Reset. AUSTRALIA

Susan McKinnon Foundation

Response to the Exposure Draft Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023

Reset.Tech Australia
August 2023

Summary

Reset.Tech welcomes the Government's ambition to tackle misinformation and disinformation. As we have argued elsewhere, Australia's regulatory framework is neither comprehensive nor rigorous enough to address the threats posed by mis and disinformation, including those emerging in the upcoming Voice referendum. Broader regulatory requirements that hold platforms accountable for the promotion of mis and disinformation, coupled with requirements for transparency to enable effective independent oversight, are urgently needed.

Our feedback to the Exposure Draft includes the following points:

1. **Weak foundation.** The Voluntary Code on which the proposed framework would be built is reliably *ineffective*, according to our active empirical research (see pp. 9-11). **Coverage issues.** There are gaps in the current Code's scope which have opened the gates for significant content and distribution based harms, even at this early stage of Voice referendum campaigning.

2. **The graduated approach envisaged is unnecessarily circuitous.** Rather than wait for inevitable harm that occurs while co-regulation fails, we urge a direct route to proper legislation that imposes clear duties on platforms.

3. **Legislated requirements.** The Bill should entrench key activities in legislation rather than await an uncertain and lengthy co-regulation standards-setting exercise. One such activity is mandated, privacy-preserving researcher and civil society access to platform data.

4. **Role of the regulator.** ACMA's role should be explicitly supervising and incentivising platform risk mitigation, which could follow the logic of Article 34 of the *Digital Services Act*. Paired with a legislatively mandated data access regime, ACMA can draw upon the expertise of civil society and researchers to identify risks to information integrity in real time. This approach would resolve both issues of regulator capacity, as well as enable timely interventions.

Contents

Summary 2
Contents 3
About Reset.Tech Australia & this submission 4
About the Susan McKinnon Foundation 4
Background 5
    From voluntary code to comprehensive regulation: the European experience 7
The Code is a weak foundation with efficacy and coverage issues. 8
Electoral Process Mis and Disinformation: Reset.Tech Australia's investigation into platform self-regulatory efficacy 9
The graduated approach is a circuitous route that could be straightforwardly shortened. 12
Access to platform data should be legislated rather than indeterminably delayed to a codes or standards drafting exercise. 12
ACMA would be better placed to play a role assessing legislated platform risk mitigation initiatives, with timely input from research experts and civil society. 13

About Reset.Tech Australia & this submission

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, including social media 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 mis and dis information with a focus on identifying areas for regulatory intervention.

About the Susan McKinnon Foundation

We are grateful to the Susan McKinnon Foundation for supporting our work on mis and disinformation. Information on the Susan McKinnon Foundation's mission and objectives is as follows:

> *Susan McKinnon Foundation is a nonpartisan, not-for-profit organisation that works to help Australia achieve a more fit-for-purpose political, policy and service delivery system. The Foundation was established by Grant Rule and Dr Sophie Oh with the aim of making a lasting difference to Australia by helping to enhance the capability and effectiveness of our democratic institutions and government. Misinformation can have a corrosive effect on our democratic processes and institutions by misleading voters, suppressing voter turnout, and eroding trust in democratic institutions. Our primary objective is to promote the public interest and not to support a particular agenda. By supporting initiatives that work to counter the impacts of misinformation SMF seeks to provide Australians with the opportunity to make decisions based on accurate and reliable information.*

SMF provided direct funding for the monitoring of social media platforms during the Voice Referendum with the primary objective to undertake an unbiased assessment of the DIGI Code and platforms' adherence to their own Terms of Service. This initiative sought to provide independent research that would assist in the understanding of the proposed legislation. SMF's view is that platform transparency is crucial for governments and civil society groups to detect and effectively tackle potential harms stemming from social media.

Background

Mis and disinformation is rife in the Australian information architecture, especially potentially harmful electoral mis and dis information. For example;

- A QUT study examined around 54,000 Twitter accounts during and after the 2019 Australian Federal Election (looking at over 1 million tweets). It found that 13% of accounts were 'very likely' to be bots, with the majority originating from New York.[1] This is estimated to be **more than double the rate of bot accounts in the US presidential election**. These can have big impacts: research into the US election by ANU indicated that the average bot was 2.5 times more influential than the average human, measured by success at attracting exposure via retweets.[2]
- Chinese Australians have faced misinformation in the past, often in what appear to be **coordinated disinformation campaigns**.[3] Social media platforms, such as WeChat, Weibo and Douyin have been found to serve targeted misinformation to Chinese language speakers in Australia. In 2019, WeChat in particular was a site of much political campaigning in Mandarin which included mis & disinformation.[4]
- In Reset.Tech Australia's current project monitoring toxic actors in the lead-up to the referendum, **we have identified a significant proportion of the most viral and toxic narratives about the Voice are amplified by accounts with bot-like attributes**. This suggests widespread use of automation and/or coordinated inauthentic behaviour which would presumably breach platform terms of service.

Despite the risks, Australia's current regulatory framework does not have a robust nor comprehensive approach to electoral mis and disinformation. Electoral mis and disinformation lies outside the scope of the *Online Safety Act* and the eSafety Commissioner's remit, and is often beyond the reach of even the Australian Electoral Commission.

Australia's policy response to electoral mis and dis information in particular is limited, and like many of our digital platform policies, is industry-drafted and co-regulatory. It is largely left to the Disinformation and Misinformation Code of Practice ('DIGI Code'). Industry drafted, co-regulatory models suffer from two significant constraints; industry-led drafting creates sub-standard levels of protection,[5] and, the inevitably voluntary nature of these efforts create coverage issues.[6] The problems are systemic to the model, and not isolated to the DIGI Code. While we commend the eSafety Commissioner for steering a consultation process for the Online Safety Codes, numerous accepted codes fell below international standards.[7]

---

[1] See study quoted in Felicity Caldwell (2019) 'Bots stormed Twitter in their thousands during the federal election' *SMH*
www.smh.com.au/politics/federal/bots-stormed-twitter-in-their-thousands-during-the-federal-election-20190719-p528s0.html

[2] Sherryn Groch (2018) 'Twitter bots more influential than people in US election: research' *SMH*
www.smh.com.au/national/twitter-bots-more-influential-than-people-in-us-election-research-20180913

[3] Kirsty Lawson (2020) 'WeChat the channel for China disinformation campaigns' *Canberra Times*
https://www.canberratimes.com.au/story/6802076/the-social-messaging-system-helping-spread-chinese-disinformation-campaigns/

[4] Kirsty Lawson (2020) 'WeChat the channel for China disinformation campaigns' *Canberra Times*
https://www.canberratimes.com.au/story/6802076/the-social-messaging-system-helping-spread-chinese-disinformation-campaigns/

[5] Reset Tech (2022) *How outdated approaches to regulation harm children*
https://au.reset.tech/news/how-outdated-approaches-to-regulation-harm-children-and-young-people-and-why-australia-urgently-needs-to-pivot/

[6] For example BitChute, Odyssey and Telegram are not signatories despite being available in Australia and known vectors of disinformation and misinformation. See: Adobe, Apple, Google, Meta, Microsoft, Redbubble, TikTok and Twitter. See ACMA (2022) A*ustralian Code of Practice for Disinformation and Misinformation*
https://www.acma.gov.au/online-misinformation#:~:text=you%20have%20concerns.-,Australian%20Code%20of%20Practice%20for%20Disinformation%20and%20Misinformation,%2C%20Redbubble%2C%20TikTok%20and%20Twitter.

[7] Brandon How (2023) 'Concerns raised over draft online safety codes several times' *InnovationAus*
https://www.innovationaus.com/concerns-raised-over-draft-online-safety-codes-several-times/

The systematic weaknesses of industry drafted Codes are not limited to Australia. The DIGI Code closely imitates the European Union's first attempt at online content and safety policy – the *Code of Practice on Disinformation* (2018). In the European experience, policy decision makers soon discovered the Code suffered from transparency and measurability constraints, as below:

> *At present, it remains difficult to precisely assess the timeliness, comprehensiveness and impact of the platforms' actions, as the Commission and public authorities are still very much reliant on the willingness of platforms to share information and data. The lack of access to data … (along with) the absence of meaningful KPIs to assess the effectiveness of platform's policies to counter the phenomenon, is a fundamental shortcoming of the current [EU] Code.*[8]

The 2018 Code was eventually replaced by a revised version in 2022, which has been further galvanised and strengthened by provision in the EU's *Digital Services Act*. We have compiled a timeline of the European experience overleaf.

[8] European Commission (2020) 'Staff Working Document: Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement'. Found at: https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement

From voluntary code to comprehensive regulation: the European experience

This timeline summarises the European experience and shows how legislators gradually responded to the shortcomings of the voluntary industry code with a more comprehensive package. Notably, requirements for data access were consistently invoked to ensure that there were mechanisms for independent assessments of what was otherwise mere platform self-reporting.

March 2018	**April 2018**	**September 2018**	**January 2019**	**March 2019**	**April 2019**
Final report of the High Level Expert Group on Fake News and Online Disinformation	European Commission responds with a 'Code of Practice on Disinformation' which would commit online platforms and the advertising industry to provide academia with "access to platform data"	Version 1 of the Code of Practice is released	The European Commission expresses concern on the platforms' failure to benchmark and meaningfully measure progress.	The European Commission remarks platforms "didn't provide access to more granular data to assess the effectiveness of their activities to counter disinformation"	The European Commission calls for independent data access to ensure that the platforms are "not just marking their own homework"

2019-2020	**September 2020**	**2020-2021**	**June 2022**	**November 2022**	**September 2023**
An independent assessment by EU Media Regulators (ERGA) notes no sufficient progress was made on platform commitments under the Code.	Findings from the European Commission on the first 12 months of the Code of Practice released, noting "shortcomings mainly due to the Code's self-regulatory nature".	Draft *Digital Services Act* provisions construct a data access regime with a legal basis to force VLOPs/VLOSE to provide access to data to third Parties, including regulators, vetted researchers, and civil society organisations.	Roll-out of the 'Strengthened' Code of Practice on Disinformation.	The *Digital Services Act* enters into force, including risk mitigation duties on platforms and mandated data access for regulators, civil society organisations, and accredited researchers.	First risk mitigation reporting from platforms expected under the *Digital Services Act*.

The Code is a weak foundation with efficacy and coverage issues.

We commend the ambition to build regulatory oversight over platform self-regulatory efforts. However, we note that the Code is limited to only two obligatory commitments, which focus on binary assessments as to the existence of various measures and reports, without necessary benchmarks and information to make independent assessments into efficacy. As we argue in this submission, schemes for third-party data access are one way to enable independent, expert oversight. We also note the *Digital Services Act* has effectively incentivised platforms to make meaningful progress under their own voluntary disinformation code by identifying it as a 'risk mitigating measure'. This means that the code is deployed as "the carrot to offset the potential regulatory sticks in the DSA".[9]

The tables below outline the current scope of the DIGI Code. In the next section, we provide results from our current investigation into the efficacy of platform responses to misinformation and disinformation.

Obligations		**Optional Commitments**				
Develop and implement certain measures (See Box 1)	Make and publish 'transparency reports'	Disrupt ads with mis and disinformation	'Tackle' inauthentic behaviour, and 'prohibit or manage' certain types of inauthentic behaviour	Implement measures to enable users to "make informed choices" about information	Clearly mark sources of political advertising	Support independent research on mis and disinformation

Box 1: Obligatory measures by platform signatories to the DIGI Code

- Develop and implement measures that reduce the propagation of and potential exposure of Disinformation and Misinformation to users on digital platforms
- Develop and implement measures that inform users about the types of behaviour/content they consider mis and disinformation
- Develop and implement measures that allow users to report content regulated under the Code
- Publish policies and reports that users can see about how effective platforms' measures are
- Allow users to access general information about their recommender systems

Interpretative notes

a) Compliance with these measures presumably involves an objective test of whether measures were developed or not, rather than assessment of their efficacy.
b) Obligations are subject to a proportionality test. Presumably, platforms themselves get to decide which measures are proportional, and what factors to consider in deciding if a measure is proportional or not.

[9] Mark Scott and Laura Kayali, "7 things to know about Europe's plan to boost democracy" *POLITICO*, 3rd December 2020.

Electoral Process Mis and Disinformation: Reset.Tech Australia's investigation into platform self-regulatory efficacy

Taking electoral process mis and dis information as an example, the DIGI Code allows for each platform to develop their own policies and implement their own measures to address this content. These policies often differ, as we have extracted in the table below.

Platform policies on misleading content around electoral processes		
TikTok	Facebook/Meta	Twitter/X
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.[10]	*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.*[11] Examples provided by Facebook 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.	*We may label or remove false or misleading information about how to participate in an election or other civic process.*[12] Examples include procedures to participate, voter eligibility, methods of the process or actions of electoral officials. *We may label or remove false or misleading information intended to undermine public confidence in an election or other civic process.*[13] Examples include unverified information about election rigging, ballot tampering, vote tallying, or certification of election results.

How this relates to Reset.Tech's experiment		
TikTok	Facebook/Meta	Twitter/X
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 from the FYF by TikTok when it is discovered.	Facebook's policy covers **some** attacks on electoral process integrity. Claims of rigging, stolen votes or AEC malpractice would likely fall into a lower category of misinformation where Facebook focuses on **reducing its prevalence**. This requires fact checkers to have investigated content before Facebook takes action. It is unclear what measures are taken to 'reduce prevalence', but this could include labelling this content or reducing its visibility.	X's definition covers issues of claims of rigging, stolen votes or AEC malpractice via the category of misleading information about 'how to participate' or 'outcomes'. This means X should, according to their community guidelines, label or remove this information when it is discovered.

[10] TikTok 2023 *Civic and election integrity* https://www.tiktok.com/community-guidelines/en/integrity-authenticity/
[11] Meta 2023 *Community Standards: Misinformation* https://transparency.fb.com/en-gb/policies/community-standards/misinformation/
[12] X 2023 *Civic integrity misleading information policy* https://help.twitter.com/en/rules-and-policies/election-integrity-policy
[13] Ibid.

At Reset.Tech **we ran a rapid experiment for this submission to see if these three platforms were enforcing their own community guidelines, and meeting their obligations under the DIGI Code** to implement the measures they developed to provide safeguards against the harms that may arise from misinformation and disinformation.

We found 99 pieces of content that included false or misleading claims about Australia's electoral process, with a focus on the upcoming referendum. These were all found relatively quickly through searching or exploring accounts previously known to Reset.Tech Australia. This included:

a. **25 pieces of content on TikTok**. Content centred around claims of election rigging and vote stealing, but also often also elaborated and included claims that the AEC were corrupt, that the referendum is illegal, taking part is treasonous or will affect your citizenship, or that voting is invalid because Australia is governed maritime law, or a corporation or controlled by the WEF, or UN etc.
b. **24 pieces of content on Facebook.**[14] Content centred around claims that Australian elections were or were going to be rigged, stolen ballots, and two posts around voter suppression, claiming that voting a particular way would lead to voters being de-banked, or calls to boycott voting because it was treasonous.
c. **50 pieces of content on X (Twitter)**. Content centred around claims that Australian referendums had been rigged, that the Voice would be rigged, or the occasional piece suggesting that the referendum process was illegal because Australia's constitution was invalid.

The narratives that were being pushed by this body of content have been extensively fact-checked and found to be false. For example:

- **Claims that ballots have been removed or stolen** in previous elections,, often as a way to paint a picture of widespread election rigging. AAP and RMIT FactLab have deemed this to be false.[1516]
- **Claims that the referendum vote itself is illegal or fraudulent**. RMITFactLab have deemed this to be false.[17]
- **Claims that the referendum is going to be rigged**, including reference to electronic voting systems. RMITFactLab have deemed this to be false.[18]
- **Claims that the question asked in the referendum will be deceptive or disguised as multiple questions to maximise the chance of passing**. AAP have found this to be false.[19]
- **Claims that the question on the ballot will be 'rigged' or sneakily written to create a new state or end Australia's sovereignty**. These claims have been found false by both the AAP[20] and RMITFactLab.[21]

[14] We intended to track 25 pieces, but included a duplicate in error
[15] 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/
[16] 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
[17] 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
[18] 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
[19] AAP 2023 *One simple answer to five questions claim*
https://www.aap.com.au/factcheck/one-simple-answer-to-five-questions-claim/
[20] 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/
[21] 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

We monitored how platforms responded to this content, both in terms of takedown and labelling of content. We took a reading for an initial week, to establish a baseline for what platforms do more or less 'organically'. Then, we reported each piece of content via each platform's reporting mechanism.

We then tracked and monitored these 99 pieces of content for ten days. Given the deadlines for this submission, this provides the best available estimate to measure the platform's responsiveness to user reporting on breaches of electoral process misinformation rules. We found that the vast majority was still available online despite platforms being made aware of it through their online reporting systems. This strongly suggests that **platforms do not implement the measures they commit to under the DIGI Code.** We will be releasing the full data post the submission deadline, including monitoring for 14 days and growth rates over the full three week period.[22]

	Removal		**Labelling**	
Initial week **before reporting**	TikTok[23]	1	TikTok	0
	Facebook	0	Facebook[24]	1
	X	0	X	0
Total response, 10 days after reporting Please note these findings are preliminary, and we will be collecting data at the 14-day mark.	TikTok[25]	8	TikTok	0
	Facebook	0	Facebook[26]	1
	X	0	X	0

[22] The submission deadline fell a few days too early to include our final analysis, we would be happy to provide it in a further submission.
[23] We cannot be sure if this was removed by the platform or the users themselves.
[24] In addition to one post that was already labelled at the start of the research.
[25] On TikTok, 7 posts were removed during this period, on top of the one removed in the initial week. Please note, these may have been removed by the platform or the users themselves.
[26] No posts were *additionally* labelled in the 10 days after reporting. This was the post labelled in the initial week.

The graduated approach is a circuitous route that could be straightforwardly shortened.

Our rapid experiment clearly details there are shortcomings in how platforms meet their obligations under the DIGI Code. We urge consideration of a more direct route to more fulsome regulation, rather than repeating the well-known and directly analogous errors from the European experience. Arguments for industry-led tech regulation generally rest on an assumption that the alternative is burdensome and anathema to innovation. This assumption is misguided, and recent empirical evidence from similar processes indicate it is flawed in practice as well.

Self and co-regulatory processes place significant burdens on industry to step into the world of policy-making, reviewing legal requirements, interfacing with regulators, undertaking community consultations, including with vulnerable communities, and drafting nuanced public policy. Policy making is a unique skillset, and co-regulation places an additional burden on industry to assume that they either have or can recruit for these skills. This is especially true for small and medium industry actors, who end up being reliant on large international platforms to lead the process that ultimately shapes the ecosystem they develop within. This burden and reliance does not help Australian innovation take root.

As the recent experience of the Online Safety Codes outlined, co-regulation is not a trivial pursuit and requires significant time and expertise. When industry needed to repeat a stage of the process due to shortfalls in proposed codes, this burden was felt by both industry and civil society. The burden on civil society to engage constructively and productively throughout this process merits particular mention. In Reset.Tech's experience, this role was played relying wholly on external philanthropic funding, in a reactive manner to industry-led drafting, often requiring specialist input on tight timelines. It would be a pity to entrench this dynamic in Australian tech regulation, especially when more effective, resource-conscious, and evenly distributed options exist. We draw attention to the final section of this submission, where we indicate what arrangements have been constructed in Europe to draw upon the expertise of accredited researchers and civil society organisations to assist with timely and thoughtful regulatory interventions.

Access to platform data should be legislated rather than indeterminably delayed to a codes or standards drafting exercise.

Access to platform data is urgently needed by researchers and civil society to effectively and independently monitor misinformation and disinformation, particularly in the context of the Voice referendum. The turbulence at X/Twitter shows the vulnerabilities for public interest research when data access is left to the goodwill of large and offshore companies.

In August, X (formerly Twitter) announced unprecedented legal action against the Centre for Countering Digital Hate. The case, which began as a legal threat to the CCDH for making misleading claims around X's inaction on harmful content, proceeded to become legal action around the way the CCDH accesses and collects information about the public content that is shared on X. The complaint alleges that the way CCDH collects data and stores it in analysable formats—often called scraping—is a violation of their terms of service. Further, X claims that CCDH misuse an analytic tool called

Brandwatch. Both of these are ubiquitous research tools, and since X closed its third party API access, there is no other way to understand what is happening on the platform at scale.

X's legal action has serious implications for the mis and dis information research community in Australia. It creates a challenging risk environment for researchers seeking to examine public interest issues around the upcoming Voice referendum. This is not an issue that is limited to the X platform alone. Last year, Meta announced it was closing Crowdtangle, which gave researchers access to analyse what was happening on Facebook at scale. There is an unprecedented, deliberate shift from digital platforms away from transparency. This coupled with the threat of legal action for researchers who use alternative means, such as Brandwatch and scraping, makes analysing the Australian social media landscape almost impossible.

As a result, there is no clear way to monitor and analyse, for example, what misleading narratives are emerging on platforms, or detect bot networks or other signs of coordinated inauthentic behaviour—including from hostile foreign state actors. The only organisations with access to that vital public policy information are the US based platforms themselves, and we would be reliant on their goodwill to ensure that Australian platform rules are effectively implemented, and that our information architecture remains safe. Goodwill is clearly not a sustainable strategy here.

In Europe, legislators have enshrined the principle of researcher access to data where that research is in the public interest. Notably, the platform data access regime involves three key recipient groups: regulators, civil society organisations, and accredited researchers. As civil society groups and researchers have consistently argued, access to platform data is vital not only in the context of tech accountability, but also for a wide range of public interest research questions in the social sciences, such as physical health, and mental health and wellbeing. We have provided more detail on the regime in the next section.

Beyond Europe, similar proposals are supported in draft Canadian legislation (the *Online Safety Bill*) draft US legislation (the *Kids Online Safety Act*). We also note rich debates in the UK Parliament throughout the year on the matter, in the context of both the *Online Safety Bill* and the *Data and Digital Protection Bill*. There is a clear imperative to ensure that Australian researchers can access data without legal threats, to interpret and analyse uniquely Australian issues. This would require a simple amendment to the Exposure Draft, which could be modelled relatively easily from overseas provisions.

ACMA would be better placed to play a role assessing legislated platform risk mitigation initiatives, with timely input from research experts and civil society.

Our submission has detailed **a)** evidence that the platforms are not effectively enforcing their own rules, **b)** current Australian approaches to tech co-regulation are burdensome on both industry *and* civil society and have created lower standards than international benchmarks, and **c)** absent a platform data access regime, platforms are merely marking their own homework with no route for independent oversight. **Simply, Australia is where Europe was in 2019**. We must act decisively so that Australians can be availed of the comparable platform obligations as in this sophisticated and world-leading market.

Rather than crafting a framework of prolonged uncertainty for platform misinformation and disinformation obligations in Australia, we recommend legislation for ACMA to supervise two

concurrent, interconnected processes: **a platform risk mitigation framework**, with third-party input from researchers and civil society *via* **a data access regime**. We have extracted the key points from the European data access regime below.

The European data access regime[27]

There are three key initiatives around researcher access to platform data. They are:

1. A framework under the *Digital Services Act*[28] for accredited researchers and civil society to request data from very large platforms and search engines via the relevant regulator (***reactive*** data sharing).

2. Industry promises under the *Strengthened Code of Practice on Disinformation* (2022)[29] to make relevant data available to researchers (***proactive*** data sharing). Unlike the *Digital Services Act* framework, this scheme is semi-voluntary. It links to the *Digital Services Act* as an example of a 'risk mitigating measure', meaning that companies can point to their performance under the *Code of Practice* to decrease the risk of regulatory retaliatory action.

3. A draft framework for an independent, third-party intermediary body for vetting data access requests by the *European Digital Media Observatory*.[30] The existence of a future intermediary body is explicitly mentioned in the *Digital Services Act* and the *Code of Practice* includes a co-funding commitment from companies.

On risk mitigation frameworks, we note our experience monitoring for platform-based risks in the lead-up to the referendum, and the present situation of 'ad hoc' or informal escalation channels. There is no comprehensive regulatory mechanism available to seek recourse for and to tackle identified risks. In Europe, Reset.Tech affiliates have been developing metrics to support regulators in this task. We have extracted them at the end of this document to indicate what sorts of particularly distribution-based metrics could also make sense in Australia. For full context, we have also extracted the source legislation: the baseline framework from Article 34 of the *Digital Services Act*.

The central requirement of Article 34 is to observe content *and* behaviour on digital platform services that implicates the various categories of risk and submit an assessment of those where the level of severity rises to the level of "**systemic risk**." This requires answering the following question: **At what point does content or conduct on a service cross the threshold, resulting in an "actual or foreseeable negative effect" that is severe enough to be considered systemic in relation to the risk factors specified in the regulation**? We provide this framework as an alternative model for regulator oversight, which with appropriate adaptation for the Australian legislative environment, may overcome some of the challenges from the present framing of content harms.

[27] As conveyed in Mathias Vermeulen (2022) "Researcher Access to Platform Data: European Developments" *Journal of Online Trust and Safety* https://tsjournal.org/index.php/jots/article/view/84/31
[28] See:https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/09-12/p3-2020_0361COR01_EN.pdf
[29] https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation
[30] https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

Article 34

...This risk assessment shall be specific to their services and proportionate to the systemic risks, taking into consideration their severity and probability, and shall include the following systemic risks:

(a) **the dissemination of illegal content through their services**;

(b) **any actual or foreseeable negative effects for the exercise of fundamental rights**, in particular the fundamental rights to human dignity enshrined in Article 1 of the Charter, to respect for private and family life enshrined in Article 7 of the Charter, to the protection of personal data enshrined in Article 8 of the Charter, to freedom of expression and information, including the freedom and pluralism of the media, enshrined in Article 11 of the Charter, to non-discrimination enshrined in Article 21 of the Charter, to respect for the rights of the child enshrined in Article 24 of the Charter and to a high-level of consumer protection enshrined in Article 38 of the Charter;

(c) **any actual or foreseeable negative effects on civic discourse and electoral processes, and public security**;

(d) any **actual or foreseeable negative effects in relation to gender-based violence**, the **protection of public health and minors** and serious **negative consequences to the person's physical and mental well-being**.

2. When conducting risk assessments, providers of very large online platforms and of very large online search engines shall take into account, in particular, whether and how the following factors influence any of the systemic risks referred to in paragraph 1:

(a) the design of their recommender systems and any other relevant algorithmic system;

(b) their content moderation systems;

(c) the applicable terms and conditions and their enforcement;

(d) systems for selecting and presenting advertisements;

(e) data related practices of the provider.

The assessments shall also analyse whether and how the risks pursuant to paragraph 1 are influenced by intentional manipulation of their service, including by inauthentic use or automated exploitation of the service, as well as the amplification and potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions.

The assessment shall take into account specific regional or linguistic aspects, including when specific to a Member State.

Proposed metrics informing mis and disinformation regulation in the EU

Reset.Tech affiliates in the EU have been monitoring mis and disinformation in a range of global elections, including elections in the UK, the US, Kenya, Germany, France and Brazil, with a view to understanding the sorts of metrics that can be derived to enable regulatory action.

With the *Digital Services Act* taking effect across the EU, regulators are able to demand proportionate and proactive mitigation from social media platforms to reduce the risks of mis and dis information. Reset.Tech have developed a set of metrics for the European context, including the below. We would be happy to share the full list, should that be helpful. Key examples are provided below.

- Average engagement with disinformation vs genuine content
- Average growth rate for disinformation pages/actors vs genuine pages/actors
- Non follower engagement rates (on YouTube and Twitter)
- Content moderation indicator (response and notice reaction rates)
- Average toxicity score of comments, by actor or #