twitter determined breached its policy on hateful conduct before and after its acquisition on 27 October 2022.

[30] High volume, sometimes cross-platform, attacks on an individual, sometimes coordinated; also known as 'pile-on attacks' (see section 10(2)(a) of the Determination).

[31] This question sought the most up to date information on whether Twitter used any automated tools specifically designed to identify volumetric attacks in breach of Twitter's Targeted Harassment policy.

1 May 2023 to 31 May 2023	Yes/No
Did Twitter use any automated tools that are specifically designed to identify volumetric attacks in break of Twitter's targeted harassment policy?	No
Did Twitter have any other proactive measures in place to minimise volumetric attacks and breaches of Twitter's targeted harassment policy? (e.g. prioritised reporting flows for certain accounts; proactive human moderation of specific accounts; and/or events; keyword flags etc.)	No

In response to questions about alternative steps taken to ensure volumetric attacks on Twitter were identified, X Corp. stated that once it is made aware of volumetric attacks, its internal teams monitor the situation actively; share guidance to remediate against the attack as soon as possible; and monitor the situation for developments and amend strategies as developments unfold.

In response to questions about alternative measures that were in place to ensure that X Corp. had policies and processes in place to minimise volumetric attacks, X Corp. referred to Twitter's response above and added that its proactive terms of service models continually look for abusive and hateful content on the platform.

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

6.4.1 Median time to respond to user reports of hateful conduct

X Corp. was asked for Twitter's median time to respond[32] to user reports of hateful conduct for the time periods 24 January 2022 - 27 October 2022 and 28 October 2022 - 31 May 2023[33] on both tweets and direct messages. eSafety specified that if Twitter does not collect 'hateful conduct' as a specific category, that it should respond with reference to the user reporting category on Twitter of 'attacked because of my/their identity[34]'. X Corp. was also asked to outline how the median time was calculated and any other information.

In its response to the Notice, X Corp. stated that it was continuing to compile a response to the question and that it would revert to eSafety as soon as possible.

[32] Time to respond should be 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.

[33] This question sought information about the median time Twitter took to respond to user reports of hateful conduct before and after its acquisition on 27 October 2022.

[34] "Attacked because of my/their identity" are categories of report available to users on Twitter.

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 data below.

Median time taken to respond to user reports of hateful conduct	24 January 2022 – 27 October 2022	28 October 2022 – 31 May 2023
Tweets	10 hours	12 hours
Direct messages	16 hours	28 hours

X Corp. added that it calculated the median time based on all the reports it received that concerned a tweet/post or direct message and have been manually resolved by Twitter agents, measuring the time from the report's creation up to the time a case was closed in its system.

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

6.5.1 Enforcement action by Twitter

eSafety referred in the Notice to the fact that Twitter's enforcement policy states that 'when we determine that a violation is severe enough to warrant tweet removal we will require the violator to remove it before they can tweet again.'[35]

X Corp. provided the following information in response to questions about Twitter's enforcement actions taken in response to breaches of its policy on hateful conduct for the period 12 December 2022 (i.e. when 'new' Twitter Blue launched) to 31 May 2023:

Twitter enforcement action in response to breaches of Twitter policy on hateful conduct	Twitter Blue accounts	Non-Twitter Blue accounts
Number of tweets that Twitter required a user to remove for breaches of its Hateful Conduct policy	17.1K	900.2K
Number of removal requests that were not complied with*	~ 170 out of 1200	~ 14,000 out of 59,000
Number of accounts that were put into 'read only mode' as a result of failure to comply with a removal request	3.1K	198.9K
Number of decisions to put accounts in 'read only mode' that were appealed by the user*	~ 20	~ 33,000
Number of appeals by users that were successful*	0	~ 330
Number of accounts that were suspended for repeat violations of Twitter's Hateful Conduct policy	0.4K	27.3K

[35] Twitter, 'Our range of enforcement options', 2023, accessed 8 June 2023, URL: https://help.twitter.com/en/rules-and-policies/enforcement-options

*In its initial response to the Notice, X Corp. stated that data was not available for these enforcement actions. X Corp. provided this information in response to a subsequent question from eSafety, which provided X Corp. with a further opportunity to provide the information required by the Notice.

6.5.2 Twitter Blue and the enforcement of Twitter's Hateful Conduct policy

X Corp. was asked whether Twitter's enforcement policy applied to Twitter Blue subscribers in the same way as for other accounts for breaches of Twitter's hateful conduct policy.

X Corp. responded 'yes' to this question.

6.5.3 Revoking Twitter Blue status for breaching Twitter's Hateful Conduct policy

X Corp. was also asked whether Twitter Blue status was revoked for accounts that breached Twitter's policy on hateful conduct for the period 1 May 2023 to 31 May 2023[36].

X Corp. responded 'no' to this question.

X Corp. was asked questions about whether it prevented the amplification of content posted by Twitter Blue accounts which had breached its policy on hateful conduct in any other way.

X Corp. stated that any violation of Twitter's terms of service or rules, manipulation of Twitter processes, or circumvention of enforcement actions can result in the loss of the 'checkmark' [the Twitter Blue symbol by an account name] or, in some cases, suspension.

X Corp. added that for any violation that results in remediation short of suspension (restricted reach, removal of tweet), there is no impact on Twitter Blue checkmark status.

X Corp. stated that technically Twitter does not suspend a user's Twitter Blue subscription even in cases where an account is suspended for hateful conduct. It suspends the user's account that is deemed to be in violation of the policies, removes the Blue checkmark, and the user is required to end their subscription. Even though Twitter does not revoke the subscription from its end, Twitter stated that this has the end result of revoking the privileges of Twitter Blue status. It stated this prevented the amplification of content posted by Twitter Blue accounts that have breached Twitter's policy on hateful conduct.

Following a subsequent question from eSafety, seeking clarification of whether Twitter prevented the amplification of content posted by Twitter Blue accounts which had breached its policy on hateful conduct, X Corp. responded that it did prevent the amplification of content posted in violation of its policies by Twitter Blue accounts during the period 1 May 2023 to 31 May 2023. X Corp. stated:

[36] This question sought up to date information on whether Twitter Blue status was revoked for accounts that breached Twitter's policy on hateful conduct.

> When we take action against content that breaches our Terms of Service, at a minimum we apply visibility filtering which has the effect of preventing the amplification of the relevant content. Other measures can also be applied, and we have listed these below. If a post shared by an account – irrespective of its Blue/Premium status – is found to be in violation of our hateful conduct policy, we would apply the applicable remediation measure.
>
> Other remediation measures we can take, all of which would make the content on X less visible on X include:
>
> - Removing the post from search results, in-product recommendations, trends, notifications, and home timelines
> - Restricting the post's discoverability to the author's profile
> - Downranking the post in replies
> - Restricting Likes, replies, Reposts, Quote, bookmarks, share, pin to profile, engagement counts or Edit post.
> - Excluding the post from having ads adjacent to it
> - Excluding Posts and/or accounts in email or in-product recommendations.

If the amplification of content on Twitter Blue accounts was reduced or revoked as a result of breaches of Twitter's policy on hateful conduct, X Corp. was asked to provide the amount that the amplification reduced, or the period of time it is revoked for. X Corp. responded that its enforcement for Twitter Blue users and non-Twitter Blue users is the same.

X Corp. also stated:

> We're sharing the statistics from our recent Freedom of Speech and Not Reach enforcement progress update which includes both:
>
> - Labels have been applied to more than 700,000 violative posts that fall under our Hateful Conduct policy.
> - We also proactively prevent ads from appearing adjacent to content that we label.
> - Compared to a healthy Tweet, Tweets with these labels—or restricted tweets—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.
> - 90 percent of all appeals receive a response within 30 minutes.

X Corp. referred to Twitter's 'Freedom of Speech, Not Reach' enforcement policy webpage[37] for more information.

[37] X, 'Freedom of Speech, Not Reach: An update on our enforcement philosophy', accessed 27 October 2023, URL: https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy#:~:text=These%20beliefs%20are%20the%20foundation,making%20the%20content%20less%20discoverable.

6.5.4 Proportion of Twitter Blue tweets removed for breaching Twitter's Hateful Conduct policy

X Corp. was asked what proportion of tweets were removed, or required by Twitter to be removed, for breaches of Twitter's Hateful Conduct policy that were posted by Twitter Blue accounts compared with other users for the period 12 December 2022 to 31 May 2023[38]. X Corp. provided the following response:

	% of removals of tweets for hateful conduct, by account type	**Number of removals** of tweets for hateful conduct, by account type	**Total number of removals** of tweets for hateful conduct for all accounts
Twitter Blue accounts	1.87%	17.1K	917.3K
Other users (Non-Twitter Blue accounts)	98.13%	900.2K	

6.5.5 Proportion of Twitter Blue account suspensions for breaches of Twitter's Hateful Conduct policy

X Corp. was asked what proportion of account suspensions for breaches of Twitter's Hateful Conduct policy were for Twitter Blue accounts compared with other users for the period 12 December 2022 to 31 May 2023[39].

[38] This question sought information on the proportion of tweets that were removed, or required by Twitter to be removed, for breaches of Twitter's Hateful Conduct policy that were posted by Twitter Blue accounts in comparison to other users since the Twitter Blue subscription model was relaunched on 12 December 2022. Forbes, 'Twitter Blue Launches Monday: Here's What's Included'; 10 December 2022; accessed 15 June 2023; URL: https://www.forbes.com/sites/johnkoetsier/2022/12/10/twitter-blue-launches-monday-heres-whats-included/?sh=5645391371fa.

[39] This question sought information on the proportion of account suspensions for breaches of Twitter's Hateful Conduct policy for Twitter Blue in comparison with other users since Twitter's account amnesty was announced on 12 December 2022.

	% of accounts suspended for hateful conduct, by account type	**Number of account suspensions** for hateful conduct, by account type	**Total number of account suspensions** for hateful conduct
Twitter Blue accounts	1.35%	375	27.7K
Other users (Non-Twitter Blue accounts)	98.65%	27.312K	

6.5.6 Thresholds for enforcement action

In the Notice eSafety referred to the fact that, in its policy on hateful conduct, Twitter sets out a range of violations it considers to be hateful conduct such as hateful references, incitement, dehumanization, hateful imagery, etc. eSafety also noted that Twitter sets out its Enforcement Options[40] (as well as other places[41]) the action it may take in relation to content, accounts, or both, when a violation of this policy occurs.

X Corp. was asked to outline the thresholds (e.g. severity of content, number of 'strikes', surrounding circumstances/context, etc.) that would or could lead to the enforcement measures listed for the period 1 May 2023 to 31 May 2023[42]. X Corp. was also asked to provide specific examples of conduct that would lead to each enforcement measure listed. X Corp. did not provide examples of conduct. Instead, it provided examples of the enforcement action.

[40] Twitter, 'Our range of enforcement options', 2023, accessed 8 June 2023, URL: https://help.twitter.com/en/rules-and-policies/enforcement-options

[41] 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

[42] This question sought up to date information on what the thresholds were (e.g. severity of content, number of 'strikes', surrounding circumstances/context, etc.) that would or could lead to the enforcement measure listed.

Measures in Twitter's enforcement policy[43]	**Thresholds to determine the enforcement option**
Limiting tweet visibility	Posts that violate hateful conduct policy, but do not target specific individuals
Excluding the tweet from having ads adjacent to it	All hateful conduct violations irrespective of whether their reach is restricted.
Requiring tweet removal	Posts that contain dehumanizing language, incite fear against protected categories, reference a hateful event (for e.g. the holocaust) AND target specific individuals by mentioning them, AND Twitter receives a report from the person mentioned.
Labelling a tweet*	Posts that violate the hateful conduct policy, but do not target specific individuals
Notice of Public Interest exception	When an account representing a current or potential member of a local, state, national, or supra-national governmental or legislative body violates Twitter's Hateful Conduct policy, Twitter may apply the Public Interest Information label.*
Placing a direct message behind a notice*	Twitter does not place direct messages behind a notice if it is reported to X in a 1:1 setting. In a group direct message conversation, if the content posted in the group conversation violates Twitter's policy and a member and a member of the group reports it, the violating direct message may be placed behind a notice to ensure no one else in the group can see it again.*
Placing a tweet behind a notice*	Twitter may place some forms of sensitive media like adult content or graphic violence behind an interstitial advising viewers to be aware that they will see sensitive media if they click through.
Suspending an account*	In its response to the Notice, X Corp. stated that the threshold was when an account repeatedly targets another user with dehumanizing language, incites fear against protected categories, references a hateful event (for e.g. the holocaust) AND targets specific individuals by mentioning them, AND Twitter receives a report from the person mentioned. X Corp. subsequently clarified that such targeted behaviour is considered harassment and when the user repeatedly harasses (defined as a "two strike" policy) someone with hateful conduct and the target reports the post or the account, such accounts will be suspended
Add all additional enforcement options that Twitter has not published*	Direct message-level-enforcement: In a private direct message conversation, when a participant reports the other account, Twitter stops the violator from sending messages to the account who reported them. The conversation is also removed from the reporter's inbox. Withholding a post based on age: Twitter restricts views from specific forms of sensitive media such as adult content for viewers who were under the age of 18 or viewers who do not include a birth date on their profile with interstitials. Withholding a post or account in a country: Twitter may withhold access to certain content in a particular country if it receives a valid and properly scoped request from an authorised entity in that country.

* In its initial response to the Notice, X Corp. did not provide a response to these questions on Twitter's enforcement actions. 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 information included in the table.

[43] Twitter, 'Our range of enforcement options', 2023, accessed 8 June 2023, URL: https://help.twitter.com/en/rules-and-policies/enforcement-options

6.5.7 Determining suspension length

In the Notice, eSafety referred to the fact that Twitter's enforcement policy states that it will 'suspend an account if we determine that a user has engaged in repeated violations of our policies and/or violated specific policies that cause significant risk to Twitter'.[44]

X Corp. was asked to describe the framework that Twitter had in place for determining the length of suspension for breaches of its hateful conduct policy for the period 1 May 2023 to 31 May 2023[45]. X Corp. provided the response below:

Circumstance	Length of suspension
Accounts dedicated to targeting an individual with hateful content and the account was reported to Twitter by the target (first person report)	X Corp. stated that accounts are permanently suspended. This enforcement action applied to accounts that were either majority violations accounts or sole purpose accounts.
Tweet level violations where a user targeted another individual with hateful content by specifically @mentioning them, and Twitter received a first person report	X Corp. stated that it proceeds to suspend these accounts after repeated violation(s). Twitter applies a certain period of time in read-only mode before they can post again.
Tweet level violations where a user shares tweets that incite harassment (mob harassment against a protected category group or specific individuals	X Corp. stated that accounts are suspended after repeated violations.

X Corp. was asked to provide three examples of different suspension lengths and an example of a breach its hateful conduct policy that would lead to each suspension length. X Corp. was asked to provide the suspension length applied in each example. X Corp. did not provide a response to this part of the question. Instead, it provided examples of moderation or enforcement action.

X Corp. provided the following information:

[44] Twitter, 'Our range of enforcement options', 2023, accessed 8 June 2023, URL: https://help.twitter.com/en/rules-and-policies/enforcement-options

[45] This question sought the most up to date information on the framework that Twitter had in place for determining the length of suspension for breaches of its hateful conduct policy.

Example of suspension length **(eSafety notes that none of these examples were of suspension lengths, which the notice sought information on)**	**Example of breach of hateful conduct policy that would lead to each suspension length**
X Corp. gave the example of a post removal with strike	X Corp. stated that this could occur when: 1. Posts that target specific individuals with hateful conduct were bounced (i.e. users were asked to remove the post), and accounts were suspended after repeated violation(s). 2. Posts incited harassment against people belonging to protected categories.
X Corp. gave the example of immediate suspension	X Corp. stated that this could occur when: Accounts are dedicated to targeting specific people with hateful conduct violations (i.e. where Tweets posted violated the hateful conduct policy).
X Corp. gave the example of restricting reach	X Corp. stated that this could occur when there were Tweets that shared hateful content without targeting specific individuals.

X Corp. was asked to provide any other information or context. In response, X Corp. stated that Twitter's policies had not substantially changed but that it had added more transparency and nuance to its enforcement against hateful content. This includes bouncing tweets and accounts that target specific individuals with hateful content as Twitter believes this is akin to 1:1 harassment. Twitter stated that it restricts the reach of content that makes generalised statements but don't target specific individuals.

6.5.8 Detecting recidivism

X Corp. was asked whether a user was prohibited from opening another account after it was put in 'read only mode' for hateful conduct, including as a result of a user's refusal to remove content when required to by Twitter, for the period 1 May 2023 to 31 May 2023[46]. X Corp. responded that that these users were prohibited from opening another account.

X Corp. was asked to specify whether any indicators were used to help prevent re-registration of another account by the same user.

X Corp. provided a response to this question and also indicated that there were additional indicators it had not listed. X Corp. also stated that Twitter has a ban evasion policy that prohibits circumventing a Twitter suspension or enforcement action, which includes any

[46] This question sought the most up to date information on whether a user was prohibited from opening another account after it was put in 'read only mode' for hateful conduct, including as a result of a user's refusal to remove content when required to by Twitter.

behaviour intended to evade any Twitter remediation such as creating a new account or repurposing an existing account.

Following a subsequent question from eSafety, providing X Corp. with a further opportunity to provide the information required by the Notice, X Corp. provided a list of multiple indicators Twitter used to help prevent re-registration of another account by the same user. eSafety has decided not to publish this information to avoid the information being misused.

6.5.9 Permanently banning accounts for hateful conduct

In the Notice eSafety referred to the fact that Twitter's Hateful Conduct policy and Enforcement policy do not mention any circumstances in which an account may be permanently banned for hateful conduct. X Corp. was asked if it had a policy of banning accounts permanently for hateful conduct in any circumstances for the period 1 May 2023 to 31 May 2023[47].

X Corp. responded 'yes'.'

X Corp. was asked to explain circumstances in which accounts may be permanently banned for hateful conduct. X Corp. gave the examples of

1. Accounts that were dedicated to targeting an individual with hateful content, and the account was reported to Twitter by the target (via a first person report). Permanent bans were applied to accounts that were either accounts with a majority of violations or were sole purpose accounts.
2. A user repeatedly targeted another individual with hateful content by specifically @mentioning them and the Tweet was reported to Twitter via a first person report.
3. A user shared tweets that incite harassment (including mob harassment against a group or specific individuals). Permanent bans were applied after 3 strikes/violations.

[47] This question sought the most up to date information on whether Twitter had a policy of banning accounts permanently for hateful conduct in any circumstances.

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

6.6.1 Reinstating previously banned and/or suspended accounts for breaching Twitter's rules and policies

In the Notice eSafety referred to the fact that Twitter announced on 25 November 2022 that it had reinstated accounts that had previously been banned and/or suspended for breaching its rules and policies[48].

X Corp. was asked how many accounts that were previously suspended and/or banned for breaches of Twitter's rules and policies were reinstated during the period 25 November 2022[49] to 31 May 2023. X Corp. was asked to include all accounts that were reinstated which had previously been banned and or/suspended before 28 October 2022[50] for any breach of any of Twitter's Rules and Policies.

Question	Answer
How many accounts that were previously suspended and/or banned for breaches of Twitter's rules and policies were reinstated during the period 25 November 2022[51] to 31 May 2023.	Twitter initially responded 3,172, and then subsequently informed eSafety the number was 6,103 eSafety understands this relates to the number of accounts in Australia

X Corp. stated that any accounts that were withheld based on legal requests were not subject to the amnesty and each account was reviewed before being reinstated.

X Corp. was also asked how many accounts that were reinstated during the period 25 November 2022 to 31 May 2023 had been previously banned/suspended for breaches of Twitter's policy on hateful conduct (or its predecessor policy).

[48] 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

[49] When Mr. Elon Musk announced that suspended accounts would be reinstated. Elon Musk [@elonmusk], 'The people have spoken. Amnesty begins next week. Vox Populi, Vox Dei.', 25 November 2022, accessed 14 June 2023, URL: https://twitter.com/elonmusk/status/1595869526469533701?s=20

[50] This question sought information on all accounts that were previously banned and/or suspended before Twitter's acquisition on 27 October 2022.

[51] When Mr. Elon Musk announced that suspended accounts would be reinstated. Elon Musk [@elonmusk], 'The people have spoken. Amnesty begins next week. Vox Populi, Vox Dei.', 25 November 2022, accessed 14 June 2023, URL: https://twitter.com/elonmusk/status/1595869526469533701?s=20

In its response to the Notice, X Corp. stated that its tools were not built or categorised in a way which enabled it to answer the question.

Following a subsequent question from eSafety, seeking clarification as to whether Twitter has records or maintains records of reasons for a ban or suspension, for the period 25 November 2022 to 31 May 2023, X Corp. clarified it is not the case that Twitter does not have records; rather that its enforcement tools are not built in a manner that provides the categorisation of a violation. X Corp. stated that this was because policies may overlap and have similar or identical enforcement, or because posts may otherwise have violated multiple policies. Additionally, X Corp. stated that it labels high severity violations such as child sexual abuse material, non-consensual nudity, violent threats, and terrorist accounts to ensure that the accounts are not reinstated and that its teams who worked on the Twitter account amnesty used these labels in making the decision to reinstate accounts. Twitter stated that while a breakdown was not supported by its enforcement tools, it had conducted a manual review of all violations for the purposes of responding to eSafety's question, and provided the following information, which eSafety understands relates to Australia:

Between 25 November 2022 to 31 May 2023:

- Total number of accounts suspended: 387,056
    - Total number of accounts suspended for Hateful Conduct violations: 1,196
- Total number of accounts reinstated: 6,103
    - Total number of accounts reinstated that were suspended for Hateful Conduct violations: 194

6.6.2 Number of accounts that had previously been banned and/or suspended for breach of Twitter's hateful conduct policy that had enforcement measures applied

X Corp. was asked how many reinstated accounts, which had previously been banned and/or suspended for a breach of Twitter's policy on hateful conduct (or any of its predecessor policies that had effect during the Report Period), had the following measures applied to them: account suspension; account ban; put in 'read only' mode; user required by Twitter to remove a Tweet; any other additional steps that Twitter has taken in relation to reinstated accounts.

X Corp. responded that its tools currently categorise violations based on the level of harm and severity of the violation, and that corresponds with the number of strikes and suspension remediation. X Corp. stated it was actively reviewing this and as a result, it was not able to provide the exact breakdown of reinstated accounts, previously banned, which had the aforementioned measures applied to them. Instead, X Corp. provided the following data for the total number of suspensions for all accounts made from 25 November 2022 – 31 May 2023. eSafety understands these to be global figures:

Enforcement measure	
Account suspension	142.1 million
Account ban	N/A – X Corp. stated that Twitter uses suspensions
Put in 'read only' mode	111.6 million
User required by Twitter to remove a Tweet	4.3 million
Any other additional steps that Twitter has taken in relation to reinstated accounts	X Corp. stated that any accounts that were withheld based on legal requests were not subject to the Twitter account amnesty and that each account was reviewed before being reinstated.

6.6.3 Applying additional scrutiny in reinstating accounts that have previously breached Twitter's Rules and Policies

In relation to accounts suspended and/or banned prior to 28 October 2022 for breaches of Twitter's policy on hateful conduct (or its predecessor policies), 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[52]. eSafety gave as examples, the potential for fewer strikes required before a reinstated account was downranked, placing an account in 'read only' mode, or suspending an account.

X Corp. responded that reinstated accounts were not placed under additional scrutiny.

X Corp. also 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.
>
> Each account was also reviewed before reinstatement. The relevant teams conducted a harm analysis of the violation types and high severity violations were not included in the amnesty program. These also included policies where there was a risk of physical harm such as CSE [child sexual exploitation] and CSAM [child sexual abuse material], NCN [non-consensual nudity], violent speech, and privacy violations.
>
> It is important to note that the amnesty does not preclude any account from having to comply with the Rules on an ongoing basis. All amnesty accounts are subject to the same rules as every other account.

[52] 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.

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

X Corp. was asked if the policy involved downranking, or making less visible, content previously shared by an account (which may not itself violate Twitter's policies) where the account holder repeatedly breached Twitter's policy on hateful conduct for the period 1 May 2023 to 31 May 2023[56].

X Corp. answered 'yes' to this question.

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 it had misunderstood the question. It responded that it would not take any action in respect of previous content shared by that account unless that previous content also breached its rules. X Corp. also stated that Twitter does not take account level actions when restricting the reach of content. eSafety understands this means that the enforcement action is only applied at post level; and not the account.

X Corp. was also asked in the Notice to explain the circumstances where previous content shared by an account would be downranked for the repeated breaches of Twitter's hateful conduct policy by the account.

In its response to the Notice, X Corp. stated that it restricts the reach of tweets that violate the Hateful Conduct policy when the tweet is sharing hateful content without targeting any specific individual.

X Corp. also stated:

> We proactively detect content using a combination of models and heuristics and when the tweet(s) meets the precision threshold for the remediation, we apply such filters and add a label to let the users know that the tweet violates our policy on Hateful Conduct.
>
> For reactive enforcement, that is, when we receive reports, we apply the policy as-is.
>
> Also if the content moderation reviewer finds additional tweets from the account or related accounts that violate any of our policies, they will apply the relevant policy and matching remediations or actions for those. We will also inform the user accordingly.

6.7.2 Testing and updating recommender systems to reduce the risk that hateful conduct is amplified

X Corp. was asked to describe all measures it employed to test and update its recommender systems with the goal of reducing the risk that hateful conduct is amplified for the period 1 May

[56] This question sought the most up to date information on whether Twitter's policy of restricting the reach of posts that violate its policies involved downranking, or making less visible, content previously shared by an account (which may not itself violate Twitter's policies) if the account holder repeatedly breached Twitter's policy on hateful conduct.

2023 to 31 May 2023[57]. eSafety provided examples such as internal audits, external audits, risk and impact assessments, and a/b testing.

X Corp. did not provide information specifically on how it tests and updates its recommender systems.

Instead, X Corp. provided the following information on how it prevents the amplification of hateful conduct based on data collected from 5 May 2023 – 23 May 2023 from approximately 125,000 Tweets labelled under its Freedom of Speech, Not Reach (FOSNR) enforcement approach for violations of Twitter's hateful conduct policy:

> Compared to a healthy post, restricted posts receive 81% less reach or impressions and we proactively prevent ads from appearing adjacent to this content. Restricted reach can include being excluded from trends, search results, recommended notifications, and the For You /Following timelines. We may also remove the ability for this Tweet to be engaged with. These Tweets will still be visible on the author's profile to everyone on the platform. The majority of impressions served on restricted Tweets are seen by the author's followers. So far, labels have been applied to more than 700,000 violative posts.
>
> Please note that even for policies where we restrict the visibility of a Tweet, if there are situations where we are made aware of the propensity of physical harm (e.g. crisis situations), we may remove such Tweets given the severity, immediacy and likelihood of harm. The rationale is that if/when safety risks outweigh the right to freedom of speech, we would take the necessary remediation and remove such tweets from the platform.

Following a subsequent question from eSafety, providing X Corp. with a further opportunity to provide the information required by the Notice, X Corp. responded that there were no specific tests conducted from 1 May 2023 to 31 May 2023 and that Twitter does not conduct specific tests to its recommender system.

X Corp. also stated that product assessments are conducted on an ongoing basis and that Twitter conducts risk impact assessments encompassing risk categories such as illegal content and hate speech. X Corp. stated that mitigation measures are applied across different aspects of the platform and that the output is ingested and utilised by teams working on the recommender system to model accordingly.

X Corp. also stated that Twitter conducts systemic and product risk impact assessments on an ongoing basis and that includes risk categories such as illegal content and hate speech. X Corp. stated that mitigation measures are applied across different aspects of the platform. In that sense, it stated that the output of the above is ingested and utilised by the teams that are working on the recommender systems to model accordingly.

[57] This question sought the most up to date information on all measures Twitter employed to test and update its recommender systems with the goal of reducing the risk that hateful conduct is amplified.

X Corp. referred to its statement above in response to questions about any alternative measures that Twitter had in place to ensure its recommender systems did not amplify content that breaches Twitter's hateful conduct policy.

Following a subsequent question from eSafety, providing X Corp. with a further opportunity to provide the information required by the Notice, X Corp. stated:

> We plan mitigation measures distinguishing between horizontal measures that address cross-cutting risks, and those measures that can more precisely target a specific systemic risk. This methodology is applicable to the Recommender systems.
>
> For example, X's Freedom of Speech, Not Reach enforcement approach development (FOSNR), which started in April 2023, is reflected in the Recommender systems. Recommender systems are designed to exclude harmful and violating content by integrating with Visibility Filtering and other systems. It uses content health prediction models to prevent harmful and violative content ranking higher.
>
> Another development is that in March 2023, X open-sourced our recommendation algorithm, to build more transparency and trust. This has since allowed independent reviews of Recommender systems by the public.

6.7.3 Artificial or manual amplification of any accounts

X Corp. was asked if any individual Twitter accounts were artificially or manually amplified[58], and if so, what the justification or criteria was for additional algorithmic boost. In its response to the Notice, X Corp. referred to its statements above in section 6.7.2.

eSafety noted the statements above did not provide a response to this question. Following a subsequent question from eSafety, providing X Corp. with a further opportunity to provide the information required by the Notice, X Corp. responded that

> no accounts are artificially or manually amplified.

X Corp. also pointed to information on Twitter's website regarding its recommender system.

6.7.4 Automated accounts and hateful conduct

In the Notice eSafety referred to the fact that, prior to the acquisition of Twitter in October 2022, former CEO of Twitter, Mr. Elon Musk, tweeted that a successful bid for the company

[58] Content posted by the account is promoted above other users on the basis of criteria other than those published by Twitter (i.e. not linked to the characteristics that Twitter stated impacts amplification in its response to the section 56(2) non-periodic notice which were: number of likes; number of retweets; views; number of replies; whether a user engaged with the author in the past; users' interests; recency of a tweet; user's sensitive media settings (minors cannot view sensitive no matter their settings); suspected authenticity (not spammy, twitter blue).

would 'defeat the spam bots.[59]' eSafety also referred to the fact that, after the acquisition in October 2022, Twitter reported a 'spike' in 'hate speech' partly driven by small numbers of accounts using the same slurs at significant volumes (50,000 tweets from 300 accounts[60]).

X Corp. was asked questions about the metrics it used to assess whether Twitter successfully reduced the number of total bot accounts. X Corp. stated:

> The largest success metric at this stage of the process and work is the total number of accounts and networks actioned or thwarted. It is critical to note the following on these:
>
> Twitter uses a combination of human analysts and automated detections to surface and remediate inauthenticity and manipulation of the platform. This approach works towards the primary objectives of *increasing the operating costs for bad actors to manipulate the platform and mitigating any associated harms*.
>
> As spam and platform manipulation is an adversarial space, malicious actors are constantly testing defences and evolving their tactics in response to implemented mitigations. As a result, no system or defence is foolproof, and all automations degrade over time in the face of these adversarial pivots. Consequently, all automated detections and remediations need constant maintenance and iterative updates to remain relevant against bot attacks, along with necessary product changes to address critical defence gaps that are abused over time. This is part of our ongoing work and commitment to addressing these.

X Corp. added that that between 28 October 2022 to 31 May 2023[61]:

- Twitter's automated detections and remediations suspended 130 million accounts that either specifically violated its spam and manipulation policies or were unable to pass an anti-spam challenge. In addition to these suspensions, Twitter stated that it has a suite of alternative account and tweet-level enforcement possibilities that include both reducing the visibility and impressions of content, as well as limiting the number of actions they can take on the platform that reduce the impact of spam-like activity on the platform to an order of approximately 40 million actions per day.
- Twitter's manual suspension actions against spam and platform manipulation for violative activity totalled 18 million actions.
- In total, approximately 150 million accounts were suspended for violating Twitter's Spam and Platform Manipulation policy. Whilst the majority of those accounts would be

[59] Elon Musk [@elonmusk], 'If our twitter bid succeeds, we will defeat the spam bots or die trying!', 22 April 2022, accessed 13 June 2023, URL: https://twitter.com/elonmusk/status/1517215066550116354

[60] 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

[61] Between the date Twitter was acquired and the end of the Report Period.

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