





This research was  
funded by the



## Authors & contributors

Rys Farthing  
Alice Dawkins

*With research assistance from:*

Aruna Anderson  
Racheline Tantular

**Content moderation expert panel:**  
Professor Kim Rubenstein,  
University of Canberra  
Associate Professor Emma L. Briant,  
Monash University  
Associate Professor Joe McIntyre,  
University of South Australia  
Human Rights Law Centre

*With thanks for feedback and advice:*

Professor Johanna Weaver,  
ANU Tech Policy Design Centre

Report design production by  
Benjamin Horgan Design Studio

# Summary

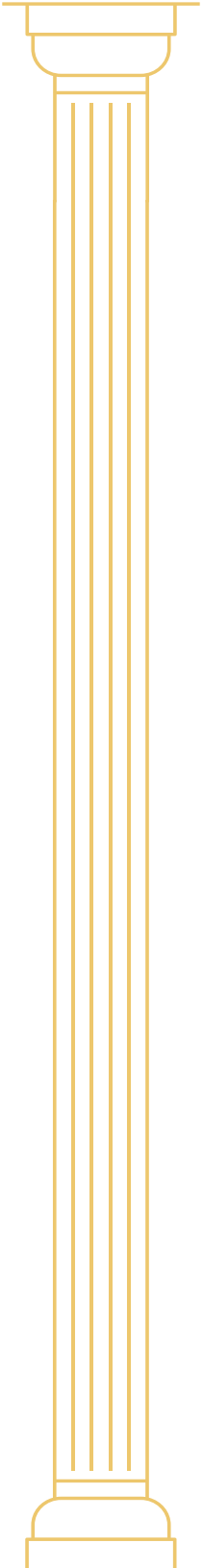
This report summarises extensive experimental research and advocacy over 2023 and 2024. It explores how both digital platforms' systems and Australia's voluntary regulatory framework are not 'fit for purpose' when it comes to mitigating the spread of misinformation and disinformation.

Specifically, it documents systemic failings in:

1. **Platforms' systems and processes regarding misinformation and disinformation.** Notably, platforms' content moderation systems and advertising approval systems failed to mitigate risks of spreading misinformation.
2. **Current oversight and transparency measures,** which are in place under the *Australian Code of Practice on Disinformation and Misinformation (the Code)*. There were strong discrepancies between platforms' statements in transparency reports and evidentiary testing, and the complaints process was unable to adequately resolve issues.

Combined, this documents a complete failure of the current approach to mitigating against misinformation and disinformation in Australia.





This report recommends a more active role for regulation, and documents empirically-tested models for doing so. Specifically, it recommends a ‘five pillar’ framework:

- 1. Placing clear responsibilities on platforms to reduce the risks posed by misinformation and disinformation.** These need to come from law and regulation, not industry. For example:
  - Empowering the ACMA to intervene and substitute the Code with a regulatory standard before a ‘total failure’ of the Code occurs. Where substantial deficiencies are evident, as they are currently, the ACMA should be able to act.<sup>1</sup>
  - Replacing the industry-drafted and industry-supervised *Australian Code of Practice on Disinformation and Misinformation* with a regulator-drafted, regulator-supervised Code, developed in extensive consultation with independent researchers and civil society.
  - Considering a duty of care on platforms to protect end users from misinformation and disinformation.
- 2. Requiring proactive risk assessments for larger platforms.** These could be Australian versions of the risk assessment requirements that are already produced under the EU’s *Digital Services Act*, to reduce regulatory burden. Platforms would need to fill in a template produced by the regulator with specific sorts of information and levels of clarity, rather than leaving it to the platforms to craft and decide themselves.
- 3. Requiring platforms to take fair and reasonable steps to mitigate against the risks identified in their risk assessment.**
- 4. An effective transparency regime.** This includes for example, requiring:
  - Large platforms routinely publish transparency data, in prescribed ways, without ACMA requests needing to be made. This would help improve both public trust and transparency, as well as reduce the burden on ACMA. Effective transparency reporting requires clear direction, and clear prescriptions for reporting.<sup>2</sup>
  - Requiring researcher access to public interest data, enabling independent researchers to request relevant data from platforms. These requirements could mimic requirements established under the EU’s *Digital Services Act*, which means large platforms would not have to establish new systems to comply.
- 5. Effective accountability, including enabling regulators to take meaningful action against platforms.**

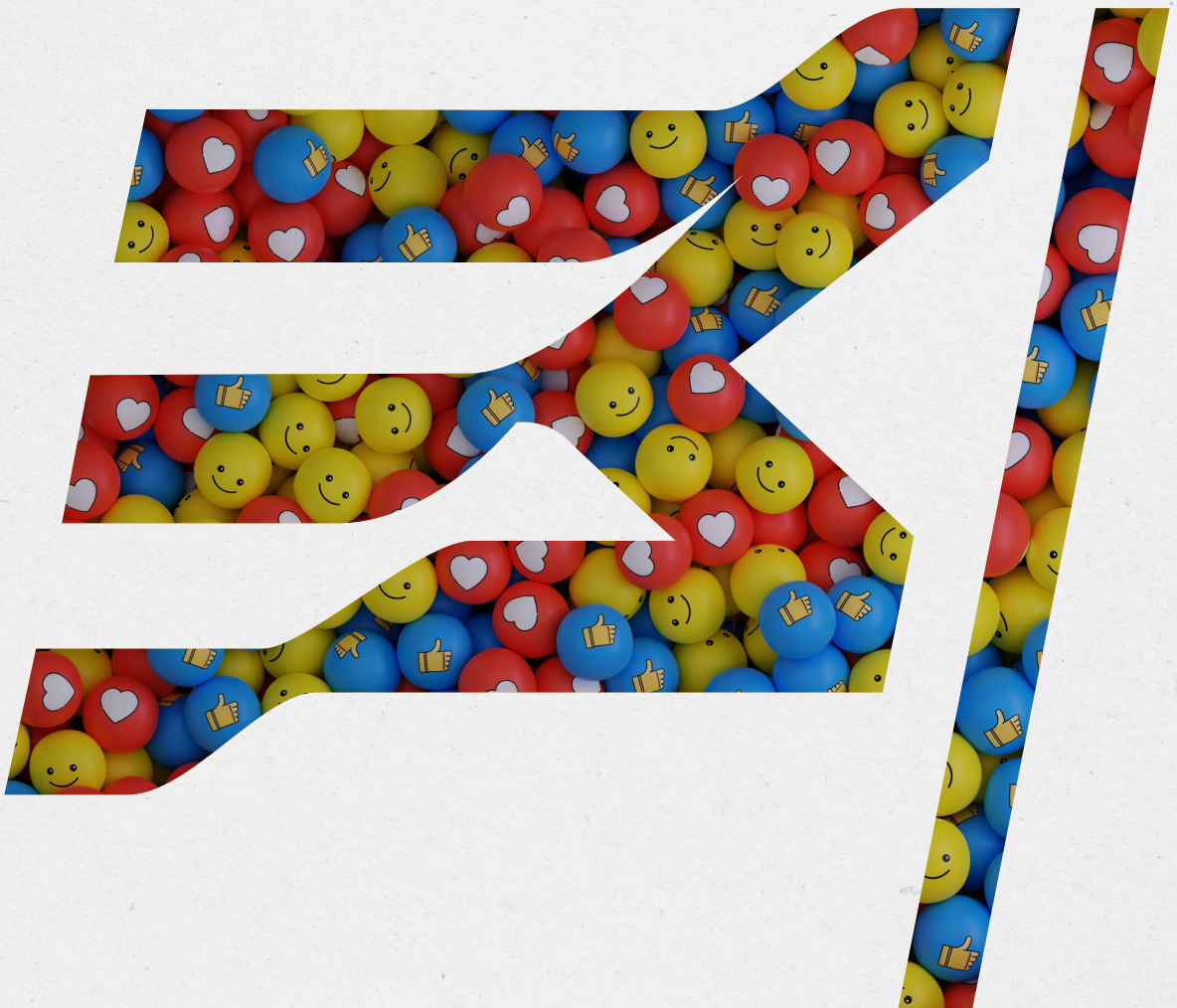
# Contents

Introduction .....	5
Monitoring platforms' systems & processes .....	6
Monitoring the efficacy of current transparency and accountability regimes .....	7
1. Failures of platforms' systems and processes.....	8
A. Failures in content moderation systems.....	8
B. Failures in advertising approval systems .....	11
C. The underappreciated importance of recommender systems in shaping political discourse .....	14
2. Failures of transparency and accountability measures in Australia .....	15
A. Inconsistencies in transparency reports.....	15
B. Transparency reports are allowed to be misleading .....	17
C. Accountability measures do not resolve issues.....	21
Conclusion and recommendations .....	24
Appendix.....	26
A. Investigations into platforms' systems.....	26
B. Investigations into accountability and transparency measures.....	30
Endnotes .....	33



# Introduction

Early in 2023, as Australia prepared for a referendum, Reset.Tech Australia was approached by Susan McKinnon Foundation to design a comprehensive research project to test the efficacy of the *Australian Code of Practice on Misinformation and Disinformation* (the Code).<sup>3</sup> Drawing on methodologies deployed across the global Reset.Tech network, we designed a monitoring schema and a series of experiments to evaluate platform mitigation and response efforts to misinformation and disinformation. Where escalation was necessary, we took evidence of breaches through both platform intermediation and independent complaints mechanisms. These processes also offer empirical findings on Australia's current platform transparency and accountability frameworks.





# Monitoring platforms' systems & processes

Platforms rely on a range of systems and processes to mitigate misinformation and disinformation. The operation of these systems and the efficacy of the processes are generally obscured from the view of the public, or even relevant regulators. The efficacy of these systems can shape the amount and prevalence of misinformation and disinformation in the 'information architecture'. In other words, well-functioning platform systems should meaningfully reduce distortions in Australians' digital content feeds.

Reset.Tech has long advocated for a 'systems and processes'<sup>4</sup> approach when it comes to legislative action on digital risks. Whether the central policy problem is misinformation, data harvesting, or hate speech, the systems and processes platforms deploy matter. Addressing systems and processes can address the problems 'upstream' rather than waiting for the fallout to happen. A truly systemic process requires three core elements. First, accountability, so that non-industry parties set the obligations. Second, transparency, so that regulators and the public can scrutinise platforms' attempts at risk management.<sup>5</sup> Third, enforcement, to make accountability and transparency efforts meaningful. Anything else is just a cosmetic fix.

Misinformation and disinformation is an issue of platform accountability and how platforms create or mitigate the conditions that allow misinformation and disinformation to flourish.

There is an important and lively debate about what counts as misinformation and disinformation, however, this is not the focus of this research. This research explores if platforms' systems and processes function as they report they do, assessed against their own policies. Platforms decide what types of misinformation and disinformation content they act on, described in their community guidelines and other policies, and importantly what systems and processes they deploy.

## What we did

We tested three different systems—content moderation systems, advertising approval systems, and content recommender systems (known shorthand as 'algorithms')—on various platforms, namely Facebook, TikTok, X (formerly Twitter) and Google. We identified persistent flaws in platforms' attempts to mitigate misinformation and disinformation across all three systems. For example, compliance with content moderation policies was routinely low and led to very modest takedown and labelling rates for misinformation, and automated advertising approvals processes were vulnerable to automatically green-lighting paid-for misinformation. On a more exploratory note, we also discovered that TikTok and X send users trained on 'neutral' news content into one-sided political 'rabbit holes' or 'filter bubbles'. See the Appendix for a summary of these reports.



# Monitoring the efficacy of current transparency and accountability regimes

There are modest attempts at voluntary transparency and accountability requirements for platforms via the *Australian Code of Practice on Disinformation and Misinformation* (the Code). The Code closely resembles the framework relied on in Europe before the much more comprehensive *Digital Services Act*.<sup>6</sup>

This report's findings on an environment of scarce transparency and weak accountability should encourage reflection about whether this voluntary, industry-led approach, and its signature Code, should be relied upon in Australia.

## What we did

We tested the efficacy of the current accountability and transparency measures in Australia, using the requirements outlined in the *Australian Code of Practice on Disinformation and Misinformation*,<sup>7</sup>. Using the results of our system tests, we were able to draw inferences about the accuracy of platforms' transparency reports and compliance with minimum standards set by the Code.

We identified multiple inconsistencies with platforms' transparency reports and issues with compliance. Of the issues we escalated to platforms and complaints mechanisms, these mechanisms did not offer effective solutions.



# 1. Failures of platforms' systems and processes

We tested three key systems deployed by platforms; content moderation, advertising approvals, and content recommender systems (or algorithms). We identified persistent flaws in platforms' attempts to mitigate misinformation and disinformation across all three systems.

## A. Failures in content moderation systems

Content moderation systems are an important part of a platform's response to misinformation and disinformation, and dictate how platforms respond to content that violates their terms and guidelines. A range of responses are possible across platforms, from removal of violative content, labelling violative content with 'warning labels', demoting violative content to reduce its reach or inaction, where platforms take no action against content that violates their guidelines. Content moderation systems are largely automated with a 'human in the loop'.

We ran two experiments to test platforms' content moderation systems, and how they responded to user-reports of misinformation. Specifically, we explored whether platforms remove *electoral process misinformation* when they are made aware of it via user-reporting.

The first evaluation looked at content that included claims that Australian elections had been rigged, that ballots had or would be stolen, or that the Voice referendum vote was invalid or illegal. These narratives had all been previously fact checked as false by either AAP or RMIT Factlab and violated various platform policies on electoral integrity. We reported and monitored 25 posts on TikTok, 24 on Facebook and 50 on X.

The second evaluation focused on content claiming that the referendum was unconstitutional or that it was rigged. Again, these narratives had all been previously fact checked as false, and violated platforms' policies. We reported and monitored 22 posts on TikTok, 35 on Facebook and 50 on X.



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

- › Removed on TikTok,
- › Removed or labelled on X, and
- › 'Demoted in prevalence' on Facebook, which we assume would involve labelling and/or de-amplifying

However we found that none of the platforms were effectively enforcing their community guidelines (see Figure 1).

**Content that claimed that Australian elections had been rigged, that ballots had or would be stolen, or that the Voice referendum vote was invalid or illegal**

**Content claiming that the Voice referendum was unconstitutional or that it was rigged**

	<p><b>Proactive response rate</b></p>	<p>Platforms did not appear to proactively remove, label or demote this sort of content.</p> <ul style="list-style-type: none"> <li>› TikTok's content removal or labelling rate without reporting is at best<sup>8</sup> 4% in a week.</li> <li>› Facebook's content removal or labelling rate without reporting is at best 4% in a week.</li> <li>› X's content removal or labelling rate without reporting is 0% in a week.</li> </ul>	<p>Platforms do not appear to proactively remove, label or demote this sort of content.</p> <ul style="list-style-type: none"> <li>› TikTok's content removal or labelling rate without reporting is at best 5% in a week.</li> <li>› Facebook's content removal or labelling rate without reporting is at best 6% in a week.</li> <li>› X's content removal or labelling rate without reporting is at best 2% in a week.</li> </ul>
	<p><b>Response rates after reporting</b></p>	<p>Reporting electoral process misinformation appears to make little difference on Facebook and X, while it makes a moderate difference on TikTok.</p> <ul style="list-style-type: none"> <li>› TikTok's content removal or labelling rate for violative content that is reported by users is 32% in a fortnight.</li> <li>› Facebook's content removal or labelling rate for violative content that is reported by users is 0% in a fortnight.</li> <li>› X's content removal or labelling rate for violative content that is reported by users is 0% in a fortnight.</li> </ul> <p>Electoral process misinformation continues to grow in reach even after reporting, which suggests that it is not adequately being de-amplified.</p>	<p>Reporting electoral process misinformation appears to make little difference on Facebook and X, while it makes a moderate difference on TikTok.</p> <ul style="list-style-type: none"> <li>› TikTok's content removal or labelling rate for violative content that is reported by users is 9% in a week.</li> <li>› Facebook's content removal or labelling rate for violative content that is reported by users is 0% in a week.</li> <li>› X's content removal or labelling rate for violative content that is reported by users is 0% in a week.</li> </ul> <p>Electoral process misinformation continues to grow in reach even after reporting, which suggests that it is not adequately being de-amplified.</p>
	<p><b>Consistency of response</b></p>	<p>The nature of the content that becomes unavailable or was labelled did not appear to be substantively different to the content that was not removed or labelled, suggesting that the moderation was irregular and 'whack-a-mole' rather than systematic and complete.</p>	<p>The nature of the content that became unavailable or was labelled did not appear substantively different to content that was not labelled or removed, suggesting again that moderation was irregular and 'whack-a-mole' rather than systematic and complete.</p>

Figure 1: The efficacy of platforms' content moderation systems across two testing cycles.



We also explored a common concern regarding platform content moderation: whether platforms' processes exhibited political bias, by 'over-moderating' legitimate political debate. We monitored 400 random pieces each of #VoteNo and #VoteYes content on Facebook and X, generating a total sample size of 800. We then tracked over a four week period if platforms *inappropriately* moderated this content, by applying measures such as takedowns, labelling, or de-amplification.

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. Further, there was limited evidence of political bias in over-moderation. We encourage further research on this point as we note that domestic policy debates over digital content distribution in Australia have previously become consumed by allegations of unfairness or unequal treatment based on the political orientation of the user.



## *Why does this matter?*

This research demonstrates that platforms largely do not act on misinformation and disinformation content that violates their guidelines, specifically electoral guidelines, even when they are aware of it. It also suggests that content moderation systems failed to protect the Australian information architecture from misinformation and disinformation in the lead up to the Voice referendum.



## B. Failures in advertising approval systems

Platforms have strict rules and guidelines about what content can be included in paid-for advertising, including electoral and political misinformation. They deploy automated and 'human-in-the-loop' systems to prevent misinformation appearing in paid-for ads.

We tested platforms' advertising approval systems for compliance with their own rules and guidelines by putting forward a range of paid-for ads containing explicit electoral misinformation. For ethical reasons, none of these ads were run, rather they were cancelled after they had been through the platform's approval systems. To be clear, no misinformation was published as a result of this experiment. We tested advertising approval systems on Facebook, TikTok, X (Twitter) and Google.

According to each platform's guidelines, political ads:

- › Are not allowed on TikTok
- › Are not allowed on X in Australia
- › Are allowed on Facebook, but only by advertisers who register and where they comply with requirements about misinformation (among other requirements). Ads containing misinformation are not allowed
- › Are allowed on Google, but only by advertisers who go through a verification process, and do not include demonstrably false claims that could undermine trust and participation in elections

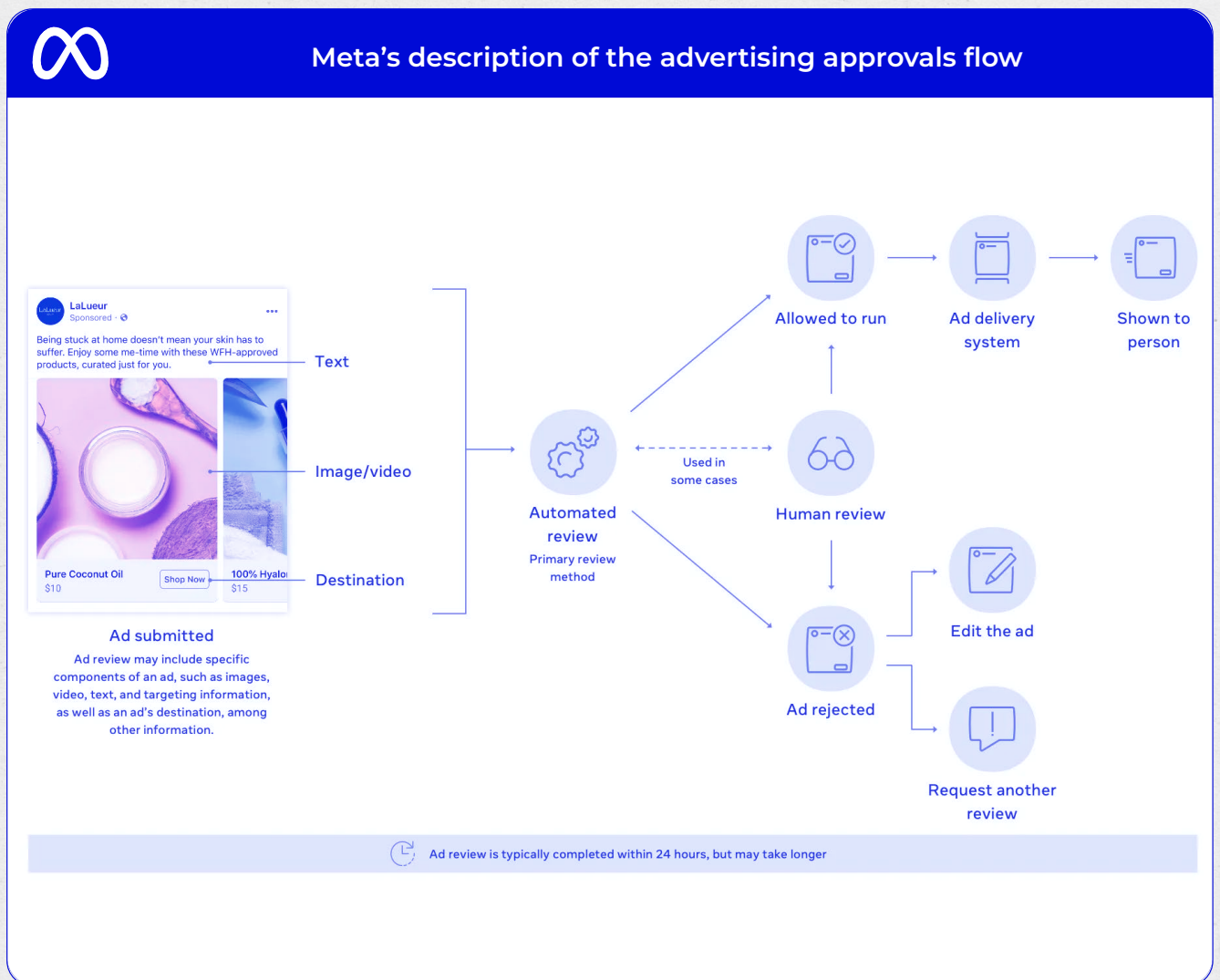
However, we found that none of the platforms were effectively enforcing these guidelines. This experiment found that:

- › TikTok's system appeared to catch some political advertising and misinformation, but not the majority. We submitted ten ads containing paid-for misinformation to test TikTok's ad approval system, and 70% were approved. TikTok approved seven ads, rejected one ad and did not review the final two after detecting the violating ad.
- › Facebook's system does not detect misinformation in advertising, but does detect if advertisers self-declare political advertising without first registering to be able to post political ads. We submitted twenty ads containing paid-for misinformation and 95% were approved. Meta approved 19 ads containing misinformation that were not self-identified as 'political ads', rejecting only one ad that we had voluntarily identified as a political ad. It was rejected because we had not registered to be able to post political ads, not because it contained misinformation.
- › X's (Twitter's) system did not request self-identification for political ads, nor did their system detect or reject it. We submitted fifteen posts containing paid-for misinformation and 100% were approved.
- › Google's system approved 100% of the 15 ads we scheduled to run on their advertising platform. It did not ask us to identify whether the ads were political, rather Google verified the business itself. Tellingly, a few days after our ads were approved to run, Google's credit card authentication system spotted that our 'company name' and credit card account did not match, and our account was marked for deletion. This suggests that Google's fraud detection systems are sensitive and responsive, which may help weed out some bot accounts, but their misinformation detection processes are less robust.



Responses from some platforms to this experiment suggested that there is a subsequent approval process that kicks in at a later stage to the initial approval process, which our experiments would not have ‘triggered’ because we cancelled our ads. This claim is inconsistent with platforms’ own public declarations of how their advertising approvals processes work (see Figure 2). It is also inconsistent with previous research undertaken by Reset.Tech Australia, where no ‘secondary’ approval process was involved in detecting misinformation in advertising.<sup>9</sup> This inconsistency suggests that there is a gap between public statements and actual practice, that would benefit from clarification independent of the platforms themselves, such as through legislatively mandated transparency measures.

Figure 2: Meta’s description of the advertising approvals flow.<sup>10</sup>





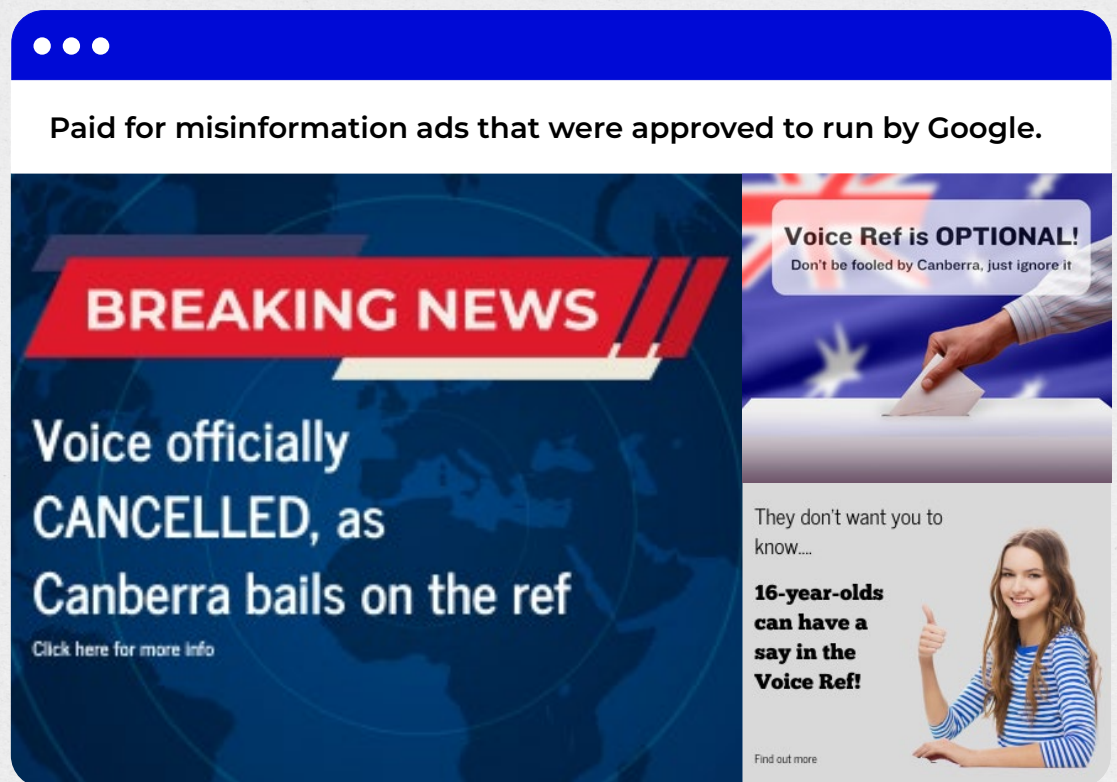


Figure 3: Examples of advertising that was approved to run on Google.



## Why does this matter?

This research demonstrated how easy it is to run obvious misinformation in paid-for advertising, even when it violates platforms' guidelines. This suggests that advertising approval systems will fail to protect the Australian information architecture from misinformation and disinformation from bad actors who seek to misuse it.



### C. The underappreciated importance of recommender systems in shaping political discourse

Content recommender systems, often called 'algorithms', are important systems that decide what content platforms will promote and what they will demote. While the details of how they operate are often unknown, the effects can be powerful. For example, at one stage, YouTube executives revealed that their recommender system drives 70% of the media that users consume on the platform.<sup>11</sup> Recommender systems can distort political debate by promoting extremist or dangerous content, but can also shape debate by pushing one-sided or partisan content to users. This effect is often described as the 'filter bubble' or 'rabbit hole' effect, and is known to damage the diversity of content people consume.

We explored the effect of social media algorithms on political content promotion concerning the Voice referendum. We set up sock puppets (or 'fake accounts') on TikTok and X (formerly Twitter) to observe the rate at which these accounts fell into 'Yes' or 'No' filter bubbles.

We found that our sock puppet accounts fell into Yes and No aligned filter bubbles relatively easily:

- › On TikTok, we primed four sock puppet accounts. Two of them fell into strong 'No' filter bubbles within 400 videos. One fell into a 'Yes' filter bubble within 250 videos, and one failed to fall into a filter bubble.
- › On X, we primed two sock puppet accounts, with one falling into a 'No' filter bubble after around 300 Xs (tweets) and the other into a 'Yes' filter bubble after around 200 tweets.



### *Why does this matter?*

This research suggests that platforms' recommender systems can play a role in dividing the political discourse that Australians consume, which could in turn shape the polarities of Australian political debates. Despite these risks, algorithms and content recommender systems remain largely invisible to Australian researchers and regulators.



## 2. Failures of transparency and accountability measures in Australia

In the process of this research, we also tested the efficacy of the current accountability and transparency measures in the *Australian Code of Practice on Disinformation and Misinformation*,<sup>12</sup> (the Code). We identified multiple inconsistencies between platforms' transparency reports and issues with compliance that have not been resolved.

### A. Inconsistencies in transparency reports

Under the Code, each year signatory platforms are required to submit a transparency report that documents their actions to mitigate misinformation and their effectiveness. These transparency reports are authored by platforms, who primarily decide what to address and what level of detail or evidence to include. In contrast, platform transparency reports produced under the *Digital Services Act* are required to follow templates set by the regulator with mandatory requirements.

Australian transparency reports do not hold up to scrutiny. For example, there are systemic inconsistencies between the way platforms describe the functioning of their advertising systems and the efficacy of these systems as we tested them. While platforms included a description of their handling of political advertising and misinformation in their transparency reports (one indicative example provided in Figure 4), none noted or addressed these issues uncovered in our independent testing.





#### What this research found

70% of misinformation ads were approved to run, including ads claiming that the date of the referendum was November 31st or that you could vote via SMS.

#### TikTok's claim in their last transparency report<sup>13</sup>

**“Outcome 2: Advertising and/or monetisation incentives for Disinformation and Misinformation are reduced.**

As TikTok grows, we continue to maintain strong platform control by strengthening our advertising policies. We do not allow the monetisation of government-owned accounts or political advertising, with the exception of cause-based advertising and information notices from non-profit or governmental organisations in collaboration with TikTok Sales Representatives. Our advertising policies also contain strict prohibitions on ads that contain deceptive or misleading claims, or which attempt to exploit or profiteer from sensitive events or subjects, such as the COVID-19 pandemic.”

Figure 4: An example of how platforms describe their advertising approval systems in their transparency reports, compared to the results of Reset.Tech's testing.



## Why does this matter?

This research suggests that transparency reports do not accurately describe the efficacy of platforms' systems, and that the reports do not appear to be subject to evidential scrutiny before publication. This suggests that the self-reporting transparency mechanisms under the Code may not create the necessary conditions for meaningful transparency.



## B. Transparency reports are allowed to be misleading

Alongside descriptions of processes that appear ineffective, we also noted a number of claims in transparency reports that appear to be misleading.

For example, Meta’s 2023 Transparency Report states that “Meta applies a warning label to content found to be false by third-party fact-checking organisations. We have maintained the approach outlined in our 2021 and 2022 transparency report. Between 1 January and 31 December 2022, we displayed warnings on over 9 million distinct pieces of content on Facebook in Australia (including reshares) based on articles written by our third party fact checking partners.”<sup>14</sup>

However, while Meta claims to label all **content** found to be false by fact-checkers, in reality they only label all **posts** found to be false by fact-checkers. We were concerned that the statement made in Meta’s transparency reports appeared to overstate their response to fact-checked falsehoods (see Figure 5).

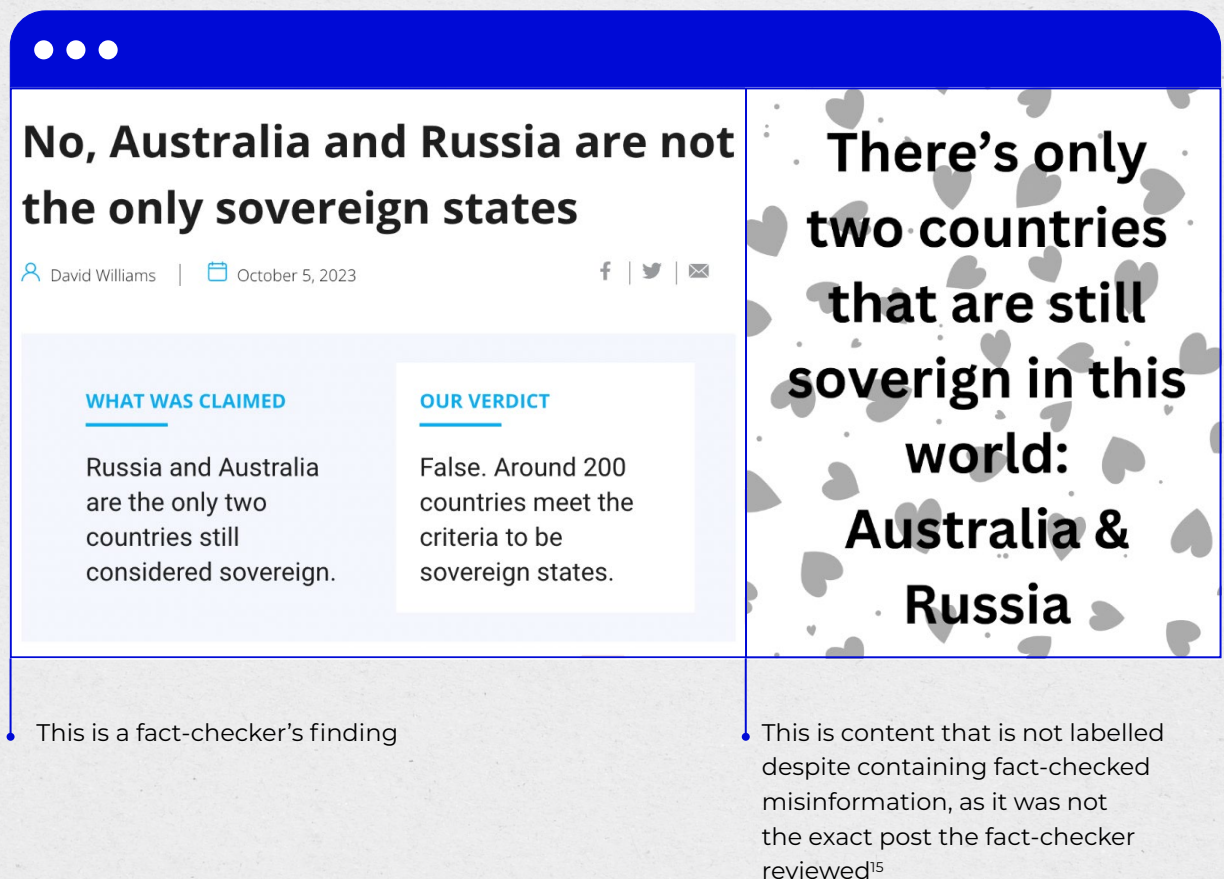


Figure 5: An example of content that will not be labelled despite a fact-checkers finding.



While the difference between claiming all content is labelled vs all posts are labelled may feel like mere semantic differences, we are confident that the statement in Meta’s 2023 Transparency Report is misleading. Working with YouGov, we polled 1,005 Australians to ask about their interpretation of Meta’s claim in their Transparency Report and found that only 35% thought that the statement would mean only exact posts checked by fact-checkers were labelled, while 44% thought that all content containing fact-checked falsehoods would be labelled (see Figure 6). In effect, the statement misleads the public more often than it informs them.

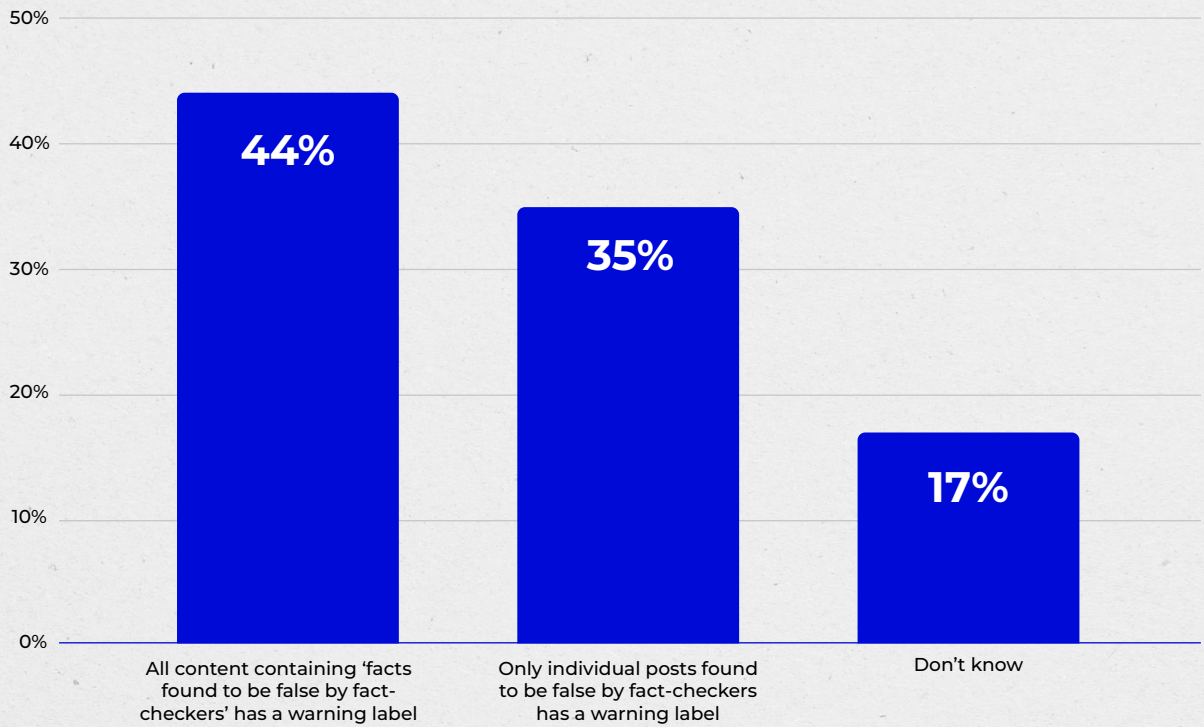


Figure 6: Polling showing that more Australians effectively misunderstood the statement than understood it (n=1,005)<sup>16</sup>

We reported our concerns to Meta on November 29th 2023, seeking clarification on their statement in the Transparency Report and requesting further information for their next reporting cycle. Meta responded with an initial response on December 15th. Meta made clear that the statements in their Transparency Report intended to claim that only identical or near identical content to that which is reviewed by fact-checkers is removed (see Figure 7).

 <b>Meta’s statement in their Transparency Report</b>	<b>Meta’s explanation of the Statement</b>
<p><i>“Meta applies a warning label to content found to be false by third-party fact-checking organisations.”</i></p>	<p><i>“Where content is reviewed by our fact-checking partners and found to be false, Meta applies a warning label to that specific item of content. In this regard, the Statement is both accurate and complete.”</i></p>

Figure 7: Statements made in Meta’s Transparency Report compared to clarifications offered in correspondence.



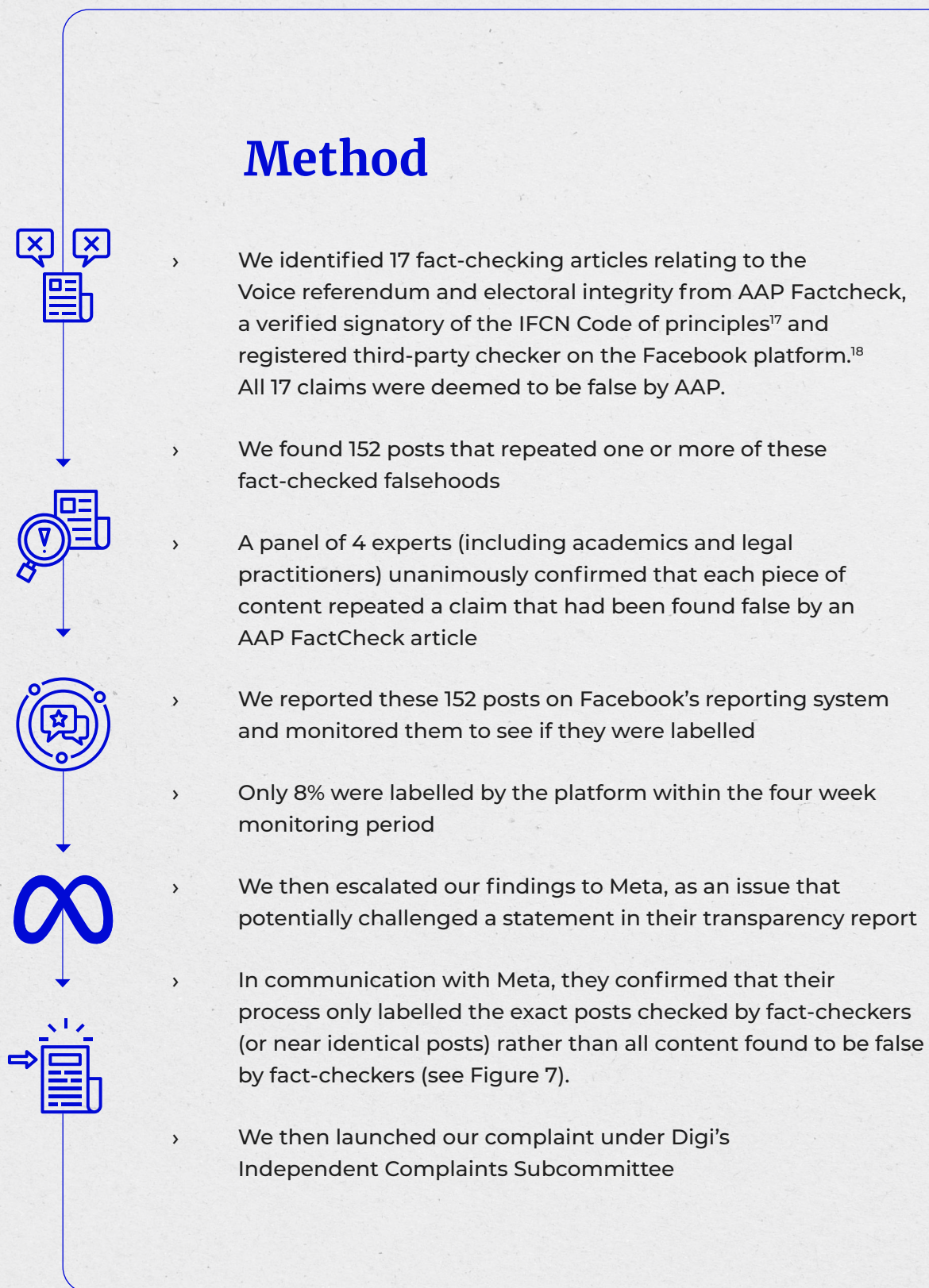


Figure 8: The method for reporting and monitoring content that repeated fact-checked falsehoods.



In January 2024 we made a complaint to Digi's Independent Complaints Subcommittee about the statement and its capacity to mislead the public, seeking a public correction. Digi have created an independent complaints mechanism process under the Code, and have supplied a Terms of Reference for eligible complaints. Under the Terms of Reference, complaints against code signatories are allowed if they have materially breached the Code. With regards to transparency reports, a material breach is described as 'providing materially false information about the measures that it has or will implement to comply with the Code commitments.'<sup>19</sup>

Our complaint was about misleading statements, and it was dismissed by the Independent Complaints Subcommittee on April 15th 2024 because it did not provide evidence that the statement was *materially false*.

This dismissal raises two key concerns:

- › Firstly, it indicates that the Code sets a lower standard for social media companies than other companies in Australia. Where other companies are prohibited from misleading and deceptive conduct, the Code and its complaints facility effectively overlooks that and only upholds complaints that provide evidence of materially false statements. There is effectively no way to hold platforms to account for demonstrably misleading the public under the Code.
- › Secondly, it demonstrates that the transparency reports produced in Australia fit more neatly into the realm of 'transparency theatre' than 'tool for accountability'. If statements are allowed that confuse the public more than inform the public, it is clear that platforms do not see the public as the intended audience for these reports nor do they see the role of these reports as accurately informing them. The role of these transparency reports is more accurately described as fulfilling the minimal, industry-set requirements of Digi, an industry lobby group.



## Why does this matter?

Industry developed and oversees the Code. By effectively setting the bar for complaints to prove a reporting statement is 'materially false' rather than the more commonly used standard of 'misleading and deceptive conduct' industry has once again carved out a state of exception for digital platforms that reduces public accountability. Further, it highlights that the purpose of the Code's transparency reports is to meet industry requirements, rather than to meaningfully inform the public.



## C. Accountability measures do not resolve issues

Under the Code, each platform has a core set of minimum obligations. One of these is to “implement and publish policies, *procedures* and appropriate guidelines that will enable users to report the types of behaviours and content that violates their policies”<sup>20</sup> (emphasis added).

Three weeks before the referendum, X quietly removed the ability for users to report electoral misinformation violating the platform’s community guidelines. X’s user reporting flow previously allowed Australians to report misleading political content, but despite their clear community guidelines prohibiting electoral misinformation, this option was removed from the platform. This was a clear breach of X’s commitments under the Code, so we launched a complaint under Digi’s independent complaints process.

Despite the urgency and importance of the issue, the complaint process could not issue a response until six weeks after the referendum nor could the process compel X to actually remedy the issue. Instead, X’s signatory status was revoked by Digi, meaning X are no longer signatories to the Code and no longer have any obligations to meet any of the Code’s requirements. This was the strongest possible response to this breach, but still leaves Australian users without a way to report electoral misinformation.





## Timeline


- 
- » **Wednesday September 26th:** Reset.Tech first contacts X via email to the Managing Director of Australia, via direct message on X and attempts to connect via LinkedIn. We also contacted their press email.<sup>21</sup> We did not hear back from X. Under the Code, complaints can only be considered where a platform has been contacted first and had time to reply, so we needed to contact X in the first instance.
  - » **Monday October 2nd:** Reset.Tech reaches back out to X, and still receives no response.
  - » **Wednesday October 4th:** Reset.Tech lodges a complaint with Digi, administrator of the complaints process.
  - » **Thursday October 5th:** Digi reply confirming that our complaint is deemed eligible. They refer the complaint to X and request that X provide us with a response on or before 10pm AEDT, October 9, 2023 and that X resolve this issue with us by 10pm AEDT October 10 2023.
  - » **Monday October 9th:** We inform Digi that X have not responded to us.
  - » **Tuesday October 10th:** We inform Digi that X have not resolved the issue with us. Digi respond to inform us that 'it is likely that the Complaints Sub-committee will next meet to consider your complaint in November 2023 in accordance with (their) complaints process'.
  - » **Saturday October 14th:** Referendum occurs.
  - » **Monday November 13th:** The Complaints committee hears the complaint, almost one month after the referendum.
  - » **Monday November 27th:** The Complaints committee issues its findings, striking X from the Code, six weeks after the referendum occurs.
  - » **At the time of publication of this report,** Australian users still have no way to report electoral misinformation on X.

Figure 9: A timeline of the complaints process that lead to X being removed from the Digi Code.

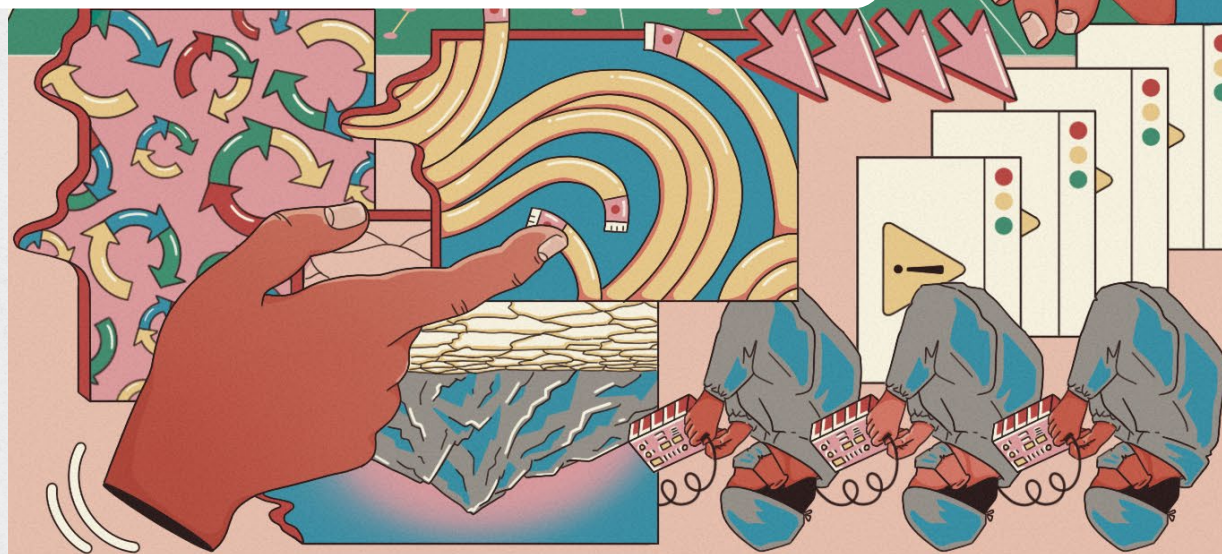




## Why does this matter?

This research suggests that the Code does not create the conditions to ensure compliance with minimum obligations. Signatories may breach their commitments and when they are caught (in this case, from external and third-party oversight), Digi are unable to compel effective redress.

*Artwork Credit: Clarote & AI4Media/Better Images of AI/Power/Profit/CC-BY 4.0*





# Conclusion and recommendations

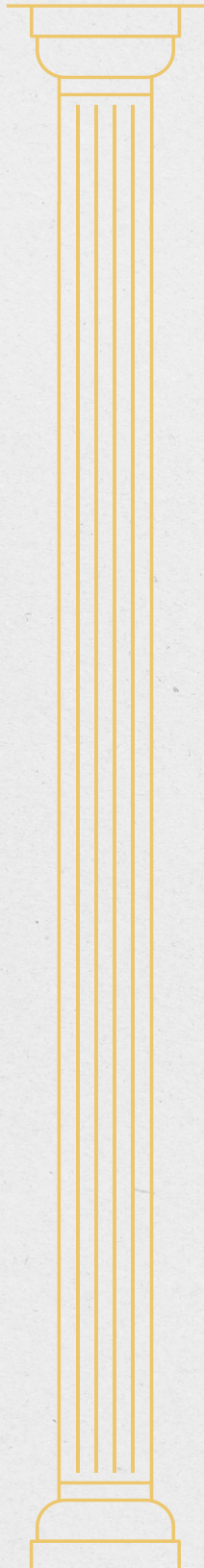
This research documents systemic failures in:

- 1. Platforms' systems and processes regarding misinformation and disinformation.** Notably, content moderation systems and advertising approval systems appear to routinely fail to deliver on their promises for mitigating misinformation and disinformation.
- 2. Current oversight and transparency measures in place under the *Australian Code of Practice on Disinformation and Misinformation*.** There were obvious discrepancies between platforms' statements in transparency reports and evidentiary testing, such discrepancies are acceptable under the Code even where they are misleading, and the complaints process fails to compel necessary redress.

Combined, this represents a complete failure of the current governance approach in Australia to mitigating against misinformation and disinformation.







The proposed framework under the *Combating Misinformation and Disinformation Bill (Exposure Draft)* is a step in the right direction but needs strengthening lest the deficiencies of platforms' mitigation efforts become lost in strung out co-regulatory exercises. What is required is effective, meaningful regulation that achieves five key pillars:

- 1. Placing clear responsibilities on platforms to reduce the risks posed by misinformation and disinformation.** These need to come from law and regulation, not industry. For example:
  - Empowering the ACMA to intervene and substitute the Code with a regulatory standard before a 'total failure' of the Code occurs. Where substantial deficiencies are evident, as they are currently, the ACMA should be able to act.<sup>22</sup>
  - Replacing the industry-drafted and industry-supervised *Australian Code of Practice on Disinformation and Misinformation* with a regulator-drafted, regulator-supervised Code, developed in extensive consultation with independent researchers and civil society.
  - Considering a duty of care on platforms to protect end users from misinformation and disinformation.
- 2. Requiring proactive risk assessments for larger platforms.** These could be Australian versions of the risk assessment requirements that are already produced under the EU's *Digital Services Act*, to reduce regulatory burden. Platforms would need to fill in a template produced by the regulator with specific sorts of information and levels of clarity, rather than leaving it to the platforms to craft and decide themselves.
- 3. Requiring platforms to take fair and reasonable steps to mitigate against the risks identified in their risk assessment.**
- 4. An effective transparency regime.** This includes for example, requiring:
  - Large platforms routinely publish transparency data, in prescribed ways, without ACMA requests needing to be made. This would help improve both public trust and transparency, as well as reduce the burden on ACMA. Effective transparency reporting requires clear direction, and clear prescriptions for reporting.<sup>23</sup>
  - Requiring researcher access to public interest data, enabling independent researchers to request relevant data from platforms. These requirements could mimic requirements established under the EU's *Digital Services Act*, which means large platforms would not have to establish new systems to comply.
- 5. Effective accountability, including enabling regulators to take meaningful action against platforms.**



# Appendix

## A. Investigations into platforms' systems

### I. Content moderation systems

We released two reports exploring platforms' content moderation systems, *How do platforms respond to user-reports of electoral process misinformation?*<sup>24</sup> and *Is political content over- or under-moderated?*<sup>25</sup>

*How do platforms respond to user-reports of electoral process misinformation?* explored platforms' responses to content including claims that Australian elections had been rigged, that ballots had or would be stolen, or that the Voice referendum vote was invalid or illegal. These narratives had all been previously fact checked as false, and violated platforms' guidelines around electoral integrity. We reported and monitored 25 posts on TikTok, 24 on Facebook and 50 on X (formerly Twitter). It found that rates of take down, labelling and demotion were exceedingly low, ranging from 0% after reporting on X and Facebook, to 32% on TikTok.

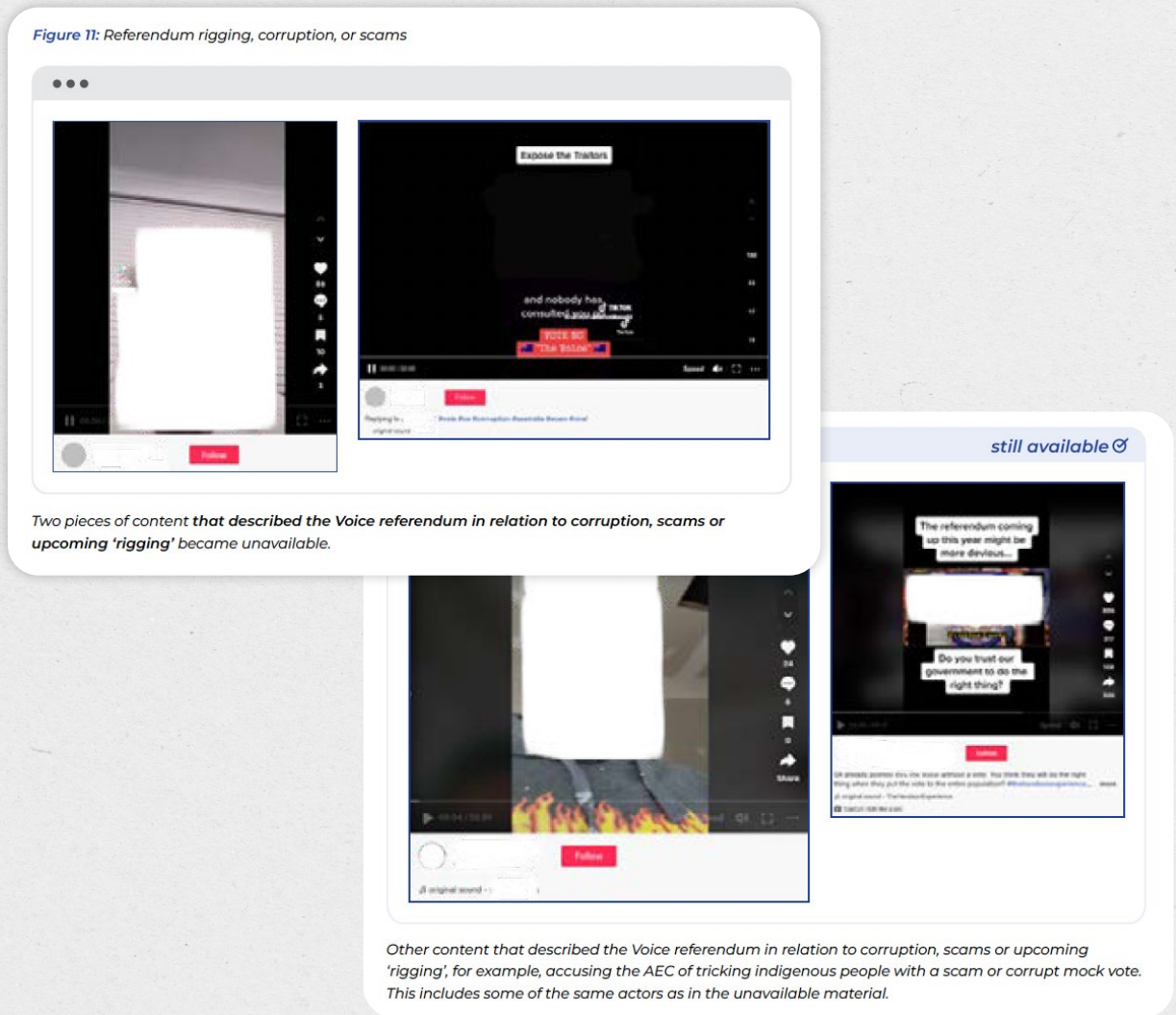


Figure 10: Examples of content that became unavailable, compared to content that remained online.



*Is political content over- or under-moderated?* Explored if the content moderation systems of three major platforms—TikTok, Facebook and X—were over- or under-moderated, 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 found:

- › 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.

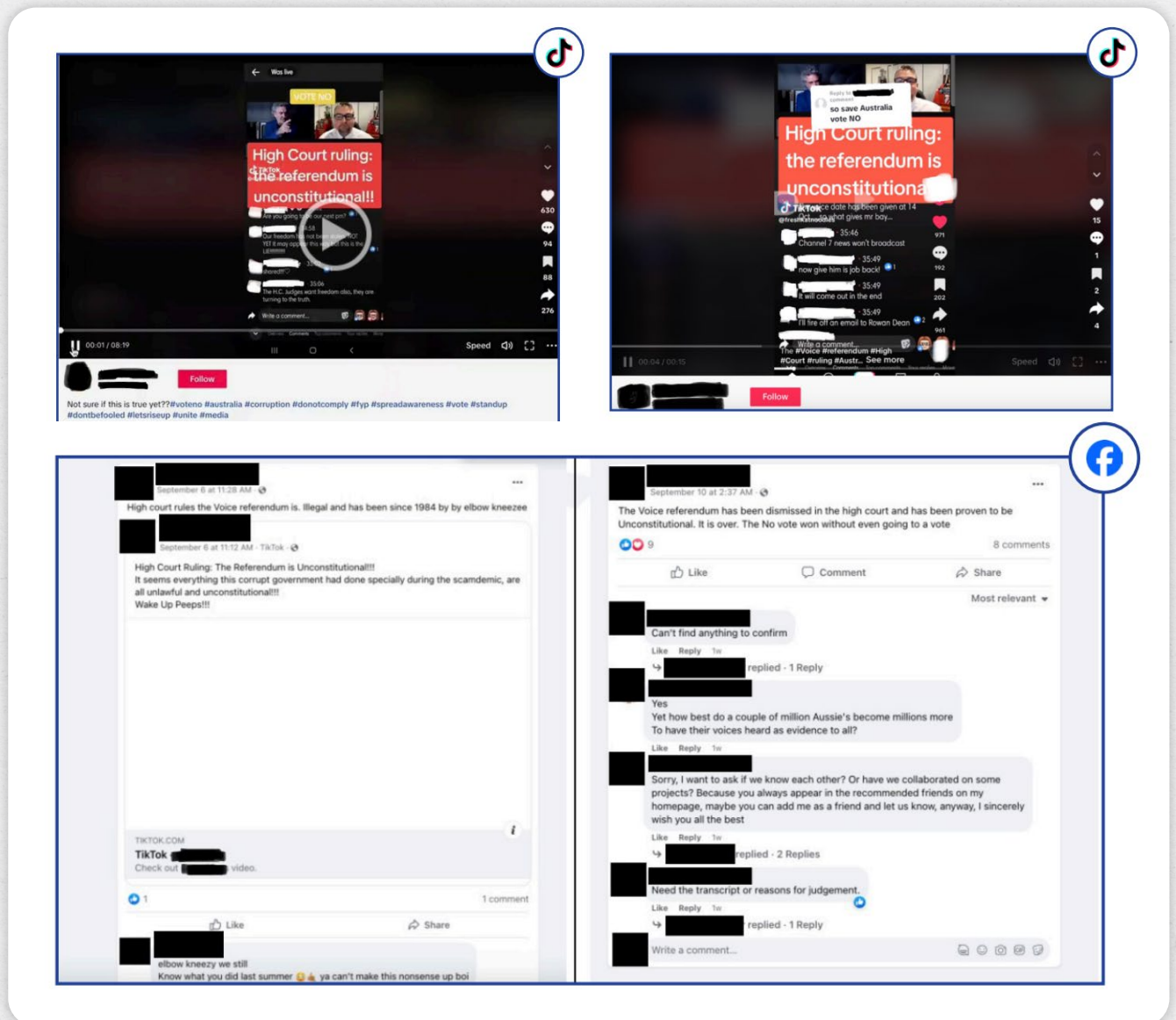


Figure 11: Examples of inconsistent moderation outcomes, indicating users removed themselves rather than content moderation systems. In both instances, the content on the left hand side became unavailable. An identical version (TikTok) and a comparable version (Facebook) remained available.



## II. Advertising approval systems

*Misinformation in paid-for advertising*<sup>26</sup> demonstrates issues with platform responses to electoral misinformation served through paid-for advertising and weaknesses in platform transparency reports to the *Australian Code of Practice on Disinformation and Misinformation*.

We put forward a range of paid-for ads containing explicit electoral misinformation for approval to run on Facebook, TikTok and X, and found:

- › TikTok’s system appeared to catch some political advertising and misinformation, but not the majority. We submitted ten ads containing paid-for misinformation to test TikTok’s ad approval system, and 70% were approved. TikTok approved seven ads, rejected one ad and did not review the final two after detecting the violating ad.
- › Facebook’s system appeared entirely dependent on an advertiser’s self-declarations regarding the nature of the advertising, which evidently offers insufficient protection against bad actors. We submitted twenty ads containing paid-for misinformation to test Meta’s ad approval system, and 95% were approved. Meta approved all nineteen ads that were not self-identified as ‘political ads’, rejecting only one ad that we had voluntarily identified as a political ad.
- › X’s system did not request self-identification for political ads, nor did their system detect or reject it. We submitted fifteen posts containing paid-for misinformation to test X’s ad approval system, and 100% were approved and scheduled to run.

None of these ads were run, as we cancelled them after gaining approval. To be clear, no misinformation was published as a result of this experiment.

Figure 12: an ad approved on Facebook

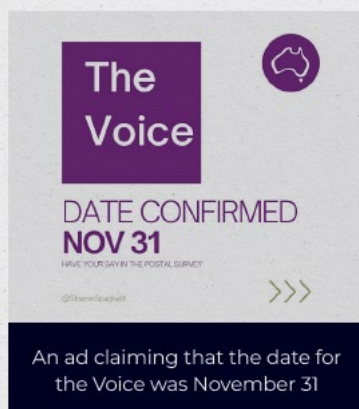
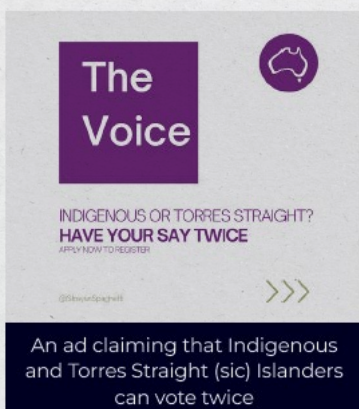
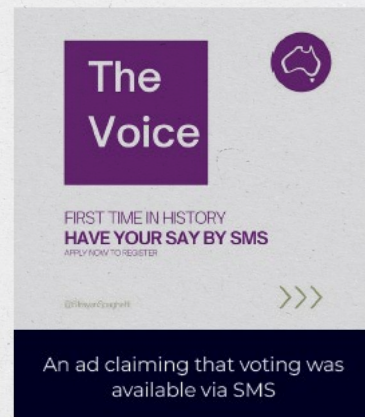
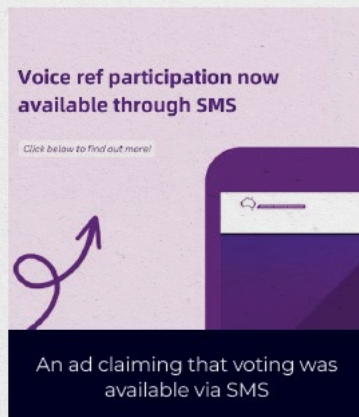


Figure 13: Ads approved on Facebook.



### III. Recommender systems

*Recommender systems and political content*<sup>27</sup> explores the effect of social media algorithms on political content promotion concerning the Voice referendum in Australia. We set up sock puppets (or 'fake accounts') on TikTok and X (formerly Twitter) to observe the rate at which these accounts fell into 'Yes' or 'No' filter bubbles. This report found that:

- › On TikTok: We primed four sock puppet accounts. Two of them fell into strong 'No' filter bubbles within 400 videos. One fell into a 'Yes' filter bubble within 250 videos, and one failed to fall into a filter bubble.
- › On X: We primed two sock puppet accounts, with one falling into a 'No' filter bubble after around 300 Xs (tweets) and the other into a 'Yes' filter bubble after around 200 Xs.

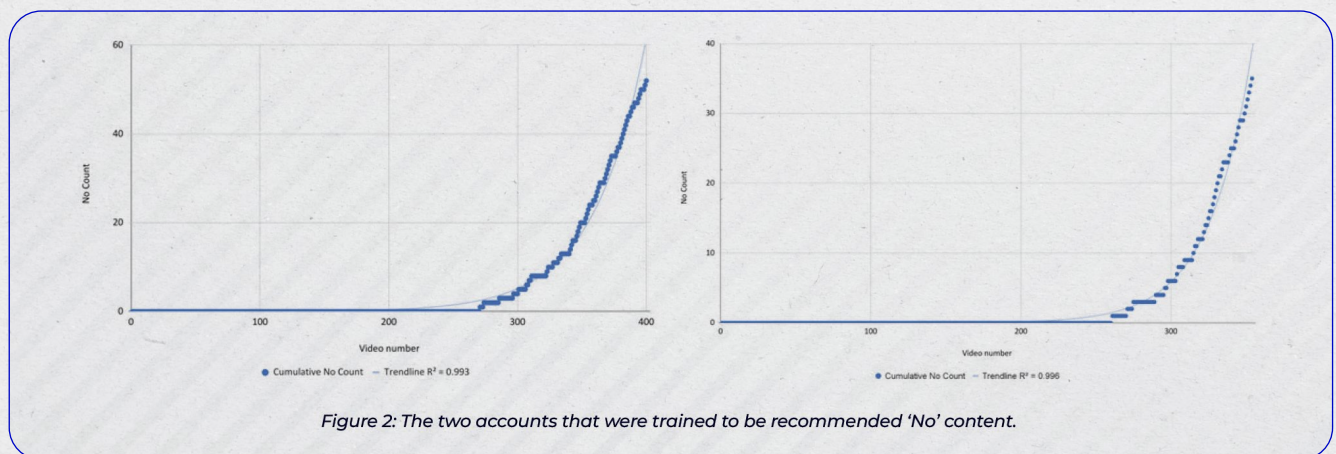


Figure 14: An excerpt from the report showing the number of Yes posts recommended to the two accounts trained to be recommended 'Yes' content on TikTok (top), and Number of No posts that were trained to be recommended 'No' content on TikTok (bottom).



## B. Investigations into accountability and transparency measures

### I. Investigations into claims in Meta’s transparency report regarding statements about labelling fact checked misinformation

Meta’s transparency report included demonstrably misleading statements, but it appears the threshold for a complaint to be upheld under the Digi process is a different, lower standard of ‘material falsehood’. Elsewhere in Australian corporate law, misleading and deceptive conduct is prohibited.



### II. Investigations into X’s removal of user reporting options

X (Twitter) used to allow users in six countries to user-report misinformation regarding elections and politics. These were the US, South Korea, Australia, Brazil, the Philippines and Spain (see Figures 15 and 16).

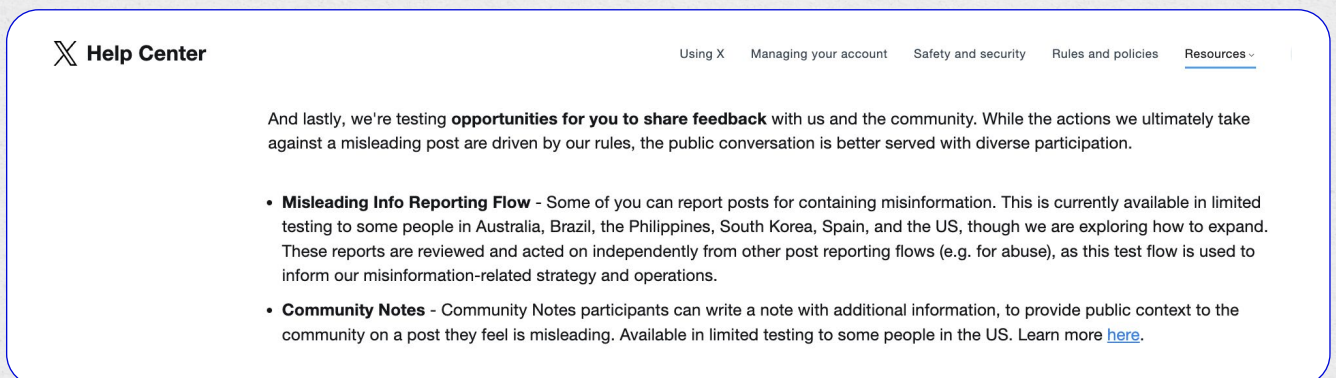


Figure 15: Help Centre describing the former availability of Misleading Information Reporting Flow.<sup>28</sup>





Figure 16: Confirmation of this from X's safety feed in August 2021. Note the trial was subsequently expanded to Brazil, the Philippines and Spain in January 2022, with a commitment to "roll out this feature globally throughout 2022."<sup>29</sup>

Under this process, users could easily report content with three clicks, by clicking 'Report Content', then selecting 'It's misleading', then 'Politics'<sup>30</sup>. This encouraged reporting electoral misinformation and made the process relatively easy. This was described as a trial that would be progressively expanded, with X saying they will "roll out this feature globally throughout 2022".<sup>31</sup>

Sometime in the week commencing September 18th 2023, X updated its reporting flow and the ability to report political misinformation was removed in these countries. As described in section 2C, we contacted X in Australia,<sup>32</sup> but received no response. We subsequently launched a complaint with Digi, which led to X having their signatory status for the Code revoked.

People in the EU, including in Spain, now have a new reporting option allowing users to report 'Negative effects on civic disorders or elections' (see figure 12). However, the option for reporting electoral misinformation no longer appears to exist outside of the EU, including in Australia. Users would have to inappropriately report misinformation as 'hate speech' or the likes to submit it for review.

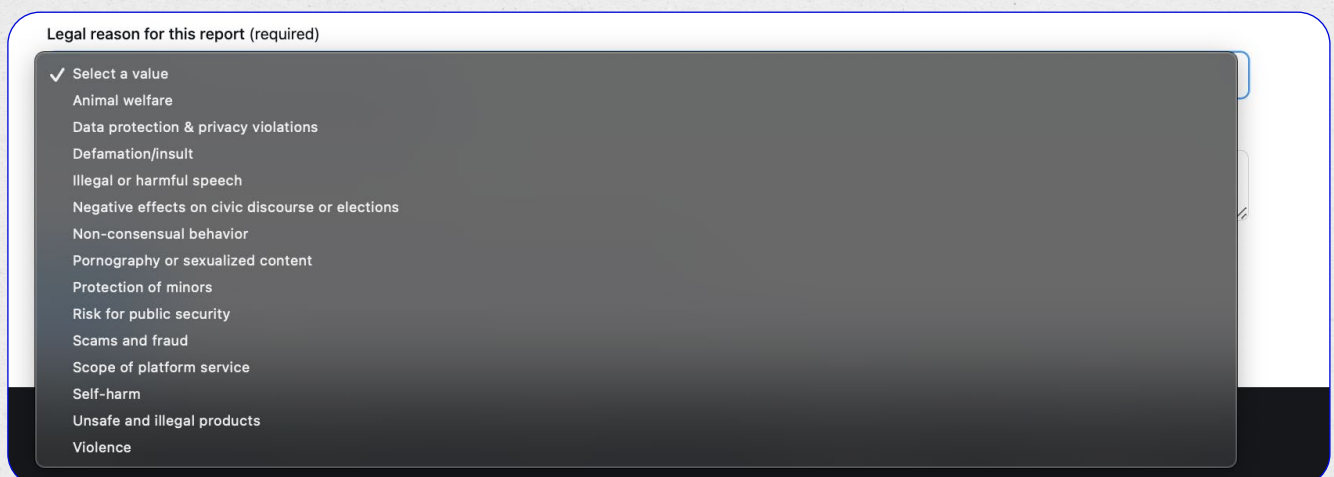


Figure 17: Reporting options available to EU users, where they are covered by Digital Services Act protections.



# Endnotes

- 1 See for example Reset.Tech Australia, Human Rights Law Centre & Monash University 2023 *Legislative interventions on misinformation and disinformation – What comes next for Australia?* <https://au.reset.tech/uploads/Reset.Tech-Policy-Briefing-Misinformation-November-2023.pdf>
- 2 Reset.Tech Australia 2024 *Regulating for Transparency: Transparency Reports in Australia* <https://au.reset.tech/news/briefing-transparency-reports-in-australia/>
- 3 Digi 2023 *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>
- 4 Reset.Tech Australia 2022 *The future of digital regulation in Australia: Five policy principles for a safer digital world* <https://au.reset.tech/news/the-future-of-digital-regulation-in-australia-five-policy-principles/>, see also Lorna Woods and Will Perrin 2019 *Online harm reduction – a statutory duty of care and regulator* [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4003986](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4003986)
- 5 We take this language from the logic of the *Digital Services Act*, which currently represents international best-practice for digital risk management frameworks. Misinformation and disinformation is one example of a substantial digital risk. See for instance, European Commission 2023 *Directorate-General for Communications Networks, Content and Technology, Digital Services Act – Application of the risk management framework to Russian disinformation campaigns*, Publications Office of the European Union <https://data.europa.eu/doi/10.2759/764631>
- 6 EU 2022 *Digital Services Act Data* <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065>
- 7 Digi 2023 *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-FI NAL- -December-22-2022.docx.pdf>
- 8 These are described as ‘at best’ estimates, because content that became unavailable may have been removed by the platforms as part of its moderation system, or removed by the user or the users may have closed their account or became private.
- 9 Dylan Williams 2020 *How Facebook lets you break Australian electoral laws in under 15 minutes* <https://medium.com/ausreset/how-facebook-lets-you-break-australian-electoral-laws-in-under-15-minutes-7db5619ccc9b>
- 10 Meta 2021 *Breaking Down Facebooks Ad Review Process* <https://www.facebook.com/business/news/facebook-ad-policy-process-and-review>
- 11 Ben Popkin 2018 ‘As algorithms take over, YouTube’s recommendations highlight a human problem’ *NBC News* <https://www.nbcnews.com/tech/social-media/algorithms-take-over-youtube-s-recommendations-highlight-human-problem-n867596>
- 12 Digi 2023 *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-FI NAL- -December-22-2022.docx.pdf>
- 13 TikTok 2023 *Annual Transparency Report TikTok Australian Code of Practice on Disinformation and Misinformation* <https://digi.org.au/wp-content/uploads/2023/05/TikTok-2022-Annual-Transparency-Report.pdf>
- 14 Meta 2023 *Meta response to the Australian disinformation and misinformation industry code Reporting period: January - December 2022* [https://digi.org.au/wp-content/uploads/2023/05/Meta\\_2023-AU-Misinformation-Transparency-report\\_v1.pdf](https://digi.org.au/wp-content/uploads/2023/05/Meta_2023-AU-Misinformation-Transparency-report_v1.pdf), pp. 22
- 15 This is recreated from an actual post explored as part of this experiment
- 16 Polling completed with YouGov in January 2024 on the question: *We are interested in knowing how people understand a statement made about safety and warning labels for social media companies. The statement is “Our platform puts a warning label on content that is found to be false by fact checkers.” If you were to see this statement, would you consider it to mean that:*
- All content containing ‘facts found to be false by fact-checkers’ has a warning label: (44%)
- Only individual posts found to be false by fact-checkers has a warning label: (35%)
- Don’t know: (17%)
- 17 International Fact-Checking Network 2024 *Signatories* <https://ifcncodeofprinciples.poynter.org/signatories>
- 18 Meta 2024 *Where We Have Fact Checking* <https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking/partner-map>
- 19 Digi 2021 *Terms of reference for Complaints Facility and Complaints Sub-committee* <https://digi.org.au/wp-content/uploads/2021/10/DIGI-TOR-for-Complaints-Facility-and-Complaints-Sub-committee--ACPDM--FINAL-NE-1.pdf>
- 20 Outcome 1.C, Section 5.11, Digi 2023 *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-FI NAL- -December-22-2022.docx.pdf>



- 21 We also released our concern publicly, which gathered global media coverage. It is extremely unlikely that X were not aware.
- 22 See for example Reset.Tech Australia, Human Rights Law Centre & Monash University 2023 *Legislative interventions on misinformation and disinformation – What comes next for Australia?* <https://au.reset.tech/uploads/Reset.Tech-Policy-Briefing-Misinformation-November-2023.pdf>
- 23 Reset.Tech Australia 2024 *Regulating for Transparency: Transparency Reports in Australia* <https://au.reset.tech/news/briefing-transparency-reports-in-australia/>
- 24 Reset.Tech Australia 2023 *How do platforms respond to user-reports of electoral process misinformation?* <https://au.reset.tech/news/report-electoral-process-misinformation/>
- 25 Reset.Tech Australia 2023 *Is political content over- or under-moderated?* <https://au.reset.tech/news/report-is-political-content-over-or-under-moderated/>
- 26 Reset.Tech Australia 2023 *Misinformation in paid-for advertising* <https://au.reset.tech/news/report-misinformation-in-paid-for-advertising/>
- 27 Reset.Tech Australia 2023 *Recommender systems and political content* <https://au.reset.tech/news/report-recommender-systems-and-political-content/>
- 28 X nd *How we address misinformation on Twitter* <https://help.twitter.com/en/resources/addressing-misleading-info>
- 29 Twitter Philippines 2022 *An update on reporting potential misinformation on Twitter* [https://blog.twitter.com/en\\_sea/topics/company/2022/update-on-reporting-potential-misinformation-on-twitter](https://blog.twitter.com/en_sea/topics/company/2022/update-on-reporting-potential-misinformation-on-twitter)
- 30 Images of this are available at Andrew Hutchinson 2021 'Twitter Tests New Misleading Info Reporting Option to Further Combat Misinformation in Tweets' *SocialMediaToday* <https://www.socialmediatoday.com/news/twitter-tests-new-misleading-info-reporting-option-to-further-combat-misinf/605151/>
- 31 Twitter Philippines 2022 *An update on reporting potential misinformation on Twitter* [https://blog.twitter.com/en\\_sea/topics/company/2022/update-on-reporting-potential-misinformation-on-twitter](https://blog.twitter.com/en_sea/topics/company/2022/update-on-reporting-potential-misinformation-on-twitter)
- 32 Reset.Tech 2023 *Open Letter to X* <https://au.reset.tech/news/open-letter-to-x/>



