

How do platforms respond to user-reports of electoral process misinformation?

An experimental evaluation from the lead-up to Australia's referendum



Summary

This rapid investigation set out to explore whether platforms remove electoral process misinformation when they are made aware of it via user-reporting. We found, reported and monitored a small number of posts on TikTok (25), Facebook (24) and X, formerly Twitter (50), that contained clear electoral process misinformation. This content largely centred around claims that Australian elections had been rigged, that ballots had or would be stolen, or that the Voice referendum vote was invalid or illegal. Electoral process misinformation stands to harm both the Yes and No campaigns.


According to each platform's community guidelines, this type of content, once detected, should be:


- **TikTok:** Removed.
- **Facebook:** Demoted in prevalence.¹
- **X:** Either removed or labelled.

However, we found that none of the platforms are effectively enforcing their community guidelines, nor are they implementing meaningful responses based on their requirements under the Australian Code of Practice on Disinformation and Misinformation. Specifically:

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

This research suggests that:


 TikTok's content removal or labelling rate without reporting is at best **4%**² in a week.

 Facebook's content removal or labelling rate without reporting is at best **4%**³ in a week.

¹ We would assume this would include labelling and deamplifying.

² These are 'best case' estimations, as it is unclear if content that became unavailable at any stage of the research was taken down by users or the platforms themselves, or if other users had reported the content first.

³ These are 'best case' estimations, as it is unclear if other users had reported the content first.

 X's content removal or labelling rate without reporting is **0%** in a week.

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

This research suggests that:

 TikTok's content removal or labelling rate after reporting is at best **32%** in a fortnight.

 Facebook's content removal or labelling rate after reporting is **0%** in a fortnight.

 X's content removal or labelling rate after reporting is **0%** in a fortnight.

The findings also show that 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.

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.

Note, the majority of the misinformation content reported to the platforms is still available online and is unlabelled at the time of publication. Further, this content continues to grow in views.



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.**

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

Funding
 Susan McKinnon Foundation

Table Of Contents

- Summary 2
- Contents 3
- Introduction 4
- What we did: Methods 5
- What should happen: Code of Practice 8
- What should happen: Platform rules 10
- What happened: Findings 12
 - Content that became unavailable and labelled 16
- Conclusions 22
- Appendix 1: Digi’s Australian Code of Practice on Disinformation and Misinformation 23
- Appendix 2: Platforms’ community guidelines, in more detail 25

Introduction

Under the voluntary *Australian Code of Practice on Disinformation and Misinformation* ('the Code'),⁴ platforms that sign on have obligations to develop and implement measures that aim to reduce the propagation of and exposure of users to misinformation and disinformation. This includes:

- developing and publishing policies around misinformation and disinformation to inform users about the types of content that will be prohibited and how the platform will manage this content.⁵
- providing users with tools to report content that violates these policies.⁶

In this rapid investigation, we set out to explore how platforms respond to user-reports of misinformation and disinformation that violates their stated policies, using electoral process misinformation in the context of the Voice referendum as a case study. Electoral process misinformation stands to harm both the Yes and No

campaigns, and this research was not designed to document bias in response.

This is a timely investigation in the context of the exposure draft of the *Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill*. As it is currently drafted, the Bill proposes providing ACMA with additional powers to gather information -- but the exercise of these powers rest somewhat on the assumption that the Code is working. This is another example of a co-regulatory approach, where tech companies are allowed to draft their own 'rules' and codes and regulators then enforce around them. Previous research has documented the risks and failures of a co-regulatory approach in tech regulation.⁷

This investigation explores the impact of the Code, specifically using three platforms' responses to user-reports of misinformation and disinformation as a case study. It raises questions around whether the Code is providing meaningful protections to Australians.



Why monitor during the Voice referendum?

The Voice referendum is both a uniquely important event in Australia's history, and provides a valuable, timely case study for evaluating platform responses to misinformation and disinformation. Specifically:

- It is distinctly Australian, which means we can monitor international platforms' responses to an Australian issue, meaning there is less potential conflation with global responses.
- It is an Australian electoral process, as that all the features of electoral misinformation and disinformation will apply, and learnings can be made for future elections.
- To an extent, it is more narrowly defined than a broader election, where 'electoral content' and 'general current affairs content' can become harder to differentiate between.

⁴ 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 The Code, 5.10

⁶ The Code, 5.11–5.12

⁷ Reset.Tech Australia 2023 *How outdated approaches to regulation harm children and young people and why Australia urgently needs to pivot* <https://au.reset.tech/news/how-outdated-approaches-to-regulation-harm-children-and-young-people-and-why-australia-urgently-needs-to-pivot/>

What we did: Methods

This rapid experiment set out to explore how platforms respond to user-reports of misinformation about electoral processes in Australia.

1

Finding content to report and monitor:

[We found 99 pieces of content to monitor that included false or misleading claims about Australia's electoral process.](#) This content was largely focused around claims of rigged elections, stolen votes or AEC malpractice, and many of the posts had a focus on the referendum; they did not relate to discussion around ticks or crosses on ballot papers. The misinformation explored included posts that could support the Yes or No campaigns; although, in many cases it was unclear which campaign stood to benefit (e.g. when the validity of polling booths is questioned). This content was identified relatively quickly by searching or exploring accounts previously known to Reset.Tech. The content included:

- A. **25 pieces on TikTok.** In total, these 25 pieces had 246,123 views between them. The content centred around claims of election rigging and vote stealing, but it also included associated conspiracy claims that the AEC was corrupt, polling booths were invalid, ministers were using mind control techniques on voters, the referendum is illegal and that taking part is treasonous and will affect your citizenship because Australia is governed maritime law or a corporation or controlled by the WEF or UN, etc.
- B. **24 pieces of content on Facebook.**⁸ Only videos on Facebook include view counts publicly; three of the pieces of content monitored were videos, which had 169,166 views between them. The 24

posts on Facebook centred around claims that Australian elections were rigged or were going to be rigged and that ballots had been stolen, and a small number of posts focused on voter suppression, claiming that voting a particular way would lead to being de-banked, or contained calls to boycott voting because it was treasonous.

- C. **50 pieces of content on X.** In total, these 50 pieces had 70,238 views between them. The content centred around claims that Australian elections had been rigged and that the referendum would be rigged, with the occasional piece suggesting that the referendum process was illegal because Australia's constitution was invalid.

We looked at posts, reposts and duets, as well as comments, to monitor if the take-down approaches differed between the types of content.



Note on the method:

This was a rapid research investigation. We monitored a small sample of posts for a short period of time to deliver indicative findings in time to meet the Government's consultation deadline for the *Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill*. It provides, to our knowledge, the best available estimates regarding the response rates from platforms.

The small sample size presents limitations, and more extensive research, following more content for longer periods of time, would be needed to understand how generalisable these results are.

⁸ We intended to track 25 pieces, but a duplicate was included in error.



The content monitored in this experiment

The content monitored in this experiment across all three platforms relates to narratives that have been largely fact checked in Australia and routinely deemed to be false. Australian fact checkers regularly disprove claims of widespread corruption or ‘rigging’ within the Australian electoral system, for example;

- **Claims that ballots had been removed or stolen** in the NSW election, often used as a way to paint a picture of widespread election rigging, which have been deemed false.⁹ Likewise, claims about the Victorian election being fraudulent, often included in content that suggests all Australian elections have been rigged, has also been deemed false.¹⁰
- **Claims that the referendum is illegal or fraudulent**, which have been deemed false.¹¹
- **Claims that the referendum is going to be rigged** using electronic voting systems, which have been deemed false.¹²
- **Claims that the question asked in the referendum will be sneakily disguised as multiple questions**, so that if only one passes ‘all of them will pass’, which have been deemed false.¹³
- **Claims that the question on the ballot will be ‘rigged’ or sneakily written to create a new state or end Australia’s sovereignty**, so that voters do not really know what they are voting for etc. These claims have been deemed false.¹⁴
- **Claims that the referendum question will be sneakily written to trick voters into agreeing to end Indigenous sovereignty**, or that by ending Indigenous sovereignty, it will trick voters into **handing sovereignty to the UN**. These claims have been deemed false.¹⁵

Some posts monitored included multiple inaccuracies alongside electoral misinformation, such as frequently suggesting that the Yes and/or No campaigns received public funding,¹⁶ or that if passed, the referendum would lead to new states or supersede existing states.¹⁷

⁹ AAP 2023 *Removal of NSW ballot boxes isn’t evidence of election fraud* <https://www.aap.com.au/factcheck/removal-of-nsw-ballot-boxes-isnt-evidence-of-election-fraud/>

¹⁰ RMIT Fact Lab 2023 *No substance to claim that Victorian state election was rigged* <https://www.rmit.edu.au/news/factlab-meta/no-substance-to-claim-that-victorian-state-election-was-rigged>

¹¹ RMIT Fact Lab 2023 *The Voice referendum is not illegal* <https://www.rmit.edu.au/news/factlab-meta/voice-opponents-wrong-on-legality-of-referendum>

¹² RMIT Fact Lab 2023 *Electronic vote rigging and Voice-by-legislation claims prove baseless* <https://www.rmit.edu.au/news/factlab-meta/electronic-vote-rigging-and-voice-by-legislation-claims-baseless>

¹³ AAP 2023 *One simple answer to five questions claim* <https://www.aap.com.au/factcheck/one-simple-answer-to-five-questions-claim/>

¹⁴ AAP 2023 *Section 122 claim confuses voice’s proposed place in constitution* <https://www.aap.com.au/factcheck/section-122-claim-confuses-voices-proposed-place-in-constitution/>

¹⁵ RMIT Fact Lab 2023 *Indigenous Australians will not cede sovereignty under the Voice due to 1973 “change” to constitution* <https://www.rmit.edu.au/news/factlab-meta/voice-will-not-be-impacted>

¹⁶ AAP 2023 *Voice campaign funding claim short-changes the facts* <https://www.aap.com.au/factcheck/voice-campaign-funding-claim-short-changes-the-facts/>

¹⁷ AAP 2023 *Counter claims confuse Voice proposals* <https://www.aap.com.au/factcheck/section-122-claim-confuses-voices-proposed-place-in-constitution/>

2

Monitoring before reporting:

The content was monitored for a week. During this week, content was monitored to uncover:

- A. Removal rates, i.e. how much of the content was removed.
- B. Labelling rates, i.e. how much of the content was labelled.
- C. Growth rates, i.e. how many more views did the content attract.
- D. During this week, platforms automated systems or other users may have flagged the content for moderation or users themselves may have taken posts down, but this represents the best available estimate of 'organic' response rates.

3

Reporting content:

We reported all 99 pieces of content using the platforms' publicly available reporting mechanisms.

4

Monitoring after reporting:

We continued to track and monitor the content for two weeks after reporting, for growth, labelling and removal rates.

What should happen: Code of Practice

The **Australian Code of Practice on Disinformation and Misinformation** ('the Code')¹⁸ is the key policy instrument underpinning multiple large platforms' responsibilities towards Australian users. The obligations for platforms under the Code largely relate to the development of measures decided by platforms, with these measures subject to a proportionality test.

Content that is misleading around electoral processes, such as claims of rigged elections, stolen votes or AEC malpractice, presents a credible and serious threat to the integrity of Australia's democratic processes and is considered misinformation under the Code. However, the Code does not create an obligation to remove or demote misinformation and disinformation content about the electoral process.

The Code defines misinformation as:

"Misinformation means:

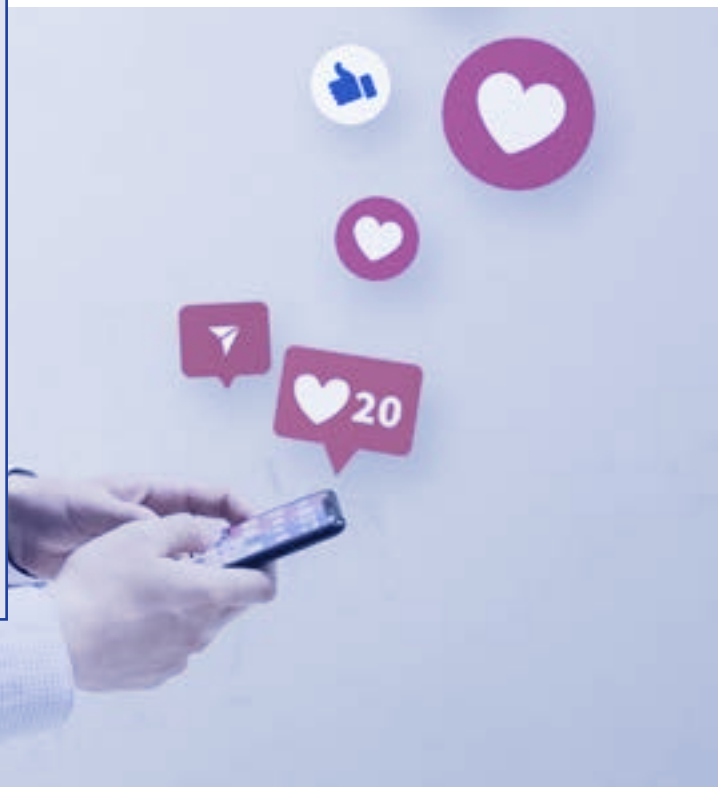
1. Digital content (often legal) that is verifiably false or misleading or deceptive;
2. Is propagated by users of digital platforms; and
3. The dissemination of which is reasonably likely to (but may not be clearly intended to) lead to harm".

"Harm here is defined as harm which poses a credible and serious threat to:

- I. Democratic, political and policy-making processes such as voter fraud, voter interference or voting misinformation or
- II. Public goods such as the protection of citizens' health, protection of marginalised or vulnerable groups, public safety and security of the environment".



Rather, according to the Code, signatories must simply 'develop and implement measures' that 'aim to reduce' the propagation of and potential exposure to misinformation and disinformation. This may include measures such as having policies and processes that require the human review of user behaviour and content, labelling false information or removing content propagated by inauthentic behaviours. It is left to individual platforms to decide. **(See Appendix 1 for more details.)**



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>

Mandatory commitments (see Figure 2)	Provide safeguards against harms that may arise from misinformation and disinformation
	Make and publish 'transparency reports'
Optional Commitments	Disrupt ads and monetisation of misinformation and disinformation
	'Tackle' inauthentic behaviour, specifically 'prohibit or manage' certain types of inauthentic behaviour
	Implement measures to enable users to 'make informed choices' about information
	Develop policies around making political advertising more transparent
	Support independent research on misinformation and disinformation, with Australian universities

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



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 to users on digital platforms. These may include, for example:

- Policies and process 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 how effective platforms' measures are, and progress made to realise the obligations under the Code

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

Figure 2: A summary of the mandatory commitments for signatories of the Code

What should happen: Platform rules

The first compulsory obligation under the Code envisages that each platform develops its own policies and implements its own measures to address misinformation. These policies often differ, but they are summarised below with excerpts included in Appendix 2.

- **In general**, TikTok pledges to remove electoral process misinformation or make some ineligible for the FYF; Facebook pledges to label or remove some of it, and

'reduce in prevalence' others after fact checking, and; X pledges to label or remove it (see Figure 3).

- Specifically for the **nature of the content in this experiment**, we would expect; TikTok to remove electoral process misinformation; Facebook to 'reduce it', and; X to label or remove it (see Figure 4).




Platform policies on misleading content around electoral processes in general		
 Remove, ineligibility for FYF	 Label, remove or reduce in prevalence	 Label or remove
<p><i>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 FYF if it contains unverified claims about the outcome of an election.</i>¹⁹</p>	<p><i>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.</i>²⁰</p> <p>Examples provided by Facebook of content that directly contributes to a risk of interference include dates, locations, times and methods for voting, voter eligibility, government involvement in the ballot measures (including sharing voter data), and whether votes are counted.</p> <p>Further Facebook states: <i>For all other misinformation, we focus on reducing its prevalence or creating an environment that fosters a productive dialogue.</i></p>	<p><i>We may label or remove false or misleading information about how to participate in an election or other civic process.</i>²¹</p> <p>Examples include procedures to participate, voter eligibility, methods of the process or actions of electoral officials.</p> <p><i>We may label or remove false or misleading information intended to undermine public confidence in an election or other civic process.</i></p> <p>Examples include unverified information about election rigging, ballot tampering, vote tallying, or certification of election results.</p>

Figure 3: A summary of platform policies regarding electoral process misinformation and disinformation (see Appendix 2 for more detail)

¹⁹ TikTok 2023 Civic and election integrity <https://www.tiktok.com/community-guidelines/en/integrity-authenticity/>

²⁰ Meta 2023 Community Standards: Misinformation <https://transparency.fb.com/en-gb/policies/community-standards/misinformation/>

²¹ X 2023 Civic integrity misleading information policy <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>




Platform policies on misleading content around the specific types of electoral process explored in this research		
 Removal	 Reduce in prevalence	 Label or remove
<p>TikTok has an expansive definition of electoral process misinformation. Misleading content around electoral processes, such as claims of rigged elections, stolen votes, or AEC malpractice on TikTok would fall into the category of civic and electoral process misinformation and, according to their community guidelines, should be removed when it is discovered.</p>	<p>Claims of rigging, stolen votes or AEC malpractice is likely to fall into the 'other' category of misinformation where Facebook focuses on reducing its prevalence.</p> <p>This requires fact checkers to have investigated content before Facebook takes action, which has happened for this body of content.</p> <p>It is unclear from the platform's guidelines what measures are taken to 'reduce prevalence', but we would assume this includes labelling this content, reducing its visibility and de-amplification.</p>	<p>X's definition of content that should be labelled or removed covers claims of rigging, stolen votes or AEC malpractice, as content intended to undermine public confidence in an election or other civic process. This means that X should, according to its community guidelines, label or remove this information when it is discovered.</p>

Figure 4: A summary of platform policies regarding the specific electoral process misinformation and disinformation included in this investigation.



What happened: Findings

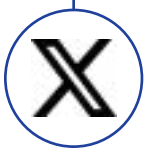
No platform responded adequately to user-reports. Violative content was still available, unlabelled and growing on each platform, after being reported (*see Figure 5*).



On TikTok: one post was removed before reporting (**4%**) and eight were removed after reporting (**32%**). This suggests that TikTok removes posts after reporting, in line with their policy. However, the response is inadequate as removal is the platform's stated response to electoral process misinformation and disinformation, as described in their community guidelines, and **the majority of posts were still available two weeks after reporting**. Further, the body of posts continued to grow relatively equally before and after reporting, which suggests there was no noticeable de-amplification. As far as this rapid experiment could detect, TikTok was the strongest performer in terms of meeting its commitments to users, as outlined in its community guidelines



On Facebook: one post was labelled at the time of reporting (**4%**), and none were labelled in the latter two weeks of monitoring. The majority of posts remained available and unlabelled after reporting. **The videos monitored continued to grow after reporting, albeit at a significantly slower pace**. This may be indicative of de-amplification, but it could also be a reflection of older video content spreading less quickly; a more detailed experiment with comparative videos is necessary to determine which of these options might be the case. However, the continued growth suggests that the platform is not entirely effective in responding to electoral process misinformation and disinformation by 'reducing its prevalence', as described in its community guidelines.



On X: no posts were labelled or removed at the time of reporting, and none were removed in the latter two weeks of monitoring. This suggests that X's response to reports and overall moderation is inadequate as the **majority of posts remained available and unlabelled after reporting**; labelling or removal is the platform's main response to electoral process misinformation and disinformation, as described in their community guidelines. The posts monitored continued to grow relatively equally before and after reporting, which suggests there was no noticeable de-amplification. As far as this rapid experiment could detect, X was the worst performer in this experiment in terms of meeting its commitments to users as outlined in its community guidelines.













		 Removal (See Figure 6)	 Labelling (See Figure 7)	 Growth (See Figure 8)
In total , across three weeks of monitoring	 25	9 ²²	0	Growth of 825 views
	 24	0	1 ²³	Facebook does not make view counts available for content, apart from videos. The three videos monitored grew by 4,345 views
	 50	0	0	Growth of 457 views
Initial week before reporting		1 ²⁴	0	The remaining 24 posts grew by 229 views
Estimate of 'organic' response rates		0	1	The 3 videos monitored grew by 2,425 views
		0	0	Growth of 132 views
Two weeks after reporting		8 ²⁵	0	The remaining 16 posts grew 298 views per week on average
Estimate of response to reporting		0	0	The three videos monitored grew 960 views per week on average
		0	0	Growth of 162 views per week on average

Figure 5: Platforms' response to the content

²² We cannot be sure if these were removed by the platform or by users.

²³ In addition to one that was already labelled at the start of the experiment.

²⁴ We cannot be sure if these were removed by the platform or by users.

²⁵ We cannot be sure if these were removed by the platform or by users.

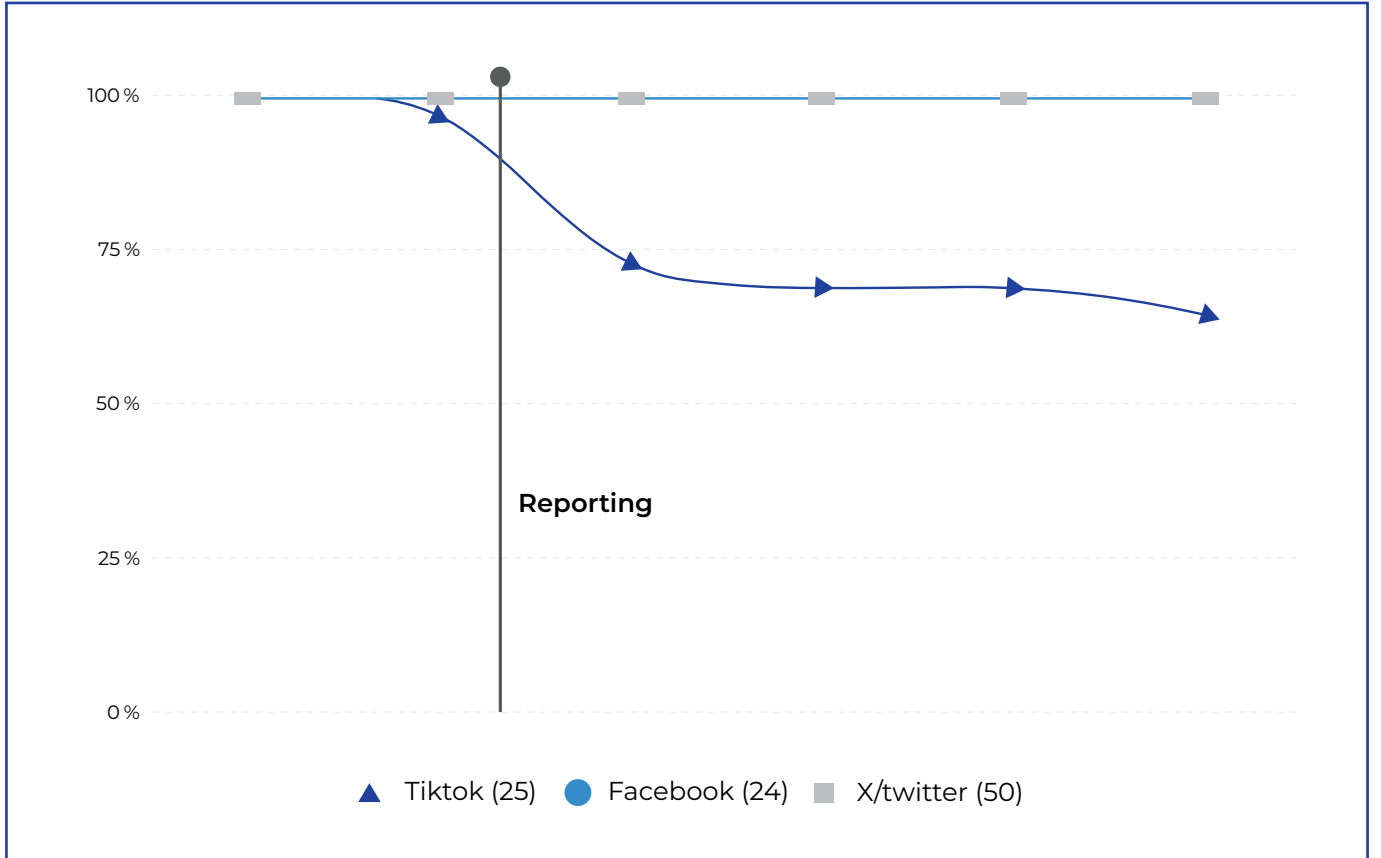


Figure 6: Content removal. The percentage of content available after reporting, on each platform

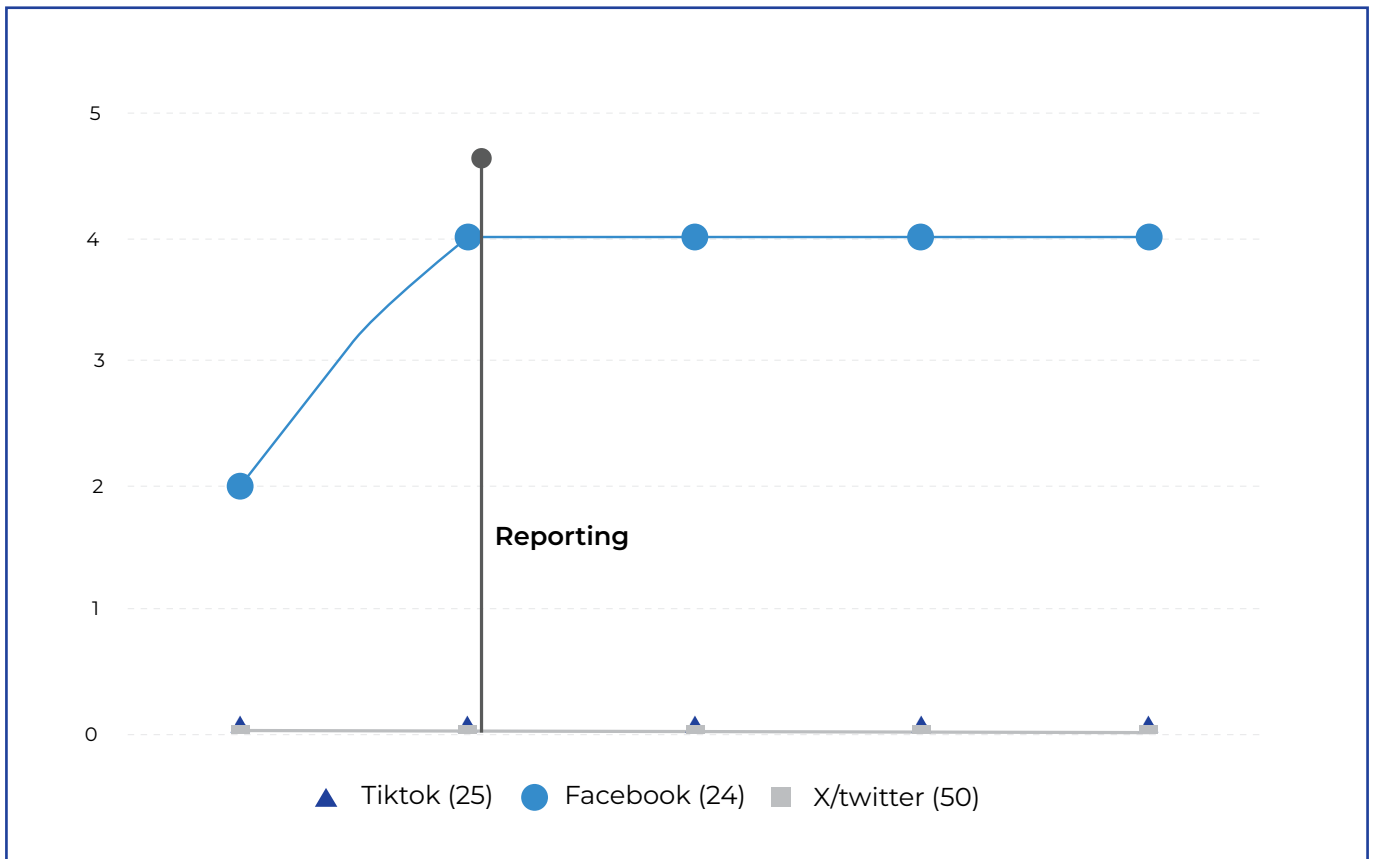


Figure 7: Content labelling. The number of posts labelled each week, with reporting dates indicated, on each platform

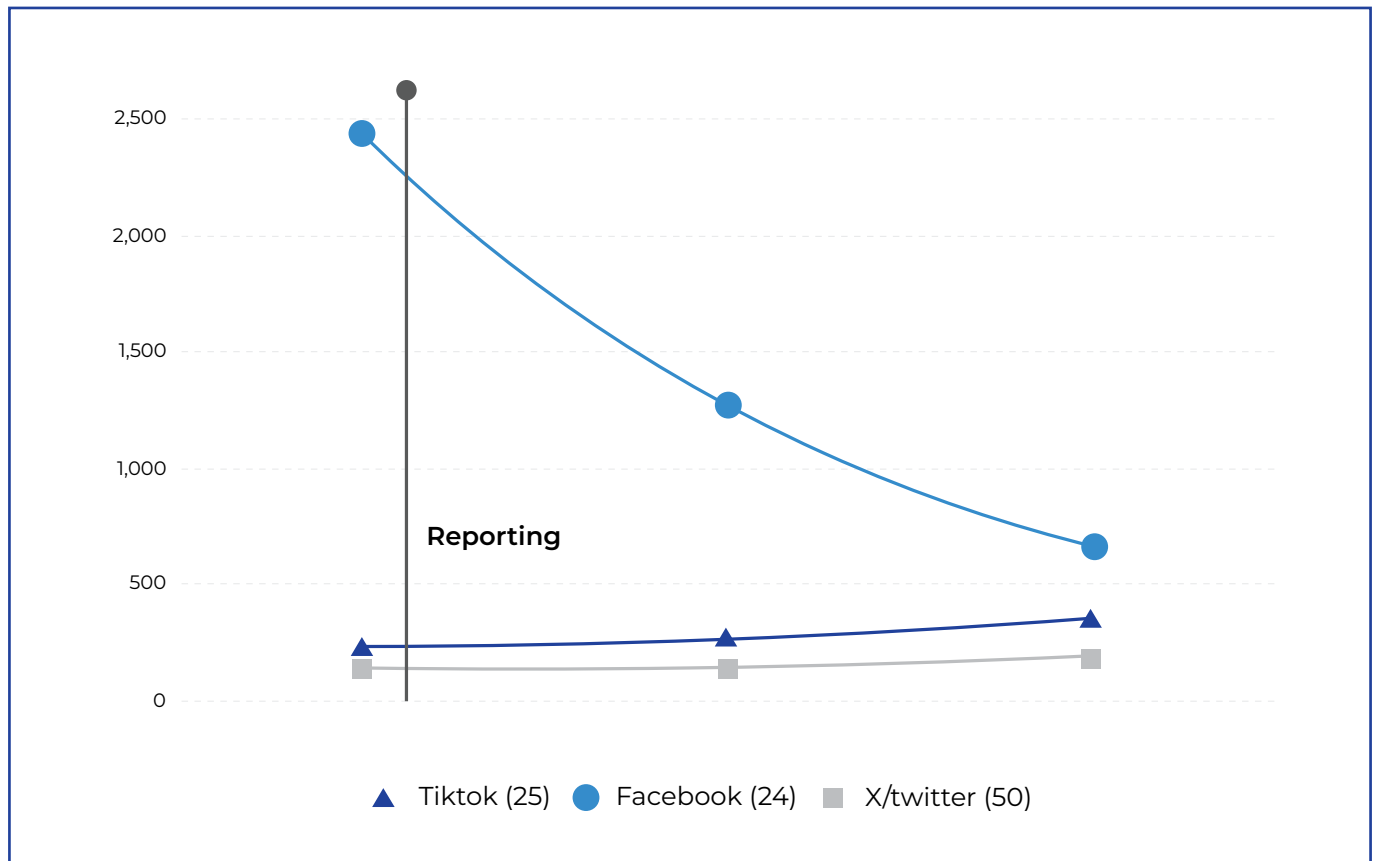
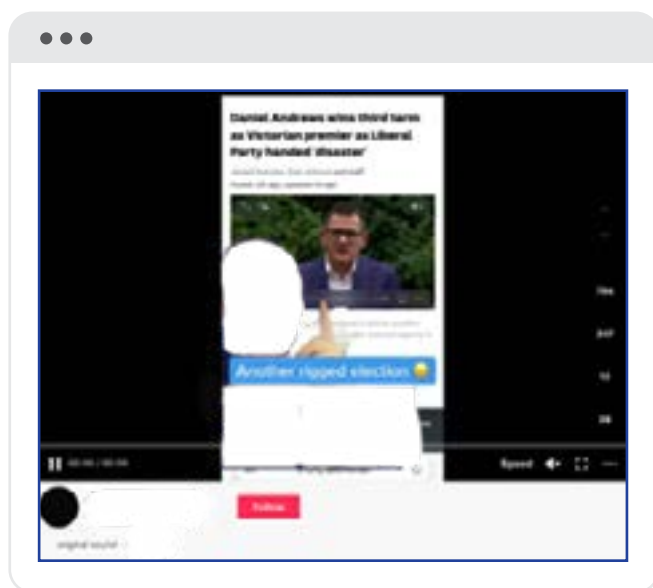


Figure 8: Content growth. The growth of the content each week, with reporting dates indicated, on each platform. Please note: all content continued to grow after reporting, albeit at a significantly less rapid rate on Facebook.

Content that became unavailable and labelled

The content that was removed does not appear to be substantively different to content that was left available (see *Figure 12a*), nor does the content that was labelled (see *Figure 12b*). Numerous other pieces of content that violate platforms' community guidelines remain available and unlabelled.

Figure 9: 'Rigged elections' content

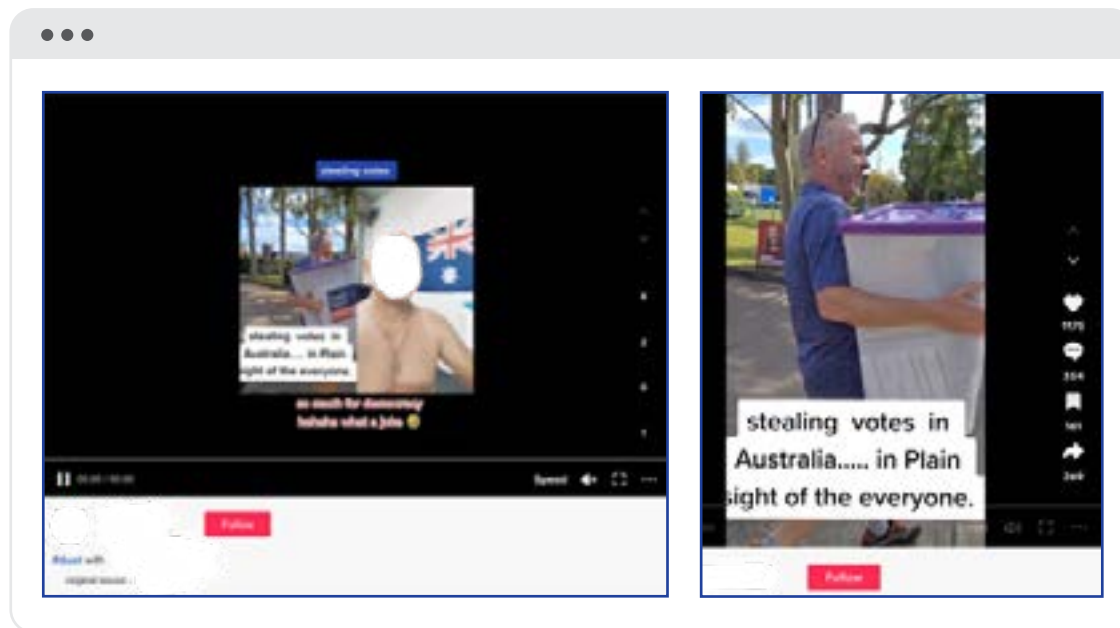


One piece of content that described the Victorian election results as rigged became unavailable before we reported it.

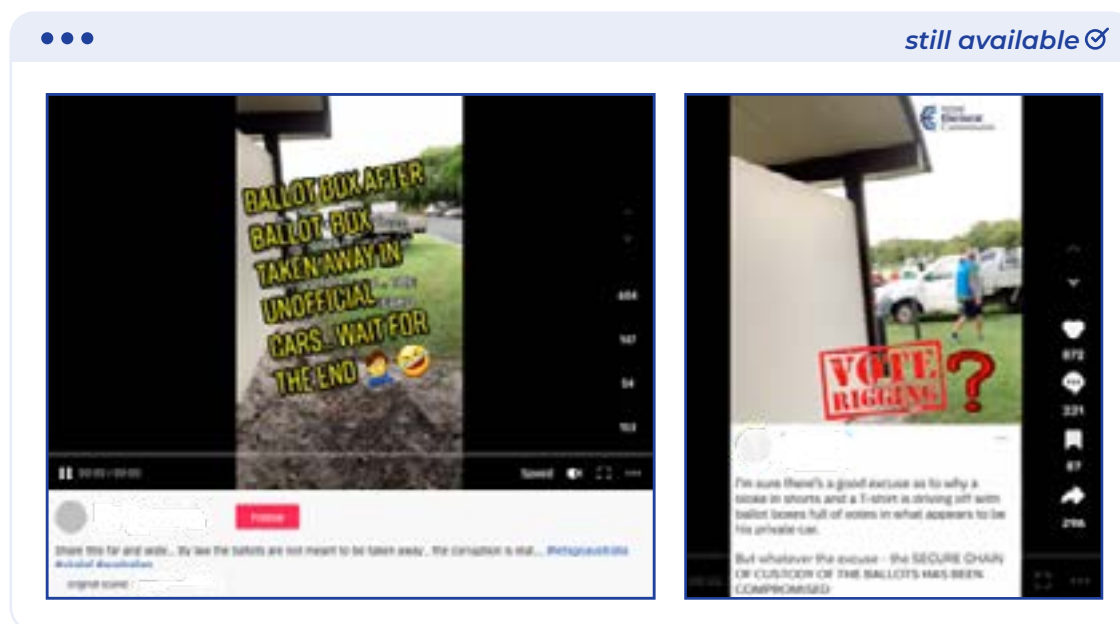


Content alleging the Victorian, NSW and Queensland elections were rigged.

Figure 10: 'Stolen ballots' content

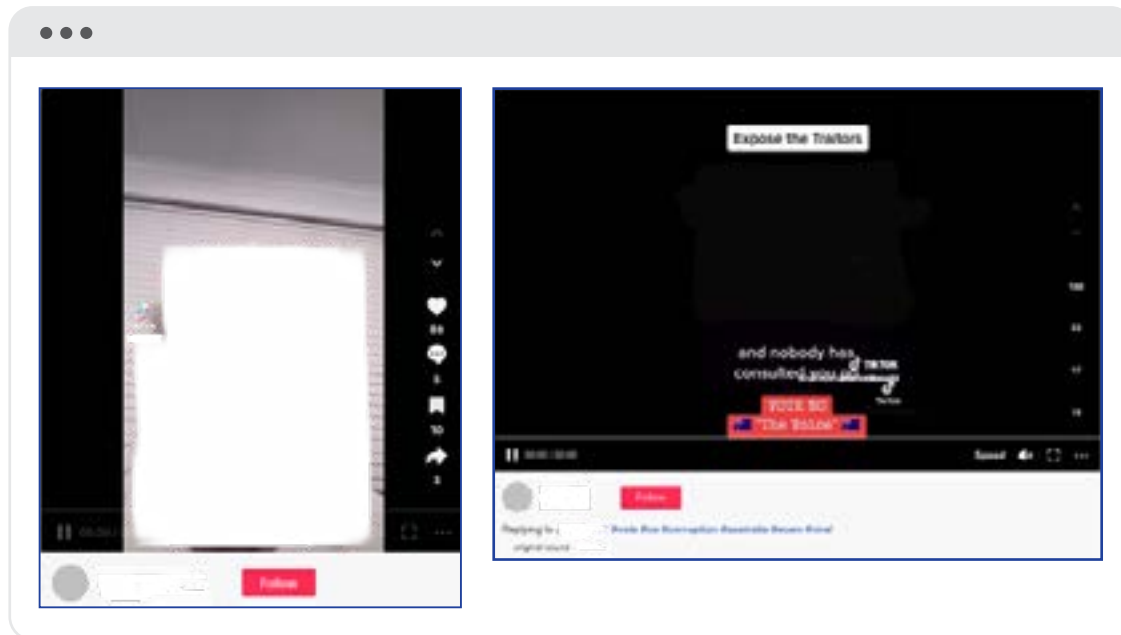


Two pieces of content that suggest that ballot boxes were stolen from Australian elections became unavailable (Note: one was a duet and the original source post was still available online.)

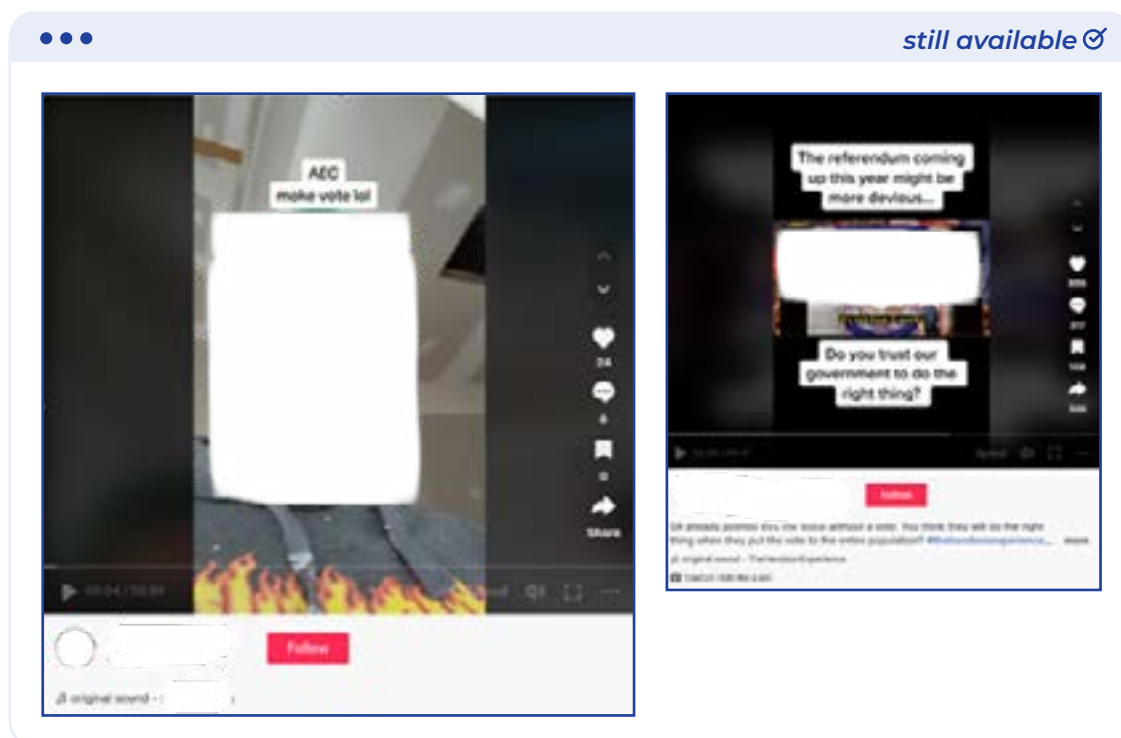


Other content that suggests that ballot boxes were stolen from Australian elections is still available.

Figure 11: Referendum rigging, corruption, or scams

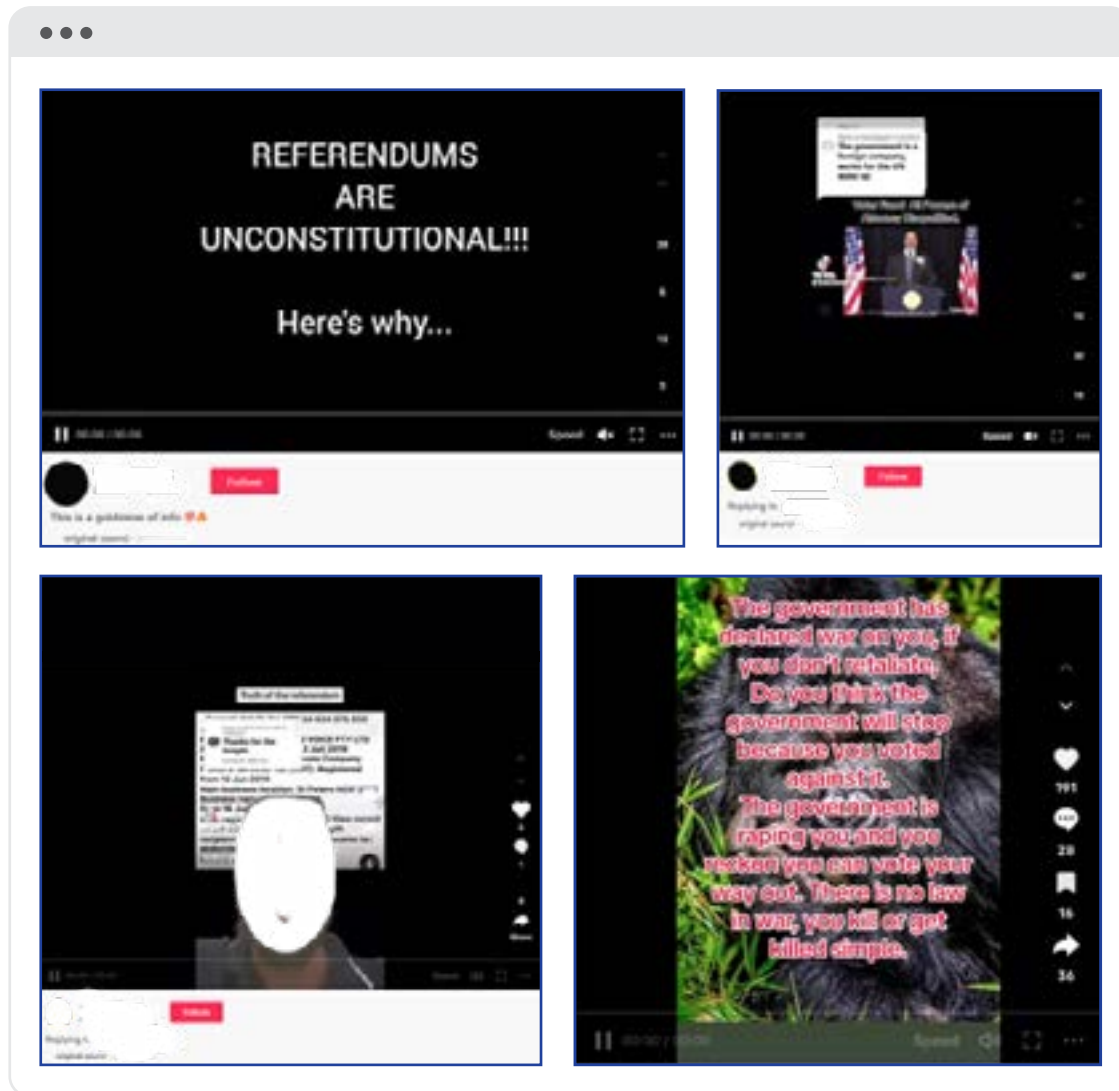


Two pieces of content **that described the Voice referendum in relation to corruption, scams or upcoming 'rigging'** became unavailable.



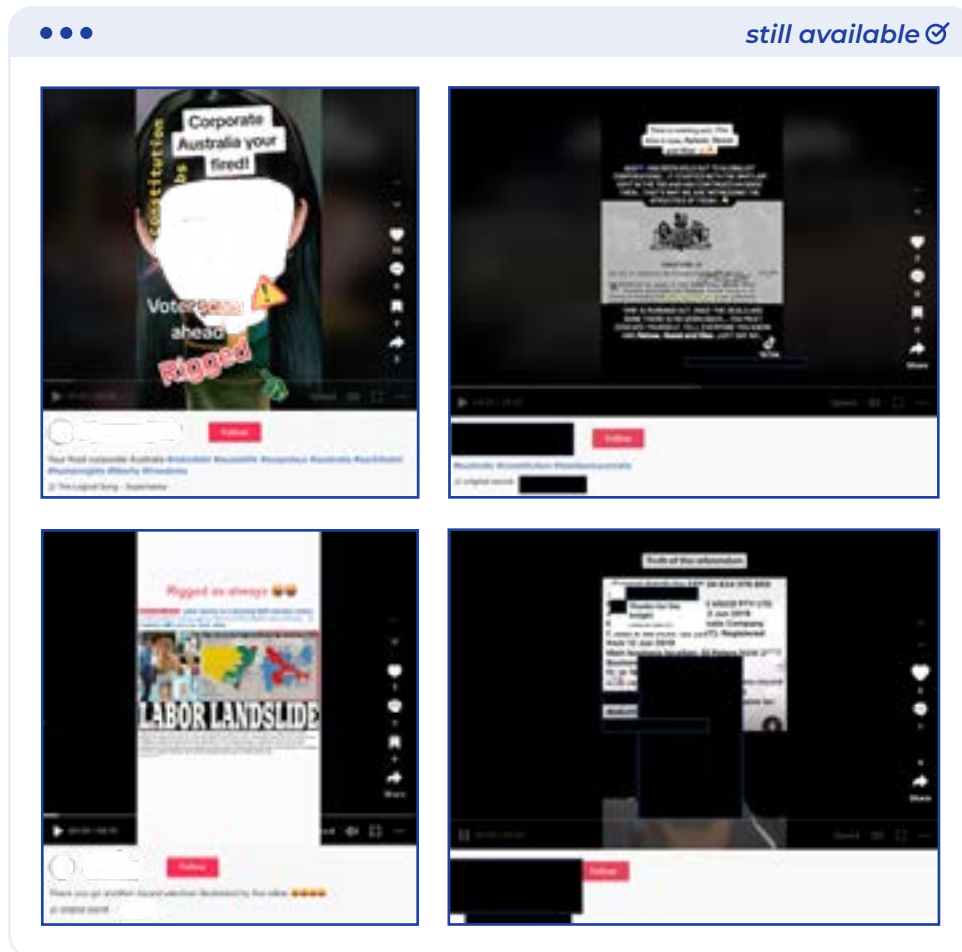
Other content that described the Voice referendum in relation to corruption, scams or upcoming 'rigging', for example, accusing the AEC of tricking indigenous people with a scam or corrupt mock vote. This includes some of the same actors as in the unavailable material.

Figure 12a: Electoral misinformation and conspiracies (Content taken down)



Four pieces of content that blended electoral process misinformation and disinformation with conspiracy theories became unavailable. One described the Voice referendum as unconstitutional because Australia's constitution was invalidated in the '70s and described voting as an unconstitutional act of treason; one described voting as irrelevant or illegal because the Government was out to get us no matter how or if we vote; and two others stated that the Australian government was a foreign company run by the UN, including US narratives about voter fraud as a way to prop up governments in the context of the referendum.

Figure 12b: Electoral misinformation and conspiracies (Content still available)



Content that blended electoral process misinformation and disinformation with conspiracy theories is still available, such as content that described the referendum as rigged by the UN, the 'company of Australia', 'the elites', or that it was illegal or invalid because of globalist forces. This group includes some of the same actors as in the unavailable material.

Figure 13: Examples of unlabelled and labelled content on Facebook

Below is content that was labelled on Facebook, along with a similar post that was not labelled. All other content remains available and unlabelled.



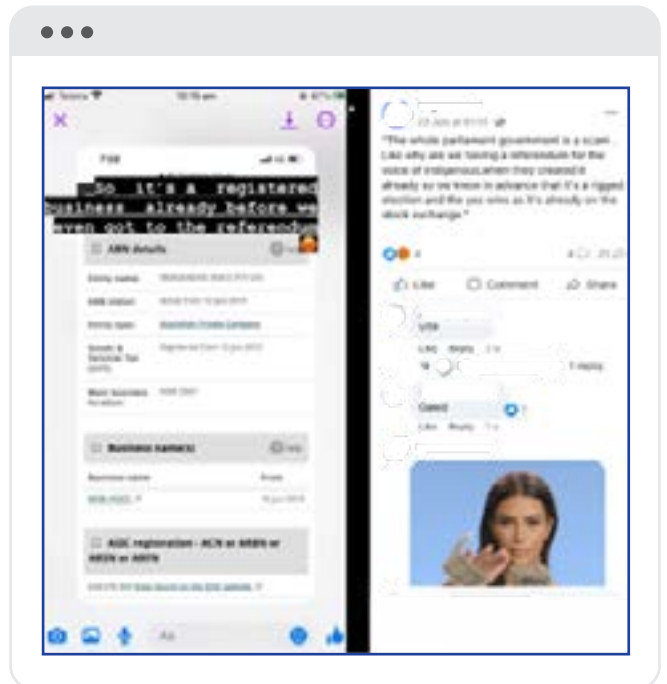
Content that was labelled on Facebook

One piece of content that was labelled inferred that the referendum could be rigged. It linked to an Andrew Bolt interview and asked **'Referendum votes can't be rigged can they??? Asking for a friend'**.



Content that was not labelled on Facebook

Other content that suggested the referendum is rigged remains unlabelled. This example describes the government as a scam and says **'... we know in advance that it's a rigged election'**



Conclusions

This was a small and rapid study, exploring 99 pieces of content, but it suggests there may be a significant problem. While platforms have published their guidelines regarding their definitions and management of misinformation and disinformation (as per their requirements under section 5.10 of the Code) and provided users with tools to report violative content (as per their requirements under section 5.11 of the Code), they fail to respond adequately to these user-reports.

This suggests that these provisions of the Code are having little meaningful impact.


The small data sample here presents limitations. This speaks to the need for greater oversight and access to data from the platforms themselves. To the best of our knowledge, this is the most accurate independent estimate of platforms' responses to user-reports of electoral misinformation and disinformation in Australia.


Platforms do not adequately respond to user-reports of electoral process misinformation and disinformation in the way that they claim to do so in their community guidelines, despite being 'aware of' this content.


Specifically;

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

It suggests:

 TikTok's content removal or labelling rate without reporting is at best²⁶ **4%** in a week


 Facebook's content removal or labelling rate without reporting is at best²⁷ **4%** in a week

 X's content removal or labelling rate without reporting is **0%** in a week.

Reporting electoral process misinformation appears to make little difference on Facebook and X when it comes to labelling or removing content, although it makes a moderate difference on TikTok.

After reporting, this research suggests that:

 TikTok's content removal or labelling rate after reporting is at best **32%** in a fortnight

 Facebook's content removal or labelling after reporting is **0%** in a fortnight

 X's content removal or labelling after reporting is **0%** in a fortnight

The findings also show that electoral process misinformation continues to grow in reach even after reporting, which suggests it is not adequately being de-amplified.

The rate of growth accelerates slowly after reporting on TikTok, but the rate of growth decelerates on Facebook.

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

²⁶ These are 'best case' estimations, as it is unclear if content that became unavailable at any stage of the research was taken down by users or the platforms themselves.

²⁷ These are 'best case' estimations, as other users may have reported these posts.

Appendix 1: Digi's Australian Code of Practice on Disinformation and Misinformation

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

Objective 1: Provide safeguards against Harms that may arise from Disinformation and Misinformation.²⁸

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 which 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. suspension or disabling of accounts of users which 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 algorithmic review of content and/or user accounts;
- H. the provision or use of technologies which assist digital platforms or their users to check authenticity or accuracy or to identify the provenance or 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
- M. providing users with tools that enable them to exclude their access to certain types of Digital Content.

²⁸ 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

[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) whose primary purpose is to disseminate information to the public and which use recommender systems, commit to :

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

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

Objective 7: (The final compulsory objective) Signatories publicise the measures they take to combat Disinformation and Misinformation²⁹

[Outcome 7: The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.](#)

All Signatories will make and publish a transparency report information (regarding their measures 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 under this Code.

Signatories may fulfil their commitment by providing additional reports and/or public updates on areas such as 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.

²⁹ 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

Appendix 2: Platforms' community guidelines, in more detail

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 the platform has committed to take.



TikTok's community guidelines state that it removes violative content from the platform that breaks their rules.³⁰ This includes misinformation content 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 FYF if it contains unverified claims about the outcome of an election."

Content on TikTok that is misleading regarding electoral processes, such as claims of rigged elections, stolen votes or AEC malpractice should fall into the category of civic and electoral process misinformation. According to its community guidelines, TikTok should remove this content.

³⁰ TikTok 2023 Civic and election integrity <https://www.tiktok.com/community-guidelines/en/integrity-authenticity/>



Facebook's community guidelines³¹ suggest that it removes misinformation and disinformation where:

- *It is likely to contribute to the risk of imminent physical harm, including risk of violence to people, harmful health mis information including vaccine misinformation or the promotion of miracle cures for example*
- *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.*



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 rigged elections, stolen votes or AEC malpractice, should fall into the final category of 'all other misinformation' where Facebook focuses on reducing its prevalence. This requires fact-checkers to have investigated content before triggering. The content included in this rapid experiment addresses narratives that have been fact-checked and determined to be false. According to its community guidelines, Facebook should 'reduce the prevalence' of this content.

³¹ Meta 2023 Community Standards: Misinformation <https://transparency.fb.com/en-gb/policies/community-standards/misinformation/>



X's community guidelines³² state that it removes or labels political mis and dis information content 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);
- 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).

³² X 2023 Civic integrity misleading information policy <https://help.twitter.com/en/rules-and-policies/election-integrity-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 rigged elections, stolen votes or AEC malpractice, should fall into the final category, of misleading information about outcomes, by undermining faith in the process itself. According to its community guidelines, X should label or remove this information.

As above, the claims made in the content covered by this research have been extensively fact checked. It is worth noting that in many of the X posts we monitored, the AEC had actively engaged and attempted to refute the contents of X posts.