

Is content over- or under-moderated in the Voice referendum debate?

An experimental evaluation



Summary

Content moderation on social media can result in the removal, demotion or labelling of content that platforms deem to have violated their rules. Moderation is an important tool in mitigating systemic risks on platforms, especially when it comes from misinformation and disinformation. However, if moderation goes wrong, it can lead to content being either over-moderated (too much being inappropriately taken down) or under-moderation (not enough content that violates platform's rules being taken down). Platforms can also create conditions of potential political bias if they over- or under-moderate in particular ways.

This research set out to see if the content moderation systems of three major platforms—TikTok, Facebook and X—produced over- or under-moderation, and if they displayed political bias when it came to content relating to the Voice referendum in Australia. We tested for differing levels of 'over-moderation', or where platforms had inappropriately removed, demoted or labelled Yes-aligned or No-aligned content. We also examined differing levels of 'under-moderation', which involved instances where platforms had failed to remove, demote or label misleading Yes-aligned or No-aligned content that violated their guidelines. Subsequently, we uncovered the following findings:

- **Over-moderation:** we found limited evidence of platform over-moderation. The techniques used in this research encourage overestimation, but even these overestimates ranged from **0.25%** on Facebook to **2%** on X.

There is limited evidence of bias, however, we found X may over-moderate #VoteNo content, and Facebook appears to favour #VoteNo content in its video recommender algorithm to a five-fold magnitude.

- **Under-moderation:** our findings suggest misinformation was substantially under-moderated across all three platforms. Misleading content regarding electoral processes that violates each of the platforms' community guidelines was not removed when platforms became aware of it. Between **75% and 100%** of misinformation was subject to under-moderation, depending on the platform and its substance. No political bias was detected in these processes.

These findings suggest that the platforms' content moderation systems were not significantly biased in terms of moderating Yes or No-aligned content. Consistent with our earlier research, there remains a substantial, potentially systemic issue regarding under-moderation of misinformation.

Furthermore, this research suggests that the measures from the **Australian Code of Practice on Disinformation and Misinformation** might not be effectively preventing the under-moderation of content. It is also evident that the signatories' transparency reports have not identified the issues highlighted by this research.

Reset.Tech Australia is an Australian policy development and research organisation. We specialise in independent and original research into the social impacts of tech companies. We are the Australian affiliate of **Reset.Tech**, a global initiative working to counter digital harms and threats. Reset.Tech has extensive, global experience in monitoring electoral misinformation and disinformation with a focus on identifying areas for regulatory intervention. **We are not affiliated to either referendum campaign.**



Reset. AUSTRALIA

Table Of Contents

Summary	2
Contents	3
Introduction	4
Over-moderation	7
What we did: Methods	7
What should happen	9
What happened: Findings	10
Under-moderation	15
What we did: Methods	15
What should happen	18
What happened: Findings	18
Conclusions	24
Appendix 1: Digi's Code and platforms' community guidelines in more detail	25
Digi's Australian Code of Practice on Disinformation and Misinformation	25
TikTok's policies and reports to Digi regarding content moderation	27
Facebook's policies and reports to Digi regarding content moderation	29
X's (formerly Twitter's) policies and reports to Digi regarding content moderation	31

Authors and researchers
Dr Rys Farthing, Aruna Anderson, Alice Dawkins, Racheline Tantular

Funding
 **Susan McKinnon Foundation**

Cover artwork
Jamillah Knowles / Reset.Tech Australia / Better Images of AI / Connected People Australia (Blue) / CC-BY 4.0

Introduction

Toxicity on social media has sadly become an accepted feature, rather than a bug within digitally-enabled democratic processes. Misleading content (*elsewhere known as misinformation or disinformation*), online-enabled abusive speech towards individuals and groups and hostile online campaigns against public figures are now regular themes of public debate.

An unfortunate hallmark of the lead-up to the Voice referendum has been the notable levels of misinformation in relation to both the policy proposals and the electoral process itself. Downturns in trust and safety resourcing at numerous social media platforms have likely contributed to a deterioration in both cautionary investments regarding misinformation as well as less effective response mechanisms when platforms are made aware of it.

The proliferation of false and misleading information is often found to be under-moderated by social media platforms. For example, Reset.Tech has run experiments that indicate electoral process misinformation is being under-moderated and is not removed, labelled or demoted as per platforms' guidelines.¹ Meanwhile, in the adjacent domain of harmful online content, various campaign stakeholders, most prominently the First People's Assembly of Victoria, have reported how platforms have not provided necessary support or redress.² Alongside concerns surrounding under-moderation,

there are many equally disturbing instances that have been flagged regarding platforms inappropriately over-moderating. Over-moderation occurs when robust political dialogue is removed, labelled or demoted in ways that go beyond a platform's guidelines on the matter. The worry is that some robust political content is incorrectly classified as misinformation. In some cases, organisations claim advertisements are denied based on being inappropriately classified by the platform as political ads.³ These stakeholders have argued that Facebook in particular is demonstrating a bias against various political viewpoints.

A general lack of transparency in platforms' content moderation processes has routinely led to low trust outcomes from a diverse array of stakeholders. Responses to these imperfect transparency conditions vary, but they generally lead to claims of political bias in a particular direction. Some have argued that Facebook's apparent approach to under-moderating Holocaust denialism displays a dangerous tolerance towards neo-Nazi views.⁴ At the same time, others have argued that over-moderation via the "deplatforming" of various political figures, suggests that the company displays internal preferences for liberal ideas.⁵

The above allegations spring out from what tech policy scholar Evelyn Douek characterises as a 'misleading and incomplete' picture of content

¹ Reset.Tech Australia, How do platforms respond to user-reports of electoral process misinformation? An experimental evaluation from the lead-up to Australia's referendum, 2023, <https://au.reset.tech/uploads/Reset.Tech-Electoral-Misinformation-Report.pdf>

² Australian Associated Press and Josh Butler, 'Victoria's First Peoples' Assembly says Facebook must act against 'tidal wave' of racist trolls' The Guardian, May 26, 2023, <https://www.theguardian.com/australia-news/2023/may/26/victorias-first-peoples-assembly-says-facebook-must-act-against-tidal-wave-of-racist-trolls>; Dechlan Brennan, 'Meta refuses to remove anti-Indigenous racist content despite complaints' National Indigenous Times, August 23, 2023, <https://nit.com.au/23-08-2023/7343/meta-refuses-to-remove-racist-content-despite-complaints>; Jack Latimore, 'Meta rules online racism against Indigenous people meets community standards', Sydney Morning Herald, August 23, 2023, <https://www.smh.com.au/national/meta-rules-online-racism-against-indigenous-people-meets-community-standards-20230815-p5dwtq.html>

³ See, for example, claims by the Institute of Public Affairs reported by Josh Butler, 'Voice referendum no campaign accuses Facebook of "restricting democracy" over ad removal', The Guardian, March 2, 2023, <https://www.theguardian.com/australia-news/2023/mar/02/voice-referendum-no-campaign-accuses-facebook-of-restricting-democracy-over-ad-removal>. See also Peta Credlin, 'Big Tech, Yes camp censors will only reinforce No vote', The Australian, August 18, 2023, <https://www.theaustralian.com.au/commentary/big-tech-yes-camp-censors-will-only-reinforce-no-vote/news-story/b4d58fd295a9cc8ad3f87dcaad49ea5a>

⁴ See some commentary on this issue as summarised by Elizabeth Dwoskin, 'Mark Zuckerberg's reversal on Holocaust denial is a 180', Washington Post, October 12, 2020, <https://www.washingtonpost.com/technology/2020/10/12/zuckerberg-holocaust-denial-facebook/>

⁵ See complaints covered by Ashleigh Gold, 'Republicans raise bias claims to board reviewing Trump's Facebook ban', Axios, February 11, 2021.

moderation. Douek argues that the perception that platforms govern content by reference to merits of individual speech decisions obfuscates the more likely approach platforms take, which is an upstream, systematic focus on ‘patterns of change rather than static snapshots.’⁶ As Douek says,

*Content moderation is not just the aggregation of many (many!) binary decisions to take down or leave up individual pieces of content... It is a vast system of administration that includes a far broader range of decisions and decision makers than the standard picture admits.*⁷

Through their moderation systems, platforms can create rules and guidelines that see certain political speech removed, demoted or appended with warning labels. Content moderation is a necessary part of running a safe and effective platform; for example, it is key to removing pro-suicide or pro-terrorism content. It is also a vital component of mitigating misinformation.

Using the Voice referendum as a case study, we set out to explore various concerns regarding over-moderation and under-moderation and to determine if certain platforms’ content moderation systems demonstrated bias towards Yes-aligned or No-aligned content. We specifically set out to answer two interrelated research questions with small-scale monitoring:

1

Is content over- or under-moderated in the Voice referendum debate?

If general discourse is moderated, we regard this as “over-moderation”. If misinformation that violates platforms guidelines is not moderated when platforms are made aware of it, we regard this as “under-moderation”.

2

Does content moderation display any bias towards Yes- or No-aligned content with regard to either over- or under-moderation practices?

In other words, is Yes or No-related content subject to more over- or under-moderation than the other?

These findings have implications for platform compliance to the **Australian Code of Practice on Disinformation and Misinformation** (referred to hereafter as “the Code”).⁸ Certain types of misleading content that obfuscate electoral processes, such as claims that ballot measures are unconstitutional or posts that call into question the integrity of said measures, are considered misinformation under the Code. According to the Code, signatories must simply ‘develop and implement measures’ that ‘aim to reduce’ the propagation of and potential exposure to misinformation, which may include content moderation practices.

Content moderation measures may include developing clear guidelines about the nature of the content they moderate, policies regarding the removal, labelling or demotion of violative content as well as providing users with ways to report content that they believe violates platform’s policies (**see Figure 1 for a summary or Appendix 1 for more details**). Where platforms are over-moderating or under-moderating content in ways that fail to uphold their stated policies, they may be in breach of the Code.

⁶ Evelyn Douek, ‘Content Moderation as Systems Thinking’, Harvard Law Review 136 (Forthcoming): 4.

⁷ Ibid: 5.

⁸ Digi 2022 Australian Code of Practice on Disinformation and Misinformation <https://digi.org.au/wp-content/uploads/2022/12/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL--December-22-2022.docx.pdf>



Mandatory commitments for platforms who sign the Code

Provide safeguards against harms that may arise from misinformation and disinformation

- Develop and implement measures that aim to reduce the propagation of and potential exposure to disinformation and misinformation by users on digital platforms. These may include:
 - *Policies and processes that require human review of content;*
 - *Labelling false content;*
 - *Removal of content propagated by inauthentic behaviours (bots etc.);*
 - *Suspension or disabling of accounts that engage in inauthentic behaviour (see **Appendix 1 for a full list of suggestions**).*
- Develop and implement measures that inform users about the types of behaviour/content that will be prohibited and/or managed under their policies.
- Develop and implement tools and policies that allow users to report content regulated under the Code.

Publish transparency reports

- Publish policies and reports that users can see regarding the effectiveness of a platform's measures and the progress they have made to realise their obligations under the Code.

More details about these obligations are provided in **Appendix 1** for clarity.

Figure 1: A summary of the mandatory commitments for signatories of the Code.

Why monitor during the Voice referendum?

The Voice referendum is a uniquely important event in Australia's history that provides a valuable, timely case study for evaluating platform responses to misinformation and disinformation. It is particularly noteworthy for the following reasons:

- It is distinctly Australian, which means we can monitor international platforms' responses to a national issue as there is less potential conflation with global responses. Platform transparency reports to Digi tend to lack suitably granular local data on their response and mitigation measures.
- It is an Australian electoral process, meaning that all the features of electoral misinformation and disinformation will apply, and lessons can be learned for future elections.
- To an extent, it is more narrowly defined than the likes of a general election where electoral content and broader current-affairs posts would be harder to differentiate.

Over-moderation

What we did: Methods

1

Finding ‘everyday’ political content to monitor

At a random point in time, we recorded the latest 200 posts that included either #VoteYes or #VoteNo on both X (formerly Twitter) and Facebook. This created a sample of 800 pieces of Yes or No-aligned content (*see Figures 2 and 3 for examples*).

2

Monitoring content:

We followed this content for four weeks, noting each week if:

- Content became unavailable. This allows us to overestimate removal rates, i.e. how much of the content may have been removed by platforms’ content moderation systems. This overestimate is a limitation of the research.
- Content became labelled. This allows us to estimate labelling rates, i.e. how much of the content was labelled.
- Content grew or declined in terms of view counts. This allows us to estimate the impact of any demotion or algorithmic dampening efforts by platforms. Facebook videos are the only medium on the platform that provide view counts, so analysis of this company only relates to video content. Content normally organically decays in growth sharply week on week, but if content entirely stalled on growth at an unexpected point in time, we considered this demotion.

3

Analysis:

To decide if content was over-moderated, we looked at all the content that had become unavailable, labelled or noticeably demoted, and then we assessed whether it violated the specified platforms’ community guidelines. If it did not, we considered this content over-moderated. We compared these practices on a platform-to-platform basis. Please note, given the limitations described below, we can only make estimations regarding potential over-moderation based on content. Accounts may be closed for inauthentic activity, but this research cannot account for those eventualities.

To decide if there was bias in over-moderation rates, we compared the rates of over-moderation for the #VoteYes sample to the #VoteNo sample. We also compared this by platform.

Limitations

This method is limited by the lack of transparency around why content becomes unavailable on social media platforms. This may occur when it is removed by the platform (a form of content moderation), or content may become unavailable because of the actions of users themselves (which is not content moderation). Users often delete posts, close their accounts or make them private, which renders the content inaccessible for monitoring. Alternatively, when content is posted on group pages—such as on Facebook—group moderators may remove the content themselves for a number of reasons. For example, content may have been posted twice,

group moderators might not think the content meets their particular standards (images may be low quality, or they do not think group members will be interested in it) or criteria (such as posting celebrity content on a dog-appreciation page) or because they do not agree with the content of the post. Content removed by group moderators is not a form of content moderation that is overseen by

platforms. These companies are not always up-front about who deletes these posts or why they are removed, so it can be unclear whether content that becomes unavailable has content has been moderated or not. “Not available” rates represent an imperfect approximation of a platform’s level of content moderation, and they represent an overestimate of the effect of content moderation.

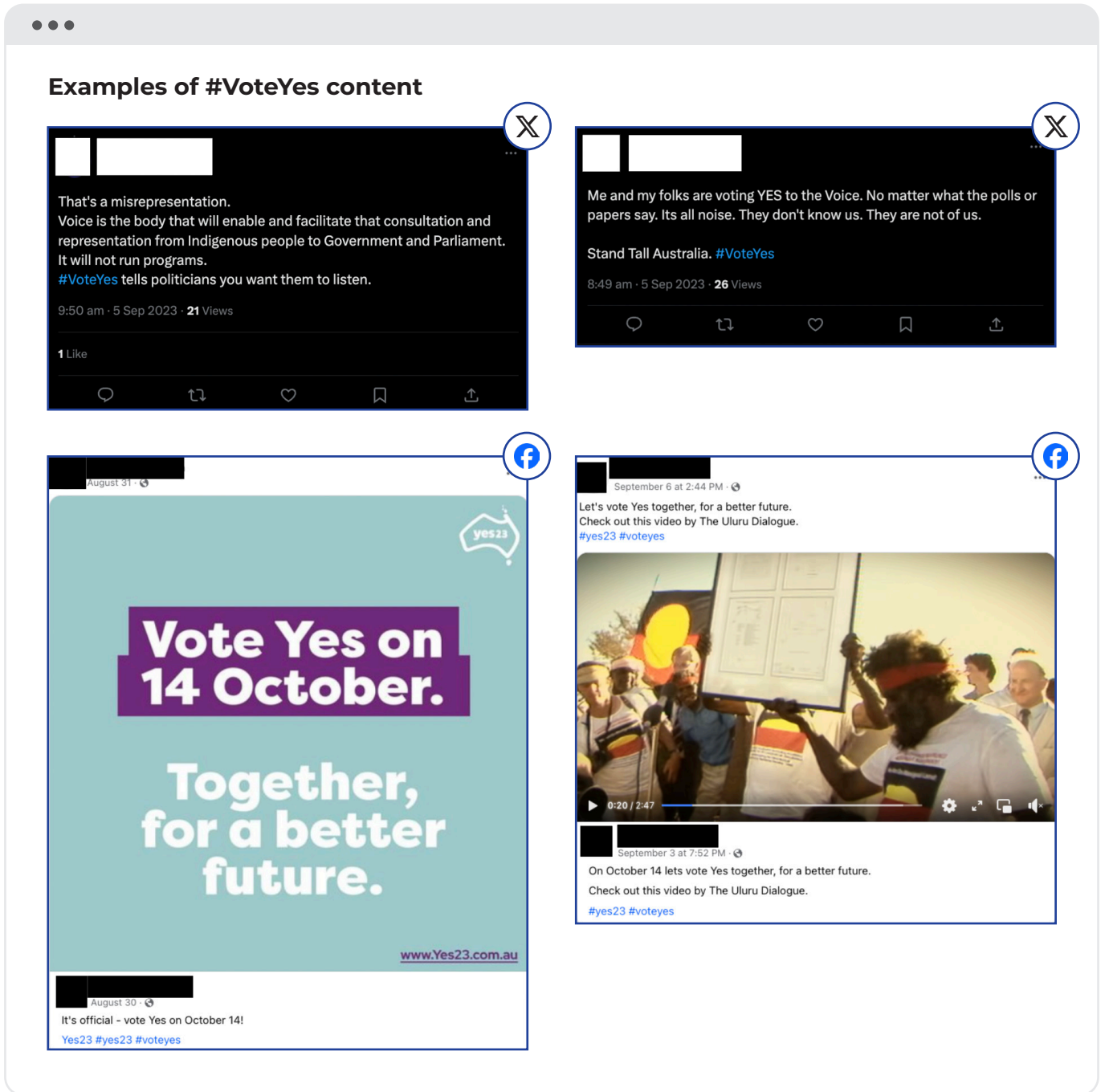


Figure 2: Examples of everyday #VoteNo political content that were found on X and Facebook.

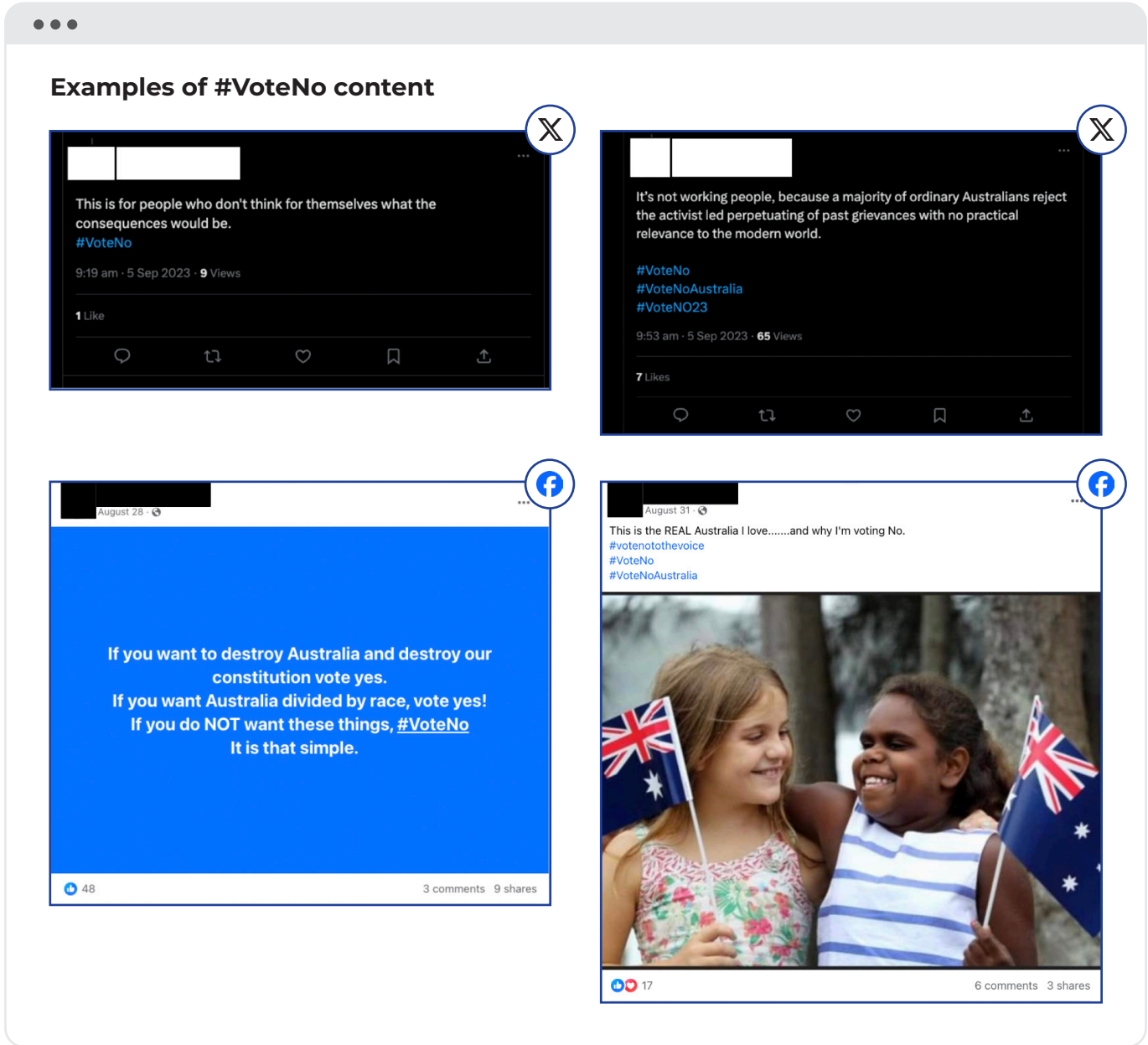


Figure 3: Examples of everyday #VoteYes political content that were found on X and Facebook.

What should happen

- If there is no over-moderation:**
- f On **Facebook**, content that is not violative should not be 'demoted in prevalence', nor should it be removed.
 - X On **X**, content that is not violative should not be labelled, nor should it be removed by the platform.

If there is no bias in over-moderation:

We should see similar levels of over-moderation in both the #VoteYes and #VoteNo sample.

What happened: Findings

These figures represent generous overestimates of the impact of over-moderation on content, and where it was unclear, we have assumed content became unavailable because of a platform's moderation. Even with these generous overestimations, the impact of potential over-moderation on content is low:

- The highest estimate for over-moderation on Facebook is **0.25%** of content.
- The highest estimate for over-moderation on X is **2%**.

In terms of bias:

- X showed a potential bias to over-moderate #VoteNo content, although we would need to understand why some of the accounts monitored were suspended before we could determine if this is true. For example, they may have been suspended for inauthentic activity.

- Facebook showed a slight bias towards over-moderating #VoteNo content, but we would again need to understand why the single piece of content in question became unavailable before we could state that this is the case. However, Facebook's algorithm appears to significantly favour #VoteNo video content over #VoteYes video content, with the former being disseminated around five times faster than #VoteYes content. The decay of growth rates was similar.

There is a significant level of dynamism with regard to #VoteNo content on Facebook (meaning a significant number of posts disappeared swiftly), yet it does not appear to be immediately evident as platform-based political bias in the form of content moderation. Most likely, this was the impact of group moderation (by pro-No groups), or differing user behaviours.



Content unavailability:

potential over-moderation is where content has become unavailable but was not deemed to have violated community guidelines.



Content labelling:

potential over-moderation is where content is labelled but not deemed to have violated community guidelines.



Content demotion:

potential over-moderation is where content is substantively demoted.

Potential content over-moderated in total (any technique)

0.25% of content was potentially over-moderated. **0%** of #VoteYes, **0.5%** of #VoteNo. Facebook's algorithm may favour #VoteNo videos over #VoteYes videos.

Content unavailability:

0.25% of content was potentially over-moderated. **0%** of #VoteYes, **0.5%** of #VoteNo

Content labelling:

0% of content was potentially over-moderated. **0%** of #VoteYes, **0%** of #VoteNo

Content demotion:

Both #VoteYes and #VoteNo videos showed substantive growth decay after one week, which is to be expected. #VoteNo content disseminated at around five times the speed of #VoteYes content.



Potential content over-moderated in total (any technique)

2 % of content was over-moderated. **0.5 %** of #VoteYes, **3.5 %** of #VoteNo

 Content unavailability:

2 % of content was over-moderated. **0.5 %** of #VoteYes, **3.5 %** of #VoteNo

 Content labelling:

0 % of content was potentially over-moderated. **0 %** of #VoteYes, **0 %** of #VoteNo

 Content labelling:

Both #VoteYes and #VoteNo posts showed substantive growth decay after one week, which is to be expected. The decay was around the same.

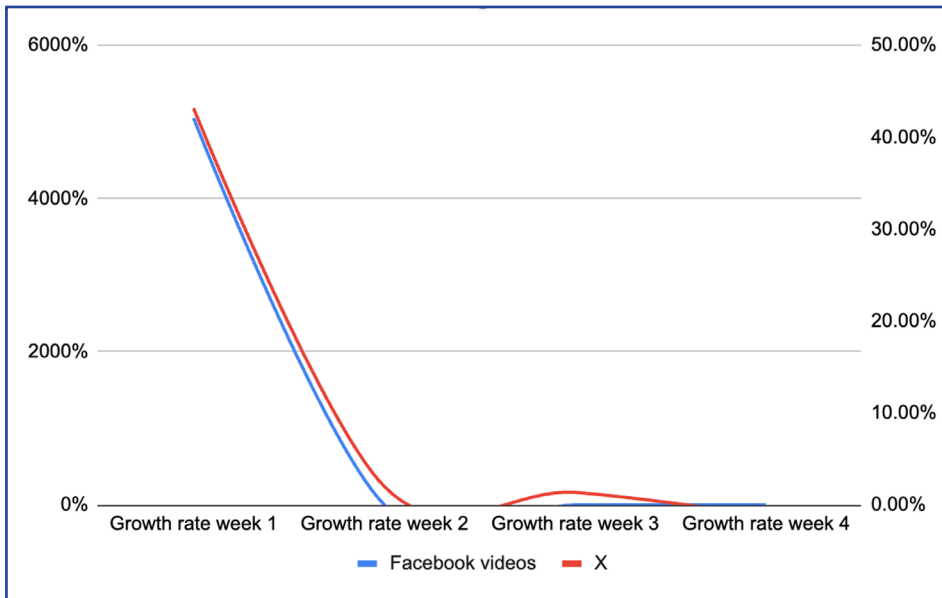


Figure 4: The week-on-week growth rates of everyday #VoteNo political content. The growth of Facebook videos is shown on the left-hand axis, while the growth of X-based content is shown on the right-hand axis.

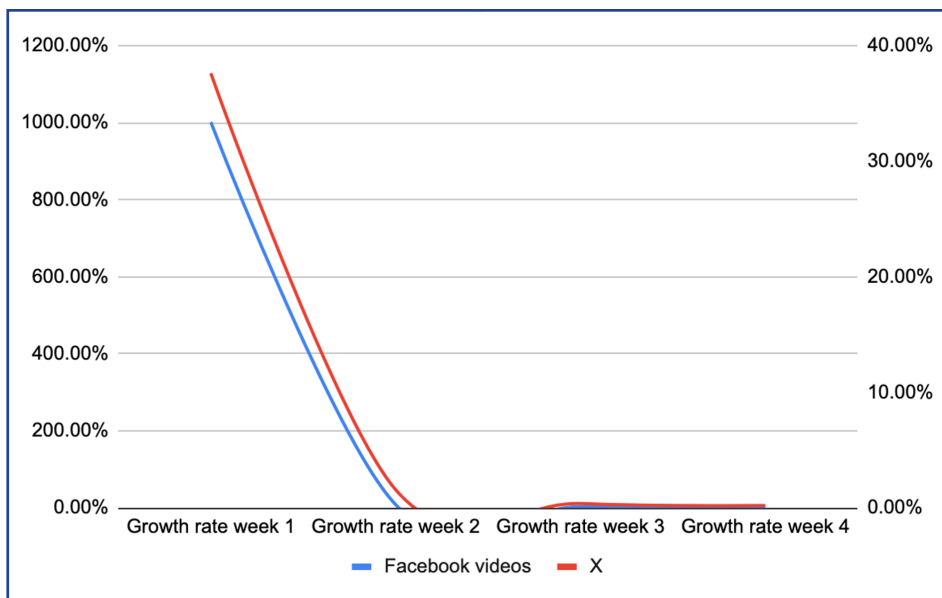


Figure 5: The week-on-week growth rates of everyday #VoteYes political content. The growth of Facebook videos is shown on the left-hand axis, while the growth of X-based content is shown on the right-hand axis.

About the content that became unavailable or was labelled

#VoteNo content

[Nine pieces of No-aligned content became unavailable on the X platform, seven of which could be indicative of over-moderation:](#)

Three became unavailable on the first day of monitoring. We are interested in content that became unavailable quickly as it is the most likely to have been detected by automated content moderation systems looking for inauthentic behaviour.

- Two were from an account that closed on the first day of monitoring. Both of these posts on X (Tweets) included the same text, including links to other posts on X that were also unavailable. While this may be an account that was closed by the user, or one that was shut down for inauthentic activity, based on content analysis alone, this could be an example of over-moderation.
- One post on X was unavailable on the first day of monitoring from an account that is still available. X displays when a post or account has been removed for violating terms, but this was not the case here, so it was most likely removed by the user.

Six instances of such content became unavailable later in the monitoring process:

- One was from an account that was closed, most likely by the user.
- Five posts were made by three accounts that were suspended by X. We are not able to determine the reason for the suspension, but frequently this is for repeated violation of community guidelines. None of the five individual pieces of content monitored here were violative, so we determine that this could be down to over-moderation.

[Twenty-four pieces of No-aligned content became unavailable on the Facebook platform, one of which could be indicative of over-moderation:](#)

Two examples became unavailable on the first

day of monitoring, and both of them were posted to groups. Their content and placement suggests that they were removed by group moderators rather than platform's content moderation. We do not count these as potentially over-moderated.

- One post became unavailable on a notice board for a community garden, which does not appear to have other Yes-Aligned or No-aligned content. This post was most likely removed by group moderators for its irrelevance.
- One post became unavailable within a group that, according to their bio, encourages opinions from both sides of the vote; however, they note all interactions must be respectful. The group hosts many No-aligned and Yes-aligned posts. The post that became unavailable was an ad hominem attack on a Yes campaigner, so it was most likely moderated by the group.

Twenty-two other posts became unavailable:

- Two posts were from individual accounts posting publicly, whose accounts have not been suspended, nor were they fact-checked. This is more suggestive of users removing old posts rather than platform moderation.
- Twenty posts shared in Facebook groups became unavailable, yet only one of them may be indicative of over-moderation. These are detailed below, but to understand this, we need to introduce two users who posted frequently to these groups who we describe as frequent 'post-then-deleters'. Both 'post-then-deleters' have Facebook accounts that have not been suspended, nor were they visibly fact-checked. If their content was moderated by platforms as frequently as we detected it becoming unavailable, their accounts would have been suspended. They do not appear to post content that is violative of platform or group rules, so these two users are more likely to be frequent Facebook users who post material and then remove it a few hours or days later themselves. It is also possible that their frequent posting has made them unpopular with group moderators.

 Group	 What happened
No-aligned group focused on Queensland	<p>Four posts became unavailable in this group:</p> <ul style="list-style-type: none"> • Two were from an account that went private. • Two were from one of the ‘post-then-deleters’.
No-aligned group focused on opposing the political left	<p>Four posts became unavailable in this group:</p> <ul style="list-style-type: none"> • Both were from one of the ‘post-then-deleters’.
No-aligned group focused on Western Australia	<p>One post became unavailable within this group:</p> <ul style="list-style-type: none"> • By a user who went private.
No-aligned group focused on avoiding the ‘doom’ of the Voice passing	<p>One post became unavailable within this group:</p> <ul style="list-style-type: none"> • By one of the ‘post-then-deleters’.
A group that hosts and encourages opinions from both the Yes and No camps, but states that all interactions must be respectful. It hosts a lot of No-aligned and Yes-aligned content.	<p>Twelve posts became unavailable in this group:</p> <ul style="list-style-type: none"> • Four posts were exactly the same and were re-posted four times within the course of three minutes; they were most likely removed by the group’s moderators. • Five posts were by one of the ‘post-then-deleters’. • One post was by the other ‘post-then-deleter’. • One post was by a user that went private. • One post may have been subject to platform moderation, but it is more likely to have been subject to group moderation. As figure X below highlights, the post claims that Australia has a ‘fake community government’. Facebook allows similar claims elsewhere across the platform, so it is unlikely to have moderated this. The Facebook group it was posted in tends not to host content that is posted in full caps or posts that call the government liars. It was most likely subject to group moderation, but as it may have been subject to moderation we have included it as potentially over-moderated in this analysis.

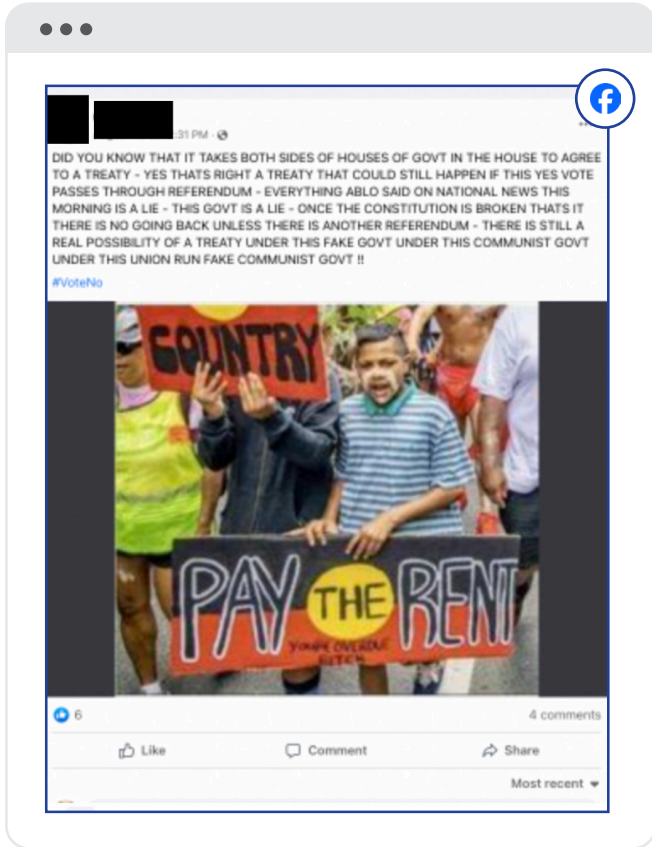


Figure 6: The post that became unavailable within a Facebook group. This may have been moderated by the group, or it may have been subject to platform moderation. To avoid under recording, we have recorded this as potentially inappropriately moderated by the platform.

There was some dynamism in terms of post availability on Facebook that was not seen on X. For example, six posts became unavailable for a week or two and then became available again. This is most likely due to accounts changing their privacy settings, but could also represent accounts that were temporarily suspended. We do not count these as subject to over-moderation because they are still available and most likely became unavailable for a short period of time.

One piece of No-aligned content was labelled on Facebook. The post claimed that Australia had two constitutions, the Australia constitution, which belongs to the corporation of Australia, and the constitution of Australia, or the 'true' constitution.

This is a falsehood that has been fact-checked by the Australian Associated Press (AAP),⁹ so it is not counted as an example of over-moderation in this analysis. This was the only piece of content in the entire 800-piece sample that was labelled.

#VoteYes content

Three pieces of #VoteYes content became unavailable on X, one of which may be indicative of over-moderation:

One became unavailable on the first day of monitoring. It included links to another post on X that was also unavailable. While this may be an account that was closed by the user or for exhibiting inauthentic activity, based on content analysis alone, this could be an example of over-moderation.

Two other posts became unavailable later in the monitoring process. One was by an account that switched to private, and the other was an account that was closed most likely by the user.

No #VoteYes content became unavailable on Facebook.

⁹ AAP FactCheck, Two constitutions claim is sovereign citizen silliness, 2023, <https://www.aap.com.au/factcheck/two-constitutions-claim-is-sovereign-citizen-silliness/>

Under-moderation

What we did: Methods

1

Finding samples of electoral misinformation content to monitor:

We searched for electoral misinformation on X (formerly Twitter), Facebook and TikTok that clearly violated each platforms' guidelines. Two types of misinformation were tracked:

A. Content that claimed that the referendum was unconstitutional. We monitored 49 pieces of content that largely rested on claims that the Voice referendum was unconstitutional because the High Court of Australia had ruled that the ballot or vote violated the constitution. This has been fact-checked as incorrect by the AAP.¹⁰ Some content also suggested that the referendum was unconstitutional because it was an attempt to change the invalid corporation of Australia's constitution, not the real Government of Australia's constitution, which relates to a broader conspiracy theory. This has also been fact-checked by the AAP and was deemed false.¹¹ This sample of content did not include debates about potential constitutional challenges or weaknesses of the referendum, which we simply regard as robust political dialogue.

Once we had identified the content to monitor, we categorised each piece of misinformation as either Yes-aligned (when it urged viewers to 'Vote Yes' in messaging or hashtags etc.), No-aligned (when it urged viewers to 'Vote No' in messaging or hashtags etc.) or unclear in instances where it did not endorse either

campaign. Misinformation content was included in the sample as it was discovered, and we discovered more unclear and No-aligned misinformation than Yes-aligned information.

In total, we found:

- Thirty-four pieces of unclear misinformation;
 - Fourteen pieces of No-aligned misinformation, and,
 - Two pieces of Yes-aligned misinformation.
- B. Content that undermined the electoral integrity of the referendum voting process, by making or amplifying unproven claims regarding electoral irregularities and casting doubt on the integrity of public institutions. We monitored 58 pieces of content derived from popular claims in the sovereign citizen movement, including that the referendum was "rigged" in some way—either through a "deep-state" style conspiracy theory, or by undermining the independence of the Australian Electoral Commission (AEC). These claims have been widely fact-checked as incorrect by the AAP.¹²

Once we had identified content we wanted to monitor, we categorised each piece of misinformation as either Yes-aligned (when it urged viewers to 'Vote Yes' in messaging or hashtags etc.), No-aligned (when it urged viewers to 'Vote No' in messaging or hashtags etc.) or unclear in instances where it did not endorse either campaign. In total, we uncovered:

¹⁰ Australian Associated Press, Unconstitutional voice claim is pure 'nonsense', 2023, <https://www.aap.com.au/factcheck/unconstitutional-voice-claim-is-pure-nonsense/>

¹¹ Australian Associated Press, Two constitutions claim is sovereign citizen silliness, 2023 <https://www.aap.com.au/factcheck/two-constitutions-claim-is-sovereign-citizen-silliness/>

¹² Australian Associated Press, Motley list of misinformation goes viral, 2023, <https://www.aap.com.au/factcheck/motley-list-of-voice-misinformation-goes-viral/>

- Fourteen pieces of unclear misinformation
- Forty-three pieces of No-aligned misinformation and
- One piece of Yes-aligned misinformation

Both samples of misinformation content violate platforms' guidelines.



On **TikTok**, this sample of content would violate their guidelines relating to electoral misinformation. TikTok states,

We do not allow misinformation about civic and electoral processes, regardless of intent. This includes misinformation about how to vote, registering to vote, eligibility requirements of candidates, the processes to count ballots and certify elections, and the final outcome of an election¹³



On **Facebook**, because this sample of content has been fact-checked, they commit to 'reducing its prevalence or creating an environment that fosters a productive dialogue.'¹⁴



On **X**, this sample of content would fall under 'misleading claims that cause confusion about the established laws, regulations, procedures, and methods of a civic process, or about the actions of officials or entities executing those civic processes,' which violate their guidelines.¹⁵

2

Monitoring content

We monitored this content for a week, noting the rates at which it became unavailable, labelled or amplified. We then reported this content and

monitored it for another week to see the rates at which it subsequently became unavailable, labelled or amplified, as described in the section on over-moderation methods.¹⁶

3

Analysis

- To decide if content was under-moderated, we looked at all the content that had not been labelled, removed or demoted. As this entire sample of content violated the platforms' community guidelines, any content that was not moderated is regarded as evidence for under-moderation.
- To decide if there was bias in under-moderation rates, we compared the rates of under-moderation for the Yes-aligned sample, the No-aligned sample and the unclear sample.

Limitations

This method is limited by the lack of transparency surrounding why content becomes unavailable on platforms, as described above. Content may be removed by the platforms, users themselves or group moderators.

¹³ TikTok Civic and election integrity, 2023, <https://www.tiktok.com/community-guidelines/en/integrity-authenticity/>

¹⁴ Meta, Community Standards: Misinformation, 2023, <https://transparency.fb.com/en-gb/policies/community-standards/misinformation/>

¹⁵ X, Civic integrity misleading information policy, 2023, <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>

¹⁶ This research is not testing the efficacy of the user -reporting system. We have done this previously and will continue this research post-referendum. Reset.Tech 2023 Report: Electoral process misinformation, 2023; <https://au.reset.tech/news/report-electoral-process-misinformation/>



A note on reporting content

During step two of this research, we noted that X had removed the ability for users to report electoral misinformation.¹⁷ Previously, Australian users were able to report electoral misinformation in three clicks by clicking on 'Report Content', then selecting 'It's misleading' and then 'Politics'. This sent the reported content into an appropriate content moderation flow, as described by X in their Help Centre.

Somewhere around September 24th—around three weeks before the Voice referendum—X turned off the ability for users to report electoral misinformation. We believe this decision will amplify under-moderation, is a threat to electoral integrity and a breach of X's commitments under the Code.

For the purposes of this research, we reported posts in the best available category, generally as a type of hate speech, abuse or incitement to violence.

We have done this before: We have previously explored under-moderation of electoral misinformation regarding similar pieces

of content. [We monitored 99 pieces of content throughout August 2023](#) and found that platform-takedown rates for serious cases of misleading content were extremely low.

Specifically, we found:

1

Platforms appear to have few effective 'organic' content moderation processes to detect and respond to electoral process misinformation and disinformation.

2

Reporting electoral process misinformation appears to make little difference on Facebook and X, while it makes a moderate difference on TikTok.

3

Electoral process misinformation continues to grow in reach even after reporting, which suggests that it is not adequately being de-amplified. Growth accelerates slowly after reporting on TikTok, but it decelerates significantly on Facebook.




4

The nature of the content that becomes unavailable or is labelled does not appear to be substantively different to the content that remains, suggesting that the content moderation process is a 'whack-a-mole' rather than a systematic process.

¹⁷ See Reset.Tech, Open Letter to X, 2023, <https://au.reset.tech/news/open-letter-to-x/>

What should happen

If there is no over-moderation:

-  On **TikTok**, all of the content should be removed
-  On **Facebook**, all of the content should be removed or 'demoted in prevalence'
-  On **X**, all of the content should be labelled or removed

If there is no bias in over-moderation:

Removal, labelling and growth rates should be the same between the Yes-aligned sample, the No-aligned sample and the Unclear sample.

What happened: Findings

This research found that misinformation was substantially under-moderated across all three platforms.

Misinformation regarding electoral processes, that violates the platforms' community guidelines, was

generally not removed when platforms became aware of it. Indeed, between **75% to 100%** of misinformation was subject to under-moderation, depending on the platform and the substance of the post. In addition, no bias in relation to No-aligned and Yes-aligned misinformation was detected.

Content that claimed that the referendum was unconstitutional

This sample of content was substantially under-moderated, and no platform demoted or labelled these types of posts. Two pieces of content became

unavailable on TikTok (one unclear-aligned and one no-aligned) and one on Facebook (unclear-aligned). This does not represent a pattern.



Content unavailability: under-moderation is where content is still unavailable.



Content labelling: under-moderation is where content is not labelled.



Content demotion: under-moderation is where content is not demoted.

Content under-moderated, in total any technique

80 % of content was under-moderated.

100 % of Yes-aligned, 83 % of No-aligned, 50 % of Unclear-aligned



🔍 Content unavailability: 80 % of content was under-moderated using this method.

100 % of Yes-aligned, 83 % of No-aligned, 50 % of Unclear-aligned

🔍 Content labelling: 100 % of content was under-moderated using this method.

100 % of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

🔍 Content labelling: 100 % of content was under-moderated using this method.

100 % of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

Content under-moderated, in total any technique

94 % of content was under-moderated.

– of Yes-aligned, 100 % of No-aligned, 90 % of Unclear-aligned



🔍 Content unavailability: 94 % of content was under-moderated using this method.

– of Yes-aligned, 100 % of No-aligned, 90 % of Unclear-aligned

🔍 Content labelling: 100 % of content was under-moderated using this method.

– of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

🔍 Content labelling (Videos only): 100 % of content was under-moderated using this method.

– % of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

Content under-moderated, in total any technique

100 % of content was under-moderated.

– of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned



🔍 Content unavailability: 100 % of content was under-moderated using this method.

– of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

🔍 Content labelling: 100 % of content was under-moderated using this method.

– of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

🔍 Content labelling (Videos only): 100 % of content was under-moderated using this method.

– % of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

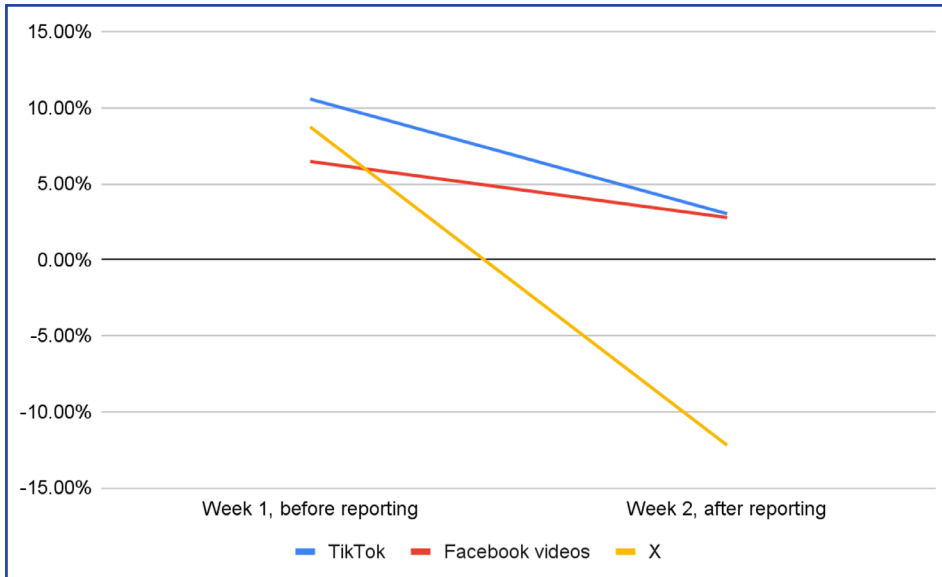


Figure 7: The week-on-week growth rate of content that contains misinformation regarding the constitutionality of the referendum according to platform. The real reduction in views on X was most likely due to the removal of bot accounts that were inflating views.

About the content that became unavailable, labelled or demoted

On TikTok, two videos became unavailable that claimed the referendum was unconstitutional; it included a video interview with a pundit who claims to have had a recent legal victory in the High Court. Three other videos that feature this interview and make the same claim were still

available (see Figure 8). This strongly suggests that TikTok is not moderating this content, meaning that the user probably removed this video themselves or deleted their account. Regardless, we have included it as a potential action from a platform in fairness to TikTok.

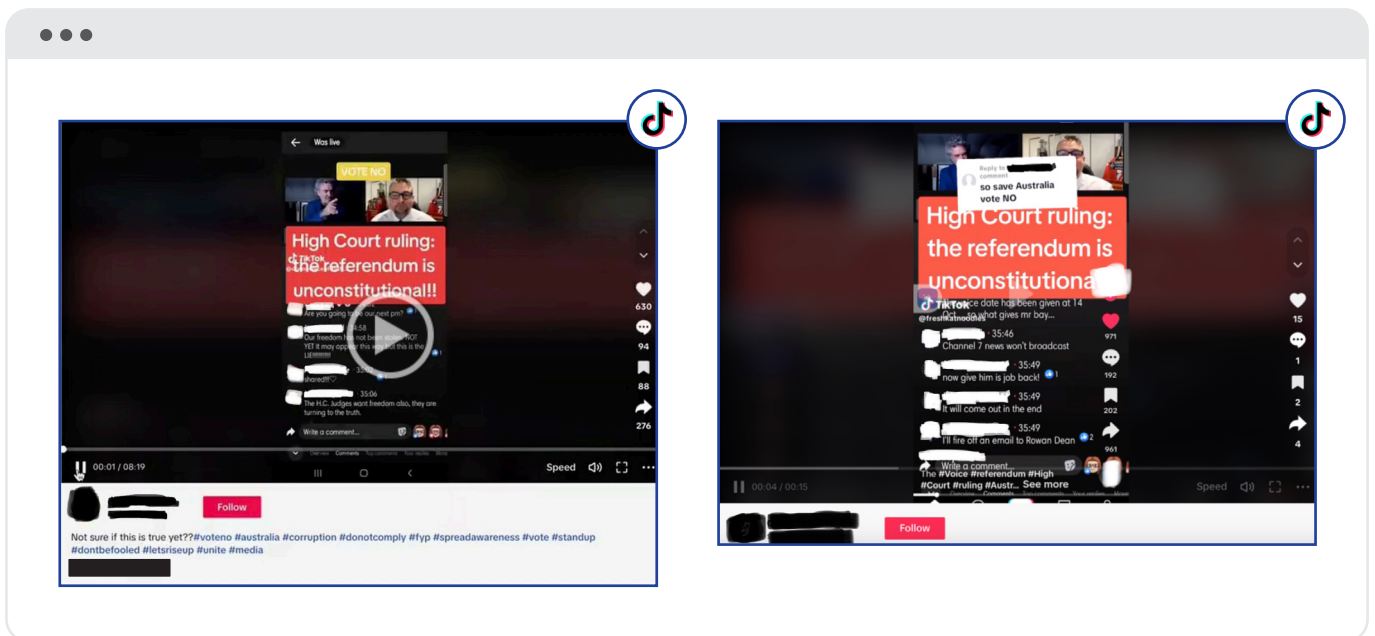


Figure 8: The video that is unavailable on TikTok (left), and a comparable video that remains on the platform (right).

On Facebook, one post was unavailable that also claimed the High Court had ruled against the referendum’s legality. Four other posts making similar claims remained visible. This strongly suggests that Facebook is not moderating this content and that

the user probably removed this post themselves or deleted their account. Regardless, we have included it as a potential action from a platform in fairness to Facebook.

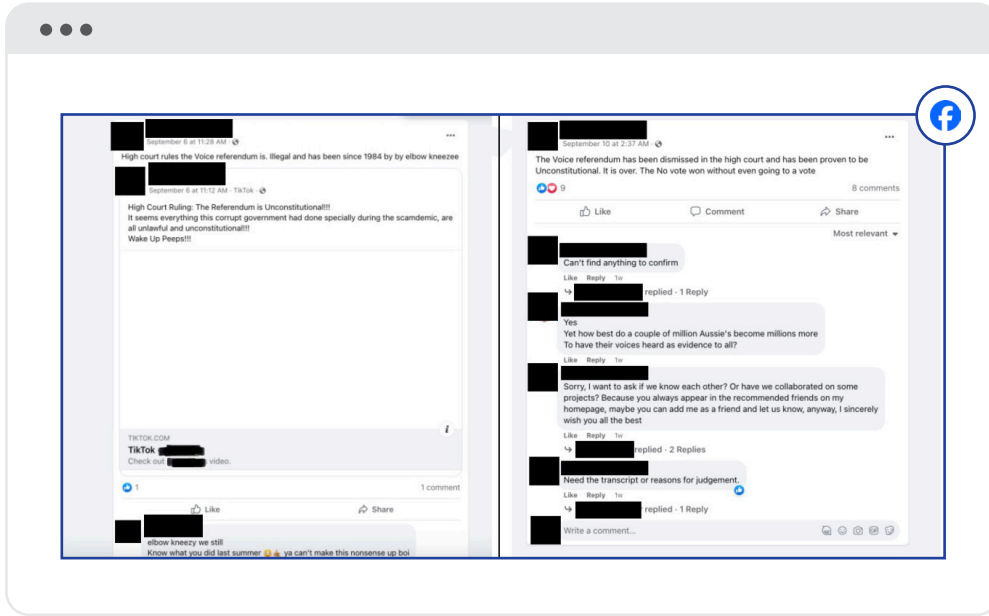


Figure 9: The post that is unavailable on Facebook (left), and a comparable video that remains on the platform (right).

Content that claimed that the referendum was unconstitutional

This sample of content was also substantially under-moderated. No platform demoted or labelled this content adequately. One piece of content became unavailable on TikTok and another on Facebook, which were No-aligned and Unclear-aligned but

this does not represent a trend. Six pieces of content on X appear to have been demoted, which were Yes-aligned, No-aligned and Unclear-aligned. One piece of content was labelled on X, which was unclear-aligned.



Content unavailability:
under-moderation is where content is still unavailable.



Content labelling:
under moderation is where content is not labelled.



Content demotion:
under-moderation is where content is not demoted.

Content under-moderated, in total any technique

92 % of content was under-moderated.

– % of Yes-aligned, **92 %** of No-aligned, **100 %** of Unclear-aligned



Content unavailability: **92 %** of content was under-moderated using this method.

– % of Yes-aligned, **92 %** of No-aligned, **100 %** of Unclear-aligned

Content labelling: **100 %** of content was under-moderated using this method.

– % of Yes-aligned, **100 %** of No-aligned, **100 %** of Unclear-aligned

Content demotion: **100 %** of content was under-moderated using this method.

– % of Yes-aligned, **100 %** of No-aligned, **100 %** of Unclear-aligned

Content under-moderated, in total any technique

94 % of content was under-moderated.

– % of Yes-aligned, 100 % of No-aligned, 94 % of Unclear-aligned



🔒 **Content unavailability:** 94 % of content was under-moderated using this method.

– % of Yes-aligned, 100 % of No-aligned, 94 % of Unclear-aligned

🔍 **Content labelling:** 100 % of content was under-moderated using this method.

– % of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

📺 **Content labelling:** No Facebook videos included in sample

Content under-moderated, in total any technique

75 % of content was under-moderated.

0 % of Yes-aligned, 79 % of No-aligned, 75 % of Unclear-aligned



🔒 **Content unavailability:** 100 % of content was under-moderated using this method.

100 % of Yes-aligned, 100 % of No-aligned, 100 % of Unclear-aligned

🔍 **Content labelling:** 96 % of content was under-moderated using this method.

100 % of Yes-aligned, 100 % of No-aligned, 88 % of Unclear-aligned

📺 **Content labelling (Videos only):** 79 % of content was under-moderated using this method.

0 % of Yes-aligned, 79 % of No-aligned, 88 % of Unclear-aligned

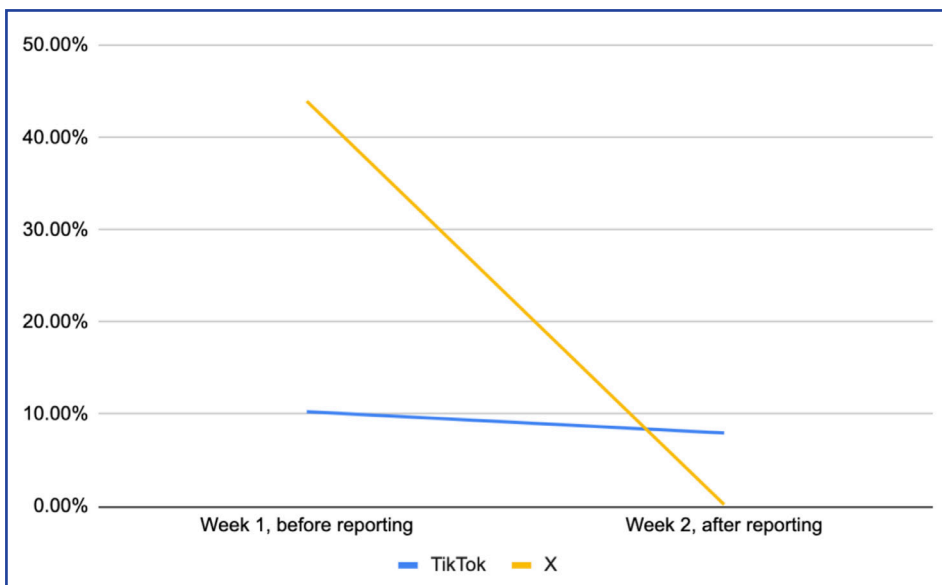
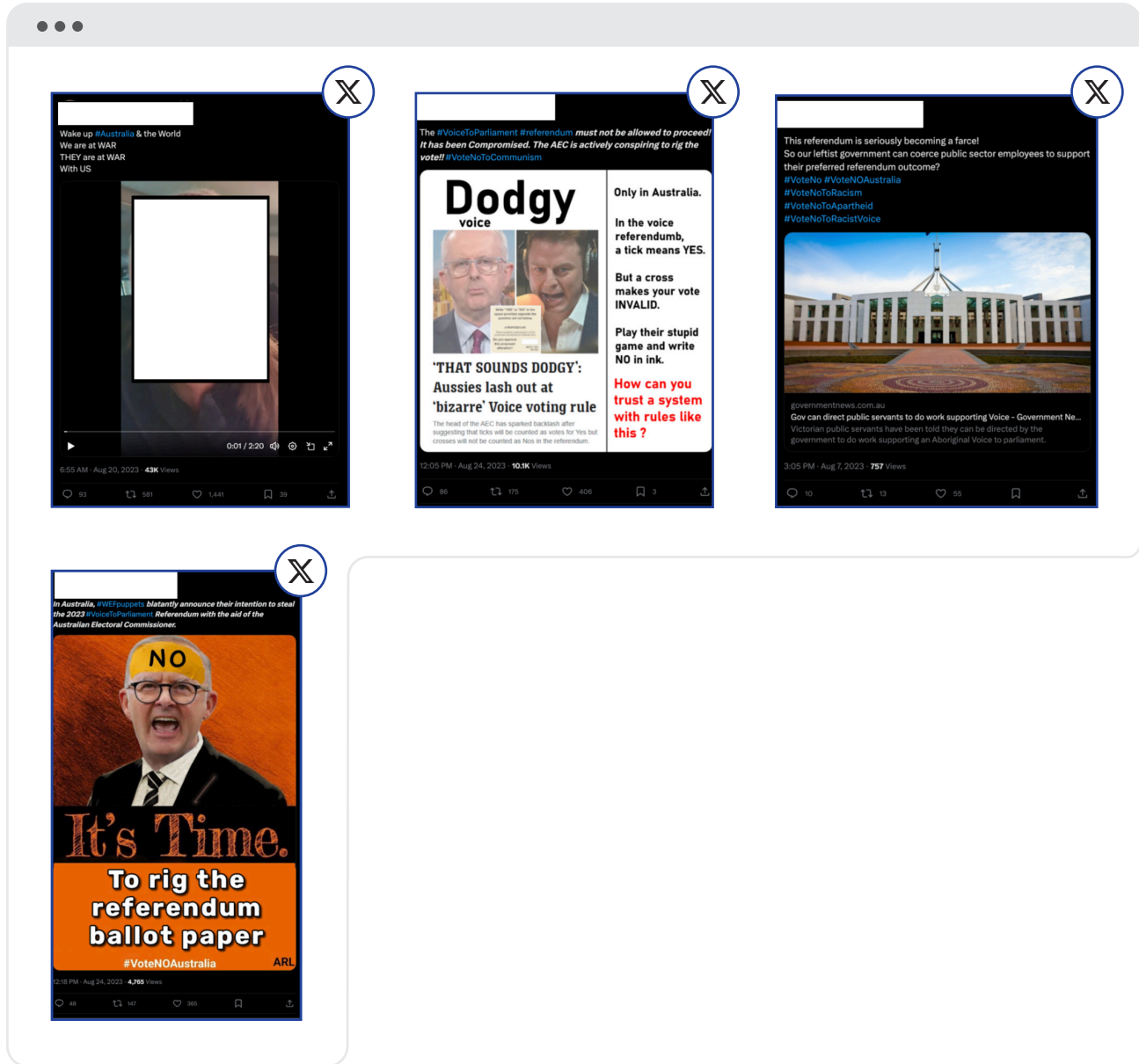


Figure 10: The week-on-week growth rate of misleading content regarding unproven claims of electoral irregularities by platform. Note, there were no Facebook videos in this sample to analyse growth.

About the content that became unavailable, labelled or demoted

Below are examples of content that **were** removed, labelled, or demoted.



Conclusions

This research found limited evidence of over-moderation on the specified platforms. The techniques used in this research produce overestimates, but even these overestimates range from **0.25%** on Facebook to **2%** on X. There is therefore limited evidence of bias, but X may over-moderate #VoteNo content, while Facebook appears to favour #VoteNo content in its video-recommender algorithm to a five-fold magnitude.

This project highlights that misinformation is substantially under-moderated across all three platforms. Misinformation regarding electoral processes, that violates each platform's community guidelines, was not removed when the companies became aware of it. In fact, between **75 to 100%** of misinformation was subject to under-moderation, depending on the platform and substance of the misinformation. There was limited evidence to support claims of systemic bias between No-aligned, Yes-aligned and Unclear-aligned misinformation, with platforms under-moderating all categories of misinformation. Caution should be used when interpreting these results however, due to the low numbers of Yes-aligned content found in this 'point-in-time' sample.

More research, which uses data provided by social media platforms themselves, needs to be conducted in order to make more verifiable claims. At best, due to the methods available to us, this investigation provides overestimates of potential over-moderation.

This research also suggests that Digi's **Australian Code of Practice on Disinformation and Misinformation** might not be effectively preventing under-moderation of content. Furthermore, as the Appendix suggests, a number of transparency reports struggle to actively characterise the scale of the issues that have been illuminated during this research project.



Appendix 1: Digi's Code and platforms' community guidelines in more detail

The Australian Code of Practice on Disinformation and Misinformation ('the Code')¹⁸

There are two compulsory objectives for signatories under Digi's Code; these are summarised by Reset. Tech below. There are also five additional optional commitments in the Code, which are not summarised here, but they are available on Digi's website.

Objective 1: Provide safeguards against harms that may arise from disinformation and misinformation.¹⁹

Outcome 1a: Signatories contribute to reducing the risk of harms that may arise from the propagation of disinformation and misinformation on digital platforms by adopting a range of scalable measures

Signatories will develop and implement measures, which aim to reduce the propagation of and potential exposure of users of digital platforms to disinformation and misinformation.

Measures implemented may include, by way of example rather than limitation:

- A. Policies and processes that require human review of user behaviours or content that is available on digital platforms (including review processes that are conducted in partnership with fact-checking organisations);
- B. Labelling false content or providing trust indicators of content to users;
- C. Demoting the ranking of content that may expose users to disinformation and misinformation;

- D. Removal of content that is propagated by inauthentic behaviours;
- E. Providing transparency about actions taken to address disinformation and misinformation to the public and/or users as appropriate;
- F. Suspending or disabling accounts of users that engage in inauthentic behaviours;
- G. The provision or use of technologies to identify and reduce inauthentic behaviours that can expose users to disinformation, such as the algorithmic review of content and/or user accounts;
- H. The provision or use of technologies that assist digital platforms or their users to check authenticity/accuracy or to identify the provenance/source of digital content;
- I. Exposing metadata to users about the source of content;
- J. Enforcing published editorial policies and content standards;
- K. Prioritising credible and trusted news sources that are subject to a published editorial code (noting that some signatories may choose to remove or reduce the ranking of news content which violates their policies);
- L. Partnering and/or providing funding for fact checkers to review digital content and

¹⁸ Digi 2022 Australian Code of Practice on Disinformation and Misinformation <https://digi.org.au/wp-content/uploads/2022/12/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL--December-22-2022.docx.pdf>

¹⁹ Digi 2022 Australian Code of Practice on Disinformation and Misinformation <https://digi.org.au/wp-content/uploads/2022/12/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL--December-22-2022.docx.pdf>, 5.8 - 5.14

M. Providing users with tools that enable them to exclude their access to certain types of digital content.

Outcome 1b: Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by signatories under this

Signatories will implement and publish policies and procedures and appropriate guidelines or information relating to the prohibition and/or management of user behaviours and/or content that may propagate disinformation and/or misinformation via their services or products.

Outcome 1c: Users can report content or behaviours to signatories that violate their policies (as above) through publicly available and accessible reporting tools

Signatories will implement and publish policies, procedures and appropriate guidelines that will enable users to report the types of behaviours and content that violates their policies (as above).

In implementing the commitment, signatories recognise that the terms disinformation and misinformation may be unfamiliar to users and thus policies and procedures aimed at achieving this outcome may specify how users may report a range of impermissible content and behaviours on digital platforms.

Outcome 1d: Users will be able to access general information about signatories' actions in response to reports made (using the tools above)

Signatories will implement and publish policies, procedures and/or aggregated reports (including summaries of user-reports made) regarding the detection and removal of content that violates platform policies, including (but not necessarily limited to) content on their platforms that qualifies as misinformation and/or disinformation.

Outcome 1e: Users will be able to access general information about signatories' use of recommender systems and have options relating to content suggested by recommender systems

Signatories that provide services (other than search engines) where the primary purpose is to disseminate information to the public and use recommender systems should commit to:

- A. Making information available to end-users about how they work to prioritise information that end-users may access on these services and
- B. Providing end-users with options that relate to content suggested by recommender systems that are appropriate for the service.

Note: for example, the comments section provided under news stories published by an on-line newspaper would be ancillary to the main service represented by the publication of news under the editorial responsibility of the publisher and is therefore not subject to this commitment.

Objective 7: (The final compulsory objective) Signatories publicise the measures they take to combat disinformation and misinformation.²⁰

Outcome 7: The public can access information about the measures signatories have taken to combat disinformation and misinformation

All signatories will conduct an investigation and publish a transparency report that contains information relating to the measures they have undertaken to combat disinformation and misinformation.

In addition, signatories will publish additional information detailing their progress in relation to **Objective 1** and any additional commitments they have made in line with this code.

²⁰ Digi 2022 Australian Code of Practice on Disinformation and Misinformation <https://digi.org.au/wp-content/uploads/2022/12/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL--December-22-2022.docx.pdf>, 5.30 - 5.32

Signatories may fulfil their commitment by providing additional reports and/or public updates on areas like content removals, open-data initiatives, research reports, media announcements, user-data requests and business-transparency reports. Examples of such information could include, by way of example rather than limitation, blog posts, white papers, in-product notifications, transparency reports, help centres or other websites.

Under the Code, platforms that sign on are required to 'develop and implement measures which aim to reduce the propagation of and potential exposure of users of digital platforms to Disinformation and Misinformation'. Below, Reset.Tech summarises the relevant sections of each platform's international policies, specifically the community guidelines of each platform, describing the measures they have committed to undertake.

TikTok's policies and reports to Digi regarding content moderation

TikTok's community guidelines state that it removes content that violates their rules.²¹ This includes misinformation that can cause significant harm, as described below:

We do not allow inaccurate, misleading, or false content that may cause significant harm to individuals or society, regardless of intent. Significant harm includes physical, psychological, or societal harm, and property damage.

We do not allow misinformation about civic and electoral processes, regardless of intent. This includes misinformation about how to vote, registering to vote, eligibility requirements of candidates, the processes to count ballots and certify elections, and the final outcome of an election. Content is ineligible for the 'For You Feed' if it contains unverified claims about the outcome of an election.



Content on TikTok that is misleading regarding electoral processes, such as claims of unconstitutionality or unproven allegations of electoral irregularities, should fall into the category of civic and electoral process misinformation. According to its own community guidelines, TikTok should remove all of the content highlighted during our under-moderation experiment.

TikTok is a signatory to the **Australian Code of Practice on Disinformation and Misinformation**.²² In their first annual transparency report,²³ TikTok described the impact of their content moderation system on election misinformation (excerpts of this are included below). TikTok describes removing 132 video-based examples of election misinformation throughout 2022, and outlined how users can report offensive content. Our research found significant levels of under-moderation that are not addressed in this transparency report and that are not consistent with these policies.

²¹ TikTok 2023 Civic and election integrity <https://www.tiktok.com/community-guidelines/en/integrity-authenticity/>

²² Digi, Australian Code of Practice on Disinformation and Misinformation, 2022, <https://digi.org.au/wp-content/uploads/2022/12/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL--December-22-2022.docx.pdf>

²³ TikTok, Annual Transparency Report, 2022, <https://digi.org.au/wp-content/uploads/2023/05/TikTok-2022-Annual-Transparency-Report.pdf>



Content removal

Election-related policy violations remain relatively rare in Australia, and they are generally limited to the period surrounding the elections themselves. This signals the robustness of both the systems we employ and the additional supporting measures that TikTok adopts to ensure election integrity, such as dedicated escalation channels and third-party fact-checking. **Fig. 5** shows the monthly number of videos removed for violating our misinformation policies over 2022 in Australia. Compared to 2021, when 12,582 medical misinformation videos were removed, TikTok removed 14,520 videos violating medical misinformation policy in AU in 2022.

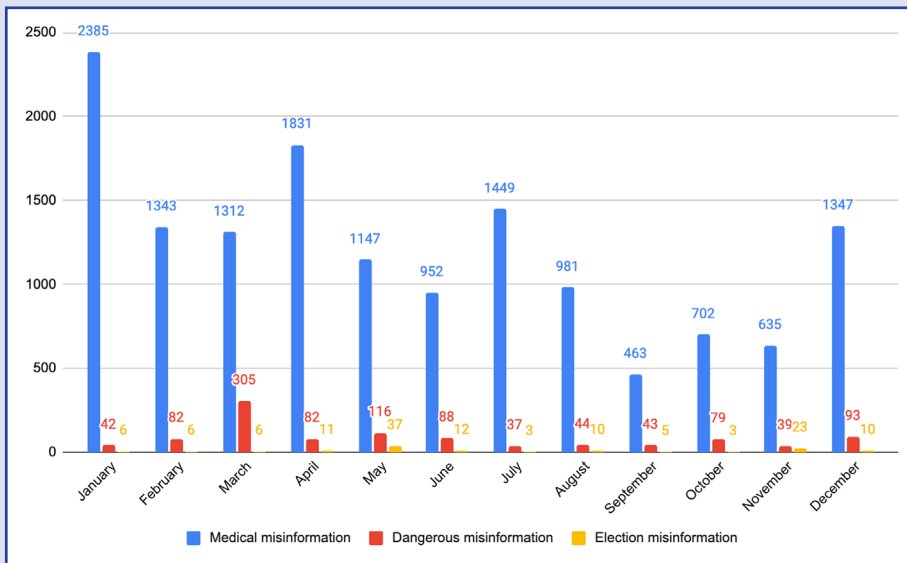


Fig. 5 (in TikTok’s report): Number of monthly removed videos violating misinformation policies in Australia in 2022.

Strengthening our resources to address misinformation and disinformation

We continue to build on our significant investments in the capacity of our teams, processes and technology to counter the spread of misinformation and disinformation and address threats that may arise from known risks, emerging violative trends, or which reflect country-specific incidents. Our Trust & Safety teams number over 40,000 people globally, including subject matter experts tasked with developing, implementing and reviewing TikTok’s “Integrity and Authenticity” policies. We continue to refine and adapt our activities and processes in [real time] to identify and combat inauthentic content. Recently, this has included implementing additional measures to detect and address violative livestreamed content. Our 24/7 moderation capacity is supported by our third-party fact-checking partners who review and verify flagged content.

These partnerships help minimise risks of [mis-moderation] and ensure that moderation decisions are based on independent assessments by accredited third-party experts. This is critical for issues involving public health or civic processes, where the consequences of disseminating misinformation are particularly acute. In all cases where information is found to be false or deceiving, we take immediate measures to remove the content in line with our Community Guidelines. In 2022, we undertook a number of additional measures to counter misinformation, including:

- improved detection of known misleading audio and imagery to reduce manipulated content spread;



- a database of previously fact-checked claims to help misinformation moderators make swift and accurate decisions; and
- a proactive detection program for our fact-checkers to flag new and evolving trends they discover across the internet. This allows TikTok to retrieve and remove violative content from our platform. Since starting this program last quarter, we have identified 33 new misinformation claims, resulting in the removal of 58,000 videos from the platform globally (as of April 2023).

Outcome 1c: Users can easily report offending content

TikTok users can report content they believe violates our Community Guidelines at any time within the app. ‘Misinformation’ is a distinct reporting category within our platform and appears as a prominent reporting option when users attempt to report content. Users are also prompted to specify the kind of misinformation they are reporting, as is shown in **Fig. 7** below [omitted by Reset.Tech Australia from this summary, but available online].

In addition to reporting misinformation in video content, users can also report misinformation across other features of the platform. Users can report comments on videos, direct messages they receive from other users, accounts, sounds, hashtags and autosuggestions generated when they search for something on TikTok. Users can also report LIVE videos and comments on livestreams if they encounter content that violates our Community Guidelines.

Non-users can submit a report to us about content on our platform using the out-of-app reporting form. We have also established systems (including the Community Partner Channel) that enable direct reports of potential misinformation for our immediate review and action. The TikTok Community Partner Channel provides additional options for users to report content via trusted community organisations, who connect with community clients by sharing expertise and support on specific issues. We also publish instructions for our publicly accessible reporting tools on our website.

Facebook’s policies and reports to Digi regarding content moderation

Facebook’s community guidelines²⁴ suggest that it removes misinformation and disinformation when:

- It is likely to contribute to the risk of imminent physical harm, including the risk of violence to people, harmful health-related misinformation like the dissemination of falsehoods about vaccines or the promotion of miracle cures;
- It is highly deceptive media, such as deepfakes or
- It is likely to directly contribute to interference with the functioning of political processes, as **detailed below**.

²⁴ Meta, Community Standards: Misinformation, 2023, <https://transparency.fb.com/en-gb/policies/community-standards/misinformation/>



In an effort to promote election and census integrity, we remove misinformation that is likely to directly contribute to a risk of interference with people's ability to participate in those processes. This includes the following:

- Misinformation about the dates, locations, times and methods for voting, voter registration or census participation.
- Misinformation about who can vote, qualifications for voting, whether a vote will be counted and what information or materials must be provided in order to vote.
- Misinformation about whether a candidate is running or not.
- Misinformation about who can participate in the census and what information or materials must be provided in order to participate.
- Misinformation about government involvement in the census, including, where applicable, that an individual's census information will be shared with another (non-census) government agency.
- Content falsely claiming that the US Immigration and Customs Enforcement (ICE) is at a voting location.
- Explicit false claims that people will be infected by COVID-19 (or another communicable disease) if they participate in the voting process."

However, they go on to state that 'For all other misinformation, we focus on reducing its prevalence or creating an environment that fosters a productive dialogue.'

Content on Facebook that is misleading regarding electoral processes, such as claims of unconstitutionality or unproven allegations of electoral irregularities, should fall into the final category of 'all other misinformation', and the platform should focus on reducing its prevalence. This requires fact-checkers to have investigated content before triggering the removal process. The content included in this brief experiment addresses narratives that have been fact-checked and determined to be false. According to its community guidelines, Facebook should 'reduce the prevalence' of the content involved in our under-moderation experiment.

However, Meta Australia suggests that Meta 'removes election-related misinformation that may constitute voter fraud or interference under our

Co-ordinating Harm and Promoting Crime policy.²⁵ According to this statement, the content involved in our under-moderation experiment should have already been deleted. It is therefore unclear if electoral process misinformation is removed or demoted, so we monitored to detect both eventualities.

Facebook is a signatory to Digi's **Australian Code of Practice on Disinformation and Misinformation**.²⁶ In their first annual transparency report,²⁷ Facebook describes the impact of their content moderation processes on the 2022 federal election (excerpts below). Our research found significant levels of under-moderation that are inconsistent with the characterisations made in this transparency report, as well as their own policies.

²⁵ Meta, Meta response to the Australian disinformation and misinformation industry code, 2023, https://digi.org.au/wp-content/uploads/2023/05/Meta_2023-AU-Misinformation-Transparency-report_v1.pdf

²⁶ Digi, Australian Code of Practice on Disinformation and Misinformation, 2022, <https://digi.org.au/wp-content/uploads/2022/12/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL--December-22-2022.docx.pdf>



In 2022, some highlights of our work include:

Implementing a suite of measures in the lead up to the 2022 Australian federal election to proactively detect and remove content that breaches our policies, combat misinformation, harmful content and coordinated inauthentic behaviour, and promote civic participation.

As a result of these efforts, during the election campaign (between April 1 and June 30, 2022):

- We took action on over 25,000 pieces of content across Facebook and Instagram for violating our Harmful Health Misinformation policies.
- We displayed warnings on over 3 million distinct pieces of content on Facebook (including reshares) based on articles written by our third-party fact-checking partners.
- We took action on over 91,000 pieces of content on Facebook and over 40,000 pieces of content on Instagram in Australia for violating our hate speech policies.
- We took action on over 200,000 pieces of content on Facebook and over 46,000 of content on Instagram in Australia for violating our Community Standards on violence and incitement.
- We rejected around 17,000 ads for not complying with our political and social issue ads enforcement policies.

X's (formerly Twitter's) policies and reports to Digi regarding content moderation

X's community guidelines²⁸ state that it removes or labels political misinformation and disinformation that misleads people about electoral participation,

that is intended to suppress turnout or intimidate or misleads about the outcomes of elections (*details below*).



Misleading information about how to participate

We may label or remove false or misleading information about how to participate in an election or other civic process. This includes but is not limited to:

- *misleading information about procedures to participate in a civic process (for example, that you can vote by Post, text message, email, or phone call in jurisdictions where these are not a possibility);*

²⁷ Meta, Meta response to the Australian disinformation and misinformation industry code, 2023, https://digi.org.au/wp-content/uploads/2023/05/Meta_2023-AU-Misinformation-Transparency-report_v1.pdf

²⁸ X, Civic integrity misleading information policy, 2023, <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>



- misleading information about requirements for participation, including identification or citizenship requirements;
- misleading claims that cause confusion about the established laws, regulations, procedures, and methods of a civic process, or about the actions of officials or entities executing those civic processes; and
- misleading statements or information about the official, announced date or time of a civic process.

Suppression and intimidation

We may label or remove false or misleading information intended to intimidate or dissuade people from participating in an election or other civic process. This includes but is not limited to:

- misleading claims that polling places are closed, that polling has ended, or other misleading information relating to votes not being counted;
- misleading claims about police or law enforcement activity related to voting in an election, polling places, or collecting census information;
- misleading claims about long lines, equipment problems, or other disruptions at voting locations during election periods;
- misleading claims about process procedures or techniques which could dissuade people from participating; and
- threats regarding voting locations or other key places or events (***note that our violent threats policy may also be relevant for threats not covered by this policy.***)

Misleading information about outcomes

We may label or remove false or misleading information intended to undermine public confidence in an election or other civic process. This includes but is not limited to:

- disputed claims that could undermine faith in the process itself, such as unverified information about election rigging, ballot tampering, vote tallying, or certification of election results; and
- misleading claims about the results or outcome of a civic process which calls for or could lead to interference with the implementation of the results of the process, e.g. claiming victory before election results have been certified, inciting unlawful conduct to prevent the procedural or practical implementation of election results (***note that our violent threats policy may also be relevant for threats not covered by this policy.***)

Content on X that is misleading regarding electoral processes, such as claims of unconstitutionality or unproven allegations of electoral irregularities, should fall into the final category, of misleading information about outcomes as it undermines faith in the process itself. According to its own community guidelines, X should label or remove this information. As underscored above, the claims made in the content covered by this research have been extensively fact-checked, and X should therefore label or remove all of the content flagged up in our under-moderation experiment.

X are a signatory to Digi's **Australian Code of Practice on Disinformation and Misinformation**.²⁹ In their first annual transparency report,³⁰ X mentioned how they avoid under-moderation by making policies and tools for users to report violative content (excerpts below). Our research found significant levels of under-moderation that are inconsistent with the characterisations made in this transparency report. We consequently believe X's report is fallacious with regard to the current referendum. Users can no longer report electoral misinformation, including examples relating to the Voice referendum.



Outcome 1c: Users can report content and behaviours to Signatories that violate their (content) policies through publicly available and accessible reporting tools.

We have a range of dedicated tools available for all of our users to report content that may violate our rules and policies and now meaningfully contribute directly to the service via Community Notes by adding or rating Notes for tweets which may be made public. As reported, Community Notes were visible in Australia in Q4 2022 and were made available for Australian contributors in Q1 2023.

On reporting Twitter has publicly available and accessible robust reporting forms, both in-app, on web and via our Help Centre where users can report 24/7, and they will be notified once our team has reviewed and taken enforcement action, where appropriate.

Users can report Tweets, Lists, and Direct Messages that are in violation of our Rules or our terms of service (TOS). For Community Notes, Twitter created the ability for tweet authors to request additional review if they disagree that a Community Note is "helpful" or provides important context to their tweet. We've made that process publicly available for review, with simple information for how and where to report.

²⁹ Digi, Australian Code of Practice on Disinformation and Misinformation, 2022, <https://digi.org.au/wp-content/uploads/2022/12/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL--December-22-2022.docx.pdf>

³⁰ X, Australian Code of Practice on Disinformation and Misinformation Twitter Annual Transparency Report, 2022, https://digi.org.au/wp-content/uploads/2023/05/ACPDM_report_2022_Twitter-052823_DIGI.pdf





Is content over- or under-moderated in the Voice referendum debate?

An experimental evaluation