The Challenge of Regulating "Misinformation" and "Disinformation"

Mark Humphery-Jenner[*†]

It is often argued that "misinformation" and "disinformation" should be suppressed as they perpetuate false or misleading narratives, and can contribute to, or cause, significant harm. However, such anti-misinformation rules face significant challenges and can undermine free speech. Australia's foray into anti-misinformation rules provides a cautionary example. The problems start with broad and uncertain definitions of "misinformation". Further, due to self-interest, lobbying, or misplaced assumptions of accuracy, powerful interests are often exempted from anti-misinformation rules. Enforcement mechanisms also risk suppressing innocuous content in service of reducing the chance of "misinformation" falling through the cracks. Anti-misinformation rules thus face significant problems that must be solved if they are to be implemented.

INTRODUCTION

Homosexuality was once considered a crime. This was premised on the erroneous belief that homosexuality was not only unnatural, but harmful to society. Of course, such assumptions are premised on misinformation. NSW only decriminalised homosexuality in 1984.[1] In 2014, NSW allowed for people to have their convictions extinguished.[2] In 2024, the NSW government issued an apology to people convicted under the old laws.[3] But, the lesson from this set of events is that the government can itself spread, believe in, and enforce misinformation. Relatedly, the government can enforce falsehoods while accusing challengers of spreading "misinformation". Thus, seeking to suppress alleged misinformation can inadvertently entrench wrong beliefs. It is against this backdrop that we must evaluate legislative attempts to suppress "misinformation" and "disinformation".

Free speech is intrinsic to challenging incorrect government policies. Consider Australia: successive governments once pursued the "White Australia" policy and implemented policies that resulted in the Stolen Generations. At the time, those governments falsely assumed—and claimed—that such policies were valid and that challenges to them would result in significant harm to society. However, challenging such policies is essential for social justice reform. But, suppose people assumed that those racist beliefs were correct, as some powerful interests once did. Now suppose that those same powerful interests suppressed challenges to their racist beliefs. Would social justice reform occur, or occur at the same rate?

---

[*] This article is based on my submission to the Senate Standing Committee on Environment and Communications in relation to the Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024. It also extends and updates the analysis on the exposure draft of that bill, as published in Mark Humphery-Jenner, "Legislating Against Misinformation: Lessons from Australia's Misinformation Bill" (2024) 45 *Statute Law Review* hmae023. The author thanks the reviewers and editors for their feedback on this article.

[†] UNSW Business School and Otso Capital. The author has no financial interest in the subject matter of this article and received no funding for it.

[1] *Crimes (Amendment) Act 1984* (NSW).

[2] NSW Government, "Extinguishment of Historical Homosexual Convictions" (6 June 2024) <https://dcj.nsw.gov.au/legal-and-justice/laws-and-legislation/royal-prerogative-of-mercy-and-reviews-of-convictions-sentences/extinguishment-of-historical-homosexual-convictions.html> .

[3] NSW Government, "NSW Government to Issue Formal Apology on the 40th Anniversary of Decriminalisation of Homosexuality" (11 May 2024) <https://www.nsw.gov.au/media-releases/nsw-government-to-issue-formal-apology-on-40th-anniversary-of-decriminalisation-of-homosexuality>.

Free speech and open debate are also essential to advance science and human knowledge. The scientific method involves challenging status quo beliefs to correct the record if the record is wrong. However, if the record is assumed to be immutably true, then challenges are *ex ante* assumed to be false and misleading, by default. If those same challenges are then suppressed, the scientific method itself is suppressed. Such knowledge advancement does not just occur in universities or in think tanks. It can occur in private enterprise. The challenges might once have been regarded as having the potential to harm health, society or the economy. Thus, attempts to suppress such challenges would entrench a status quo that is inaccurate, but is never discovered to be such.

Australia's government has attempted to suppress "misinformation" and "disinformation". To this end, the government introduced the Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth) (hereafter, the "Misinformation Bill"). This followed from a prior exposure draft. The Communications Minister highlighted that this would empower the minister to direct the relevant regulator to investigate "matters brought to the government's attention through public complaints or the media about online dis- and misinformation".[4] The Misinformation Bill would essentially deputise social media platforms to police misinformation and disinformation. To this end, they would have to develop codes relating to how they would suppress such content, and would have to back up those codes with enforcement. The regulator would oversee and monitor social media companies' activities.

The Misinformation Bill was not well received when it was introduced. The original iteration of the bill was an "exposure draft". The government solicited public feedback on the draft. The relevant regulator – ACMA – recorded a summary of the public feedback to the initial exposure draft on its website.[5] Such is the concern over the exposure draft that ACMA divided feedback into sections relating to "censorship", "freedom of expression" and the government being the "arbiter of truth". To this end, a word cloud of the feedback is in Figure 1.[6] The words "censorship" and "government" feature prominently and underscore the problem. Similarly, the comments have a linguistic negativity score that is over 40% higher than that of the explanatory memorandum.[7] The feedback on the initial exposure draft was overwhelmingly negative. The government subsequently amended the Bill, but the amended version is not significantly different from the original exposure draft.

[4] See the reported letter from Communications Minister Michelle Rowland to Prime Minister Anthony Albanese: "Minister for Misinformation's Secret Powers Prompt Free Speech Fears", *The Australian* (online) https://www.theaustralian.com.au/nation/minister-for-misinformations-secret-powers-prompt-free-speech-fears/news-story/efad79484ef4fc9d9c7d9d252353d1a5>. The mechanism of action is twofold. First, the minister can already order the regulator to investigate matters (Broadcasting Services Act 1992) and to give directions to the regulator (Communications and Media Authority Act 2005). The argued effect of the Misinformation Bill would be to extend the matters over which investigations could arise. The Minister would also be able to refer matters of misinformation by notifying the regulator, which can lead to one-sided investigations: the minister could notify the regulator of opponents' misinformation while failing to notify about pro-government misinformation.

[5] Department of Infrastructure, Transport, Regional Development, Communications and the Arts, "New ACMA Powers to Combat Misinformation and Disinformation" (2024) <https://www.infrastructure.gov.au/have-your-say/new-acma-powers-combat-misinformation-and-disinformation>; archived version: https://web.archive.org/web/20240924023930/https://www.infrastructure.gov.au/have-your-say/new-acma-powers-combat-misinformation-and-disinformation

[6] The word cloud was created using the Python package WordCloud.

[7] This was created by feeding the PDF of the explanatory memorandum and a PDF of the comment table into Python and using the SentimentIntensityAnalyzer package.



Figure 1: Word cloud from public feedback on the ACMA website.

The Misinformation Bill is one proposed method to potentially suppress "misinformation" and "disinformation". However, there are other approaches. For example, in late 2026, the Albanese government proposed a "Digital Duty of Care". Indeed, Zoe Daniel MP introduced a private member's Bill that sought to suppress "harmful" content.[8] The Bill would impose a duty of care on certain online service providers, requiring them to "take reasonable steps to prevent matters from arising that would (or would be likely to) cause harm or detriment to Australian end-users of the service".[9] This is similar to the government's proposal: it would deputise social media platforms to police certain types of content.

This raises the issue of what specific challenges arise when attempting to suppress "misinformation" and "disinformation". This article therefore analyses the problems with anti-misinformation legislation. The issues are generalisable across jurisdictions. However, this article uses Australia's Bill as a concrete example. This article focuses on three fundamental problems, which combine to undermine free discourse and create significant harm to free speech. These include issues with how to define misinformation, exemptions from anti-misinformation rules, and issues with enforcement.

The first major issue relates to how "misinformation" and "disinformation" are defined. This issue is not unique to anti-misinformation rules. All statutes have words that require definition. However, statutes should aim to avoid broad words whose breadth itself can cause harm. For example, Australia's anti-misinformation rules define misinformation as content that "is reasonably verifiable as false, misleading or deceptive" and that can cause *or contribute to* serious harm. In so doing, it raises the question of what exactly does it mean for content to be "misleading" or "false" and when does it "contribute to" harm? Indeed, in this context, it confronts the issue that if existing beliefs are deemed to be accurate, even if they are not, then anti-misinformation rules risk suppressing challenges to those false beliefs.

The next concern is in relation to who anti-misinformation rules cover. Anti-misinformation rules might exclude various parties as a result of self-interest on the part of government, lobbying from interest groups, or assumptions about whether that party is inherently accurate. However, lobbying and self-interest are not normatively valid reasons to exclude a party from anti-misinformation rules. Further, if a party is per se assumed to be accurate, it entrenches that party and chills challenges to that party or its content. Thus, while the party might currently be truthful or accurate, it need not always be the case. In turn, anti-misinformation rules can risk entrenching vested interests and rendering them insulated from criticism.

[8] Online Safety Amendment (Digital Duty of Care) Bill 2024 (Cth).

[9] Online Safety Amendment (Digital Duty of Care) Bill 2024 (Cth) s 28D(b).

The third concern addressed in this article is in relation to enforcement. If anti-misinformation rules are to exist, they should accurately suppress "misinformation" and "disinformation". But, they should also avoid inaccurately suppressing innocuous content. However, a consequence of the exemptions already noted is that misinformation from "exempt" parties can propagate. However, anti-misinformation rules – such as those in Australia – can penalise platforms for hosting "misinformation", which has a broad and uncertain definition, but do not penalise them for suppressing innocuous content. This incentivises platforms to "err on the side of caution" and suppress more content, rather than less.

The combined result of these problems is a fundamental challenge to free speech. Australia's anti-misinformation rules are a cautionary tale, providing an example of pitfalls to avoid when attempting to suppress "misinformation" and "disinformation". The net result of such rules would be to entrench content from certain parties, deter or suppress challenges to entrenched narratives, and broadly suppress content. The remainder of this article highlights the underlying issues with such rules, drawing on Australia's anti-misinformation rules as a case study.

PROBLEM 1: WHAT IS MISINFORMATION?

The problems with anti-misinformation legislation start with how we define "misinformation" or "disinformation". If the definition is vague, or open to interpretation, the law can over-suppress content or speech, thereby undermining free expression. There are several approaches to defining "misinformation". The Australian anti-misinformation bill focuses on content that is false or misleading. However, an alternative bill requires digital service providers "to take reasonable steps to prevent matters from arising that would (or would be likely to) cause harm or detriment".[10] Thus, there are common elements to anti-misinformation regulations. In turn, whether anti-misinformation regulations strike the right "balance" between protection and free speech partly depends on the definition of "misinformation".

The Australian anti-misinformation bill provides a cautionary example. The Bill defines "misinformation" broadly as content that "is reasonably verifiable as false, misleading or deceptive".[11] There is no *mens rea* element: the person need not have intended to mislead or deceive. The content is "disinformation" if, in addition, the person disseminating it either "intends that the content deceive the other person" or "the dissemination involves inauthentic behaviour".[12] In each case, it must be capable of causing serious harm. In both cases, content from specified sources is "excluded" or exempt and is effectively deemed to *not* be misinformation or disinformation.[13] There are issues throughout the definition of misinformation and these issues are endemic in attempts to suppress such content.

What is "false" content?

Factually false information is typically seen as misinformation. The Australian anti-misinformation Bill aims to curtail content that is "reasonably verifiable" as false. It requires content platforms to have codes and procedures to identify and screen "false" content. This raises the difficult question of what, precisely, constitutes "false" content. The issue is fraught because content once thought false may turn out to be true, while beliefs once fervently held to be factual may later prove false. If existing "knowledge" is entrenched and challenges to it suppressed, how can we discover that it is in fact mistaken?

Advances in knowledge happen frequently in scientific literature and in data-driven research. This can be for several reasons. For example, it could be that the "old" research used empirical techniques that have since been discovered to be flawed. This can happen even if the "old" research was conducted in good faith. For example, there have been advances in how we identify whether things are causally

[10] Online Safety Amendment (Digital Duty of Care) Bill 2024 (Cth) s 28D(b).

[11] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth) s 13(1).

[12] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth) s 13(2)(e).

[13] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth) ss 13(1)(d) and 13(2)(d).

related.[14] Thus, under "old" techniques, we might have believed that A caused B when that is not really the case. Alternatively, prior research might have believed a certain result was "universally true" when in fact the result was conditional on other factors. For example, anti-takeover provisions were once believed to reduce corporate value per se.[15] However, subsequent research suggests that this might be conditional on the nature of the firm.[16]

Anti-misinformation rules try to circumvent this by exempting universities, or academic research, from their ambit. However, this assumes that knowledge only advances within academic or established scientific communities and that it is only disseminated under their auspices. By contrast, knowledge can often advance in commercial settings. For example, IP Australia notes that "across a range of measures, the contribution of small and medium enterprise to innovation in Australia has been rising in recent decades"[17] and that "in the past two decades" the number of small and medium enterprises with patents has increased five times faster than the number of firms *per se*.[18] The documented level of patenting implies that commercial enterprises do generate additional knowledge and that such advancements are not limited to universities or established research institutions. This also applies in fields such as finance and business where there is a commercial imperative to develop new analytical techniques to improve investor returns, which itself implies advancing knowledge further than latent academic research.

The net result of this is that it is challenging to identify what is "false". What is "false" changes over time. And, assuming that challenges to the status quo are themselves "false" can simply entrench incorrect beliefs.

What is "misleading"?

Content is sometimes deemed to be misinformation if it is "misleading". The question is then what exactly is "misleading". In Australia's anti-misinformation bill, and in the prior exposure draft, the term is undefined. In Australia's case, this means that the regulator would determine what is "misleading" and then determine whether social media companies' codes adequately address that content. The definitions used in other contexts – such as consumer law – are only a partial solution. This is because the consumer-related contexts typically pertain to concerns about whether an item's, or a service's, condition, efficacy, or characteristics are misrepresented. By contrast, when applied broadly to commentary about scientific, environmental, or economic debates, value-judgments can arise. For example, the Labor government under Anthony Albanese might argue that the Greens' economic policies and rhetoric are misleading or deceptive. The Greens would presumably argue the contrary. Presumably, the Bill does not intend to suppress rhetoric from political parties. However, it would have the consequence of doing so.

Let us take an example: In Australia, the Greens demand that the government force the Reserve Bank of Australia (RBA) to lower interest rates. They specifically state "The Reserve Bank should lower interest rates tomorrow and if they don't, the Treasurer should use his existing powers to bring down

[14] AC Baker, DF Larcker and CCY Wang, "How Much Should We Trust Staggered Difference-in-Differences Estimates?" (2022) 144 *Journal of Financial Economics* 370.

[15] P Gompers, J Ishii and A Metrick, "Corporate Governance and Equity Prices" (2003) 118 *Quarterly Journal of Economics* 107; L Bebchuk, A Cohen and A Ferrell, "What Matters in Corporate Governance?" (2009) 22 *Review of Financial Studies* 783; RW Masulis, C Wang and F Xie, "Corporate Governance and Acquirer Returns" (2007) 62 *Journal of Finance* 1851; J Harford, ML Humphery-Jenner and RG Powell, "The Sources of Value Destruction in Acquisitions by Entrenched Managers" (2012) 106 *Journal of Financial Economics* 247.

[16] Mark L Humphery-Jenner, "Takeover Defenses, Innovation, and Value-Creation: Evidence from Acquisition Decisions" (2014) 35 *Strategic Management Journal* 668; William C Johnson, Jonathan M Karpoff and Sangho Yi, "The Life Cycle Effects of Corporate Takeover Defenses" (2022) 35 *Review of Financial Studies* 2879.

[17] IP Australia, *Australian IP Report 2024* (Report, 2024) 14 <https://www.ipaustralia.gov.au/tools-and-research/professional-resources/data-research-and-reports/~/-/media/Project/IPA/IPAustralia/PDF/IP-Reports/IP-Report-2024.pdf?rev=5a30f1001682404b8fb691d41f864b49>.

[18] IP Australia, n 17, 10.

interest rates and relieve pressure on mortgage holders".[19] The underlying premise is that this would be economically beneficial. That premise is misleading. The government has rejected this. Senator Gallagher is quoted as saying "I just think the Greens are out of control at the moment, full of self-importance and out seeking a populist approach to everything. It's crazy what they're saying to us, […], If that's their ultimatum … no, we won't work with that because that is crazy. It's economically irresponsible, and we won't do it".[20] Similarly, former ALP Senator Graham Richardson stated that the Greens' position was "monumentally stupid".[21] The ALP is correct as the Greens' proposal and rhetoric misleadingly asserts that this will be economically beneficial. However, it would undermine economic independence, which is the bedrock of functional central banking, and cause serious economic harm.[22] Thus, the Misinformation Bill would presumably task social media companies with suppressing policy statements from the Greens, and from Greens supporters who advocate for such positions.

Australia's anti-misinformation bill also followed from a failed referendum, which also highlights the difficulty of identifying "misleading" information. The referendum related to the creation of the "Aboriginal and Torres Strait Islander Voice" (hereafter, "The Voice"). The Voice would have been empowered to make representations to the executive and to parliament. There were disputes about the ambit of the Voice body's power, resulting in both the "yes" and "no" campaigns accusing each other to be spreading misinformation. Australia's anti-misinformation Bill specifically excludes "referendum" or "election" discussions from its purview. However, the Voice debate provides an illustrative example of how difficult it can be to identify "misleading" content in a contentious context.

One key Voice-flashpoint suffices to illustrate the conundrum: would the Voice have been "merely advisory"? Prime Minister Anthony Albanese asserted that "we have had a clear proposition of just two things - recognition and a non-binding advisory committee".[23] This was seemingly based on a literal reading of sub-section (ii) of the referendum text that the body "may make representations to the Parliament and the Executive Government of the Commonwealth". However, the comments appeared to omit sub-section (iii), which stated that parliament would make laws governing the Voice's "composition, functions, powers and procedures". Thus, Voice architects indicated that the Voice was in fact a "structural reform" and "without the Indigenous voice to parliament, a treaty is vulnerable"[24]. Further, there was the argument that unless countervailing views were given similar access and funding, the Voice would have *de facto* power, analogous to only hearing one side of a debate.[25]

The problem is then that anti-misinformation bills falsely assume that it is easy to identify what is "misleading" and suppress content accordingly. Anti-misinformation laws force content platforms to make a judgment about whether content is misleading. But, as the foregoing examples highlight, many of the examples of "misleading" content are normatively driven. Either side of the normative divide

[19] The Australian Greens, "Greens Won't Support RBA Reforms until Labor Tells the Bank to Cut Rates" (Media Release, 2024) <https://greens.org.au/news/media-release/greens-wont-support-rba-reforms-until-labor-tells-bank-cut-rates>.

[20] M Read and T McIlroy, "'Greens Are out of Control': RBA Reform Dead as Labor Rejects Demands" *Australian Financial Review* (online, 23 September 2024) <https://www.afr.com/policy/economy/greens-are-out-of-control-rba-reform-dead-as-labor-rejects-demands-20240923-p5kcm3> .

[21] G Richardson comments on *Sky News Australia* (YouTube, 23 September 2024) <https://www.youtube.com/watch?v=lemrf_K0kfg>

[22] JH Powell, "Chair Powell's Press Conference" (Federal Reserve, 18 September 2024) <https://www.federalreserve.gov/mediacenter/files/FOMCpresconf20240918.pdf>.

[23] Anthony Albanese, "Press Conference - Parliament House, Canberra" (Transcript, 15 October 2023) <https://www.pm.gov.au/media/press-conference-parliament-house-canberra-17> .

[24] M Davis and G Williams, "Without the Indigenous Voice to Parliament, a Treaty Is Vulnerable" *The Australian* (online, 29 July 2023) <https://www.theaustralian.com.au/inquirer/without-the-indigenous-voice-to-parliament-a-treaty-is-vulnerable/news-story/c981e109347357241dd87c435a89b2de>.

[25] P Swan and M Humphery-Jenner, "Six Reasons to Read the Fine Print on Indigenous Voice" *The Australian* (online, 12 September 2023) <https://www.theaustralian.com.au/commentary/six-reasons-to-read-the-fine-print-on-indigenous-voice/news-story/9358a05572076120ef5030a1bb10201e>.

would accuse the other of spreading "misleading" content. In some cases, what is "misleading" is clear cut to experts in the field. However, experts often disagree. And, should ordinary citizens' speech be suppressed, rather than corrected, it risks creating resentment and entrenching their latent misleading opinions.

Must the content be harmful to fall foul of the laws?

Content must typically be "harmful" for it to qualify as misinformation. This is to avoid an over-broad prohibition. Australia's anti-misinformation bill provides a clear example. For speech to qualify as "Misinformation" or "Disinformation" it must be "reasonably likely to cause or contribute to serious harm".[26] However, there are problems, which would be endemic to any attempt to proscribe "misinformation". Specifically:

What is "harm" in this context?

(a) it need only be "reasonably likely" to do so;
(b) it need only "contribute" to the harm; and
(c) "harm" is broadly defined.

Anti-misinformation Bills might also require that the content be "harmful", in addition to being "misleading", "false", or "deceptive". In so doing, the bills can attempt to narrow the scope of the anti-misinformation legislation by directing attention to specific types of content, such as content that could cause specific types of harm. This is the approach that Australia's anti-misinformation bill took. It required that the offending content cause or contribute to several types of enumerated harm. However, this was defined broadly, including harm to the electoral process, public health, and the economy. Such is the breadth that it is worth extracting the relevant section of the bill.

The enumerated subsections, and types of harm, all pose challenges, causing normatively legitimate content to be suppressed. We can see this in several areas.

Health-related content is a key flashpoint in anti-misinformation regulation. The anti-misinformation bill also sought to curtail content that could harm people's health, or preventative health measures. Part of the policy purpose appears to be to suppress false or misleading medical claims about purported treatments. However, the anti-misinformation bill also proscribes content that undermines "the efficacy of preventative health measures". Presumably, this would proscribe content that questions the efficacy, essentiality, or desirability of measures such as government-mandated lockdowns during the Covid pandemic. However, this raises the issue that all preventative health measures are a trade-off. In the case of lockdowns, there is a trade-off between the potential to reduce the spread of disease and (*inter alia*) the broader economic damage of closing businesses, the social issues of isolation, and educational development concerns. Raising these latter concerns would arguably undermine the efficacy of preventative health measures by reducing compliance. In so doing, capturing such content can undermine nuanced debate, thereby undermining free speech.

The Bill could suppress content that harms the "integrity" of the democratic process. This appears to encompass two main categories of content:

(a) content that is false or misleading in relation to policies or candidates; and
(b) false or misleading content about the veracity or accuracy of an election result. As indicated above, identifying what is "misleading" in the context of a contentious referendum (or election) campaign is fraught. The references to the "integrity" of the democratic process appear aimed at concerns such as "election result denial" and levelling false allegations about election fraud. The concern here is that any substantiated allegation begins life as an "unproven" allegation. Some allegations have no evidentiary support. Some do. But, where there is evidence, it must be tested. Until the evidence is tested, the allegation remains unsubstantiated. Thus, anti-misinformation rules must take care not to halt good-faith allegations of election fraud or misconduct as doing so would – ironically – undermine election integrity itself.

Anti-misinformation bills, such as Australia's, might also seek to address content that is "harmful to the economy" or undermines confidence in "financial markets". However, this is fraught for several

[26] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 ss 13(1)(c) and 13(2)(c).

reasons. Our knowledge in relation to topics, such as corporate governance and "environmental, social and governance" (ESG) investing evolves over time. Furthermore, ideological and normative value judgments might influence views on what is "harmful".

Corporate governance provides a clear example of how knowledge evolves over time. For example, there is an evolving literature in relation to "overconfident" CEOs: initial literature suggested that they were harmful per se.[27] This strand of knowledge would suggest that firms might *generally* avoid overconfident CEOs unless their other benefits outweighed their costs. Advocating for an overconfident CEO might have been seen as economically harmful. However, subsequent literature suggested that "overconfident" CEOs might help to facilitate innovation because they are less risk averse.[28] But, then it emerged that factors, such as corporate governance[29] and compensation,[30] moderate the impact of overconfident CEOs on performance. The net result is that our knowledge of what is "economically" beneficial, or harmful, might evolve over time.

Knowledge in relation to managerial entrenchment and takeovers provides another example. This has arisen in relation to "anti-takeover provisions" (ATPs, for short). Initial research suggested that ATPs *per se* worsened corporate governance by protecting managers from disciplinary takeovers.[31] However, subsequent research showed that they might yield some benefits in specific situations, such as in newly public companies where protection from takeovers might enable long-term strategic thinking.[32] Another example is in the field of ESG where some corporations have perpetuated incorrect "findings" in relation to board diversity.[33] However, these have been debunked.[34] Research evolves over time. If we assumed that the first-published research (in this case, by a corporation) were correct, then we would suppress the rebuttals. The Bill would thus suppress content that questions the status quo even if the status quo emerges subsequently to have been ill-informed.

Politicians also frequently espouse economically harmful viewpoints. Some of these would be "excluded" from the Bill's ambit, especially if they come from the government. However, there is no guarantee that opposition parties would be excluded from the Bill's ambit. This begs the normative question of whether voters should be able to hear, and vote upon, poor-quality economic proposals. An example includes the Greens' aforementioned proposal to force the RBA to lower interest rates. Thus, there is the question of whether voters should be able to hear from proponents of bad policy.

Contributing to, or causing, harm?

The next issue is whether anti-misinformation rules should relate to content that directly "causes" a harm or whether it is sufficient for it to merely "contribute to" the harm. Australia's anti-misinformation bill required only that the content "contribute to" the harm and did not require the content to "cause" the harm. However, "contributing to" a specified harm is a low threshold. For example, airing a debate between those with "harmful" opinions and those rebutting them might inadvertently contribute to that specified harm if the harmful party is more persuasive.

---

[27] U Malmendier and G Tate, "CEO Overconfidence and Corporate Investment" (2005) 60 *Journal of Finance* 2661; U Malmendier and G Tate, "Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction" (2008) 89 *Journal of Financial Economics* 20.

[28] A Galasso and TS Simcoe, "CEO Overconfidence and Innovation" (2011) 57 *Management Science* 1469; D Hirshleifer, A Low and SH Teoh, "Are Overconfident CEOs Better Innovators?" (2012) 67 *Journal of Finance* 1457.

[29] S Banerjee, M Humphery-Jenner and V Nanda, "Restraining Overconfident CEOs through Improved Governance: Evidence from the Sarbanes-Oxley Act" (2015) 28 *Review of Financial Studies* 2812.

[30] M Humphery-Jenner et al, "Executive Overconfidence and Compensation Structure" (2016) 119 *Journal of Financial Economics* 533.

[31] Masulis, Wang and Xie, n 15; Gompers, Ishii and Metrick, n 15; Bebchuk, Cohen and Ferrell, n 15.

[32] Humphery-Jenner, n 16; Johnson, Karpoff and Yi, n 16.

[33] McKinsey & Company, *Diversity Matters Even More: The Case for Holistic Impact* (Report, 2023) <https://www.mckinsey.com/featured-insights/diversity-and-inclusion/diversity-matters-even-more-the-case-for-holistic-impact>.

[34] A Edmans, "Discernment Matters Even More" (2024) <https://maycontainlies.com/discernment-matters-even-more/> .

The low threshold for proscribing content also fails to grapple with policy trade-offs: policies often have costs and benefits. A clear example is the aforementioned lockdowns during the Covid pandemic. Lockdowns had health benefits but economic costs. Thus, advocating for lockdowns could be seen as "contributing to" economic harm. Conversely, advocating against lockdowns could be seen as "contributing to" harms to preventative health. This applies similarly in the context of energy generation. There is a debate over the economic viability of renewable energy. Detractors argue that renewables are an expensive way to reduce carbon emissions. For example, some modelling purports to show that nuclear power is a more economically efficient way to generate "clean" energy.[35] Thus, they would argue that advocating for a "renewables focused" energy grid could cause economic harm. However, these costings have been disputed.[36] By contrast, renewables proponents would argue that arguing against renewables would "contribute to" environmental harm and potentially contribute to economic harm. They would also argue that the economic modelling underpinning the case for nuclear energy is erroneous; and thus, is "misinformation". The argument is then that it could "contribute to" a shift away from renewables, thereby causing environmental harm. This debate applies *mutatis mutandis* to electric vehicles. Policy debates can be contentious and involve trade-offs, suggesting that a nuanced approach to defining misinformation is required.

The net result is that policies are often trade-offs, especially in the short term. Furthermore, holding that content need only "contribute to" a specified harm sets a low bar for content to offend anti-misinformation rules. This suggests that anti-misinformation regulations should be sensitive to policy trade-offs in addition to considering a causal nexus between the content and the purported harm.

"Reasonably likely" to contribute?

The Australian anti-misinformation Bill also requires that the relevant content be "reasonably likely" to contribute to the relevant serious harm. The Bill does not require that the content actually contribute to the harm. Thus, the Bill is prospective and appears to aim to prevent hypothetical future harm. For example, if content is "reasonably likely" to incite a riot (the harm), the bill aims to stop the content before it does incite a riot. The issue is then what exactly it means for content to be "reasonably likely" to contribute to the harm.

The first possible construction of the term "reasonably likely" is that it imports a "reasonable person" standard into the exercise. Under this construction, the term "reasonably" refers to the grounds on which content is seen as likely to contribute to harm. Here, the content must still be "likely" to contribute to the harm. But, when assessing whether content is such, the media platform must consider whether a reasonable person would regard it as such. Further, the assessment must be based on reasonable grounds rather than merely speculative, hypothetical, or irrational grounds. This is akin to rules under the Information Privacy Principles, which forbid the disclosure of certain information to third parties unless (inter alia) the person is "reasonably likely" to be aware that such disclosure would occur.[37] In relation to this, the Federal Court indicated that the meaning of "reasonably likely" requires an inquiry "as to whether it is reasonable, as distinct from something that is irrational, absurd or ridiculous".[38] Similarly, the *Racial Discrimination Act 1975* (Cth) prohibits conduct that is "reasonably likely" to have certain effects on the subject group.[39] Here, the conduct must be likely to

[35] Frontier Economics, *Economic Analysis of Including Nuclear Power in the NEM* (Report, 2024) <https://www.frontier-economics.com.au/wp-content/uploads/2024/12/Report-2-Nuclear-power-analysis-Final-STC.pdf>.

[36] S Hamilton, "Economics of Coalition's Nuclear Modelling Are Worth Nothing", *Australian Financial Review* (online, 16 December 2024) <https://www.afr.com/policy/energy-and-climate/economics-of-coalition-s-nuclear-modelling-are-worth-nothing-20241214-p5kydg>.

[37] *Privacy Act 1988* (Cth), s 14, Information Privacy Principle 11.1(a).

[38] *AIT18 v Australian Information Commissioner* (2018) 267 FCR 93, [117].

[39] *Racial Discrimination Act 1975* (Cth), s 18C.

have the effect on a "reasonable" member of that subject group. [40]The qualifier "reasonably" requires that the assessment be with respect to what a "reasonable" person might perceive rather than what a specific person might subjectively believe.

This first construction has some issues, however. This is because it implies that if the content need only be "likely" to contribute to the harm, then the media platform would not need a rational reason for making such a judgment. Furthermore, the case law, and statutes, that relate to this first construction turned on whether it was sufficient for a person to subjectively believe a specific thing, or whether it is necessary for the belief to be on reasonable grounds or to be that of a reasonable person. However, in this case, there appears not to be a subjective value-judgment. Rather, the modifier "reasonably" appears to pertain to the degree of likelihood rather than the basis on which the likelihood is assessed. Further, a focus on the "reasonableness" of the assessment might not be the legislative intent in this context as it would enable media platforms to make arbitrary or capricious decisions about whether content is "likely" to contribute to harm but that they must make rational decisions about whether content is "reasonably likely" to contribute to harm. This would be an absurd construction as it is unlikely that legislators would ever countenance media platforms making irrational judgments of this nature. Rather, it is likely that media platforms would *always require* reasonable grounds for determining content contributes to the stated harm, even if the qualifier "reasonably" was not added to the term "likely".

The second construction is that the term "reasonably" in the phrase "reasonably likely" modifies the degree of likelihood of the content contributing to the harm. However, this construction is vexed as it is not clear whether "reasonably likely" is a higher or lower degree of confidence than is "likely". On the one hand, in everyday speech, "reasonably likely" is seen as less certain than "likely". Thus, there is some case law that "reasonably likely means more than merely possible but less than more probably than not".[41] Similarly, in the context of seeking redress for Child Sexual Abuse, a "reasonable likelihood" of having experienced abuse means "the chance of the person being eligible for redress is real, and is not fanciful or remote and is more than merely plausible".[42]

The net result of the drafting here is that the phrase "reasonably likely" is undefined and is open to alternative interpretations. This opacity creates uncertainty for media platforms. As indicated below, media platforms are incentivised to over-suppress content, rather than under-suppress it. Thus, there is a risk that media platforms will apply a low bar when determining what might be misinformation. Legislators must consider such issues when drafting legislation, lest the legislation become over-broad.

PROBLEM 2: WHO IS EXCLUDED FROM THE MISINFORMATION BILL?

Anti-misinformation rules should attempt to accurately identify, and address, misinformation if those rules are to exist. Thus, the rules should avoid wrongly deeming content from certain parties as inherently accurate as this would enable those parties to spread false and misleading content. Against this backdrop, it begs the question: Should the government be allowed to mislead citizens, spreading misinformation and disinformation? What about the media? If the media is allowed to spread misinformation, why not independent journalists and podcasters? Anti-misinformation legislation must grapple with these questions when determining whether, and whom, to exclude from its ambit. Australia's anti-misinformation bill provides a bad example, highlighting the issue of exempting powerful, entrenched interests.

Governments are tempted to exempt themselves from anti-misinformation legislation. Such exemptions help governments to avoid scrutiny and entrench themselves. Australia's anti-

[40] *Eatock v Bolt* (2011) 197 FCR 261, [251] (Bromberg J).

[41] See *Alvarez v Superior Court* (2007) 154 Cal App 4th 642,653-654 in the context of moving jurisdictions in order to obtain a fair trial.

[42] Department of Social Services, "What Does 'Reasonable Likelihood' Mean?", *Guides to Social Policy Law: National Redress Guide* (7 November 2022) <https://guides.dss.gov.au/national-redress-guide/3/2/1>.

misinformation bill is an example. It implicitly exempts government content from scrutiny and enables the government to exclude itself from any attempts to suppress misinformation. The exposure draft made this explicit in the definition of "excluded content".[43] The current version makes this implicit. This comes from several sections. For example, s 81 states that the relevant rules are not to be "exercised in such a way as to prevent the exercise of the powers, or the performance of the functions." Given that political communications are such a function, presumably state and territory governments are excluded. Secondly, 13(3)(d) requires social media companies to consider the "author" of information when designing their approach. Thirdly, s 13(4) empowers the minister to determine any other matter that digital services must consider, which empowers the Minister to exempt the government of the day from scrutiny but not opposition parties.

The issues with exempting the government, or politicians, from anti-misinformation laws are severalfold. First, such an exemption is logically inconsistent. The government has more resources than private citizens do. This means that the government is more able to afford fact checkers, and undertake fact-checking research, than are private individuals. Further, the government has more legal power and authority. This increases the potential damage inflicted by government misinformation as more people are prone to inherently believe and obey the government. Therefore, misinformation is costlier when it comes from the government, all else equal. This implies that the government should be held to a *higher* informational standard, not a *lower* standard.

The second issue is that of entrenchment and good governance. If a principal hires an agent to perform a task, that principal must be able to hold the agent to account. This requires that the principal can both identify any wrongdoing and take steps to remedy it. An example is in corporate governance. Managers have differing goals and incentives from shareholders.[44] Information asymmetry between the manager and the shareholders exacerbates this. Shareholders cannot observe all the manager's actions and face challenges determining ex ante whether the manager is truthful. This explains the need for shareholder class actions,[45] and SEC enforcement actions,[46] which discipline managers for lying. Managers are also often fired or otherwise disciplined after such wrongdoing. But, such enforcement actions are premised on shareholders' ability to determine whether the manager is truthful. Furthermore, if the manager is entrenched, shareholders cannot discipline the manager effectively. Thus, such entrenchment can destroy shareholder wealth[47] via shirking[48] or inefficient investments.[49] By parity of reasoning, voters must be able to effectively discipline politicians and the government. However, voters can do so only if they can properly evaluate the government and the government cannot freely deceive them. Thus, if government can spread misinformation, and social media platforms are incentivised to side with its narrative (see below) voters cannot discipline the government, which will lead to worse voter outcomes.

Excluding the news media is not sufficient to resolve this issue and can exacerbate problems. Australia's anti-misinformation Bill does this, for example.[50] However, this begs the normative question of why news media should be allowed to spread misinformation. It also raises the further issue that if the news media can promulgate incorrect information, then it will de facto be regarded as

[43] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023 (Cth), s 2 (definition of "excluded content").

[44] JS Ang, RA Cole and JW Lin, "Agency Costs and Ownership Structure" (2000) 55 *Journal of Finance* 81.

[45] S Banerjee et al, "Executive Overconfidence and Securities Class Actions" (2018) 53 *Journal of Financial and Quantitative Analysis* 2685.

[46] JM Karpoff and JR Lott, "The Reputational Penalty Firms Bear From Committing Criminal Fraud" (1993) 36 *Journal of Law and Economics* 757; JM Karpoff, DS Lee and GS Martin, "The Consequences to Managers for Financial Misrepresentation" (2008) 88 *Journal of Financial Economics* 193.

[47] Bebchuk, Cohen and Ferrell, n 15; Gompers, Ishii and Metrick, n 15.

[48] M Bertrand and S Mullainathan, "Enjoying the Quiet Life? Corporate Governance and Managerial Preferences" (2003) 111 *Journal of Political Economy* 1043.

[49] Masulis, Wang and Xie, n 17; Harford, Humphery-Jenner and Powell, n 15.

[50] See Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth) s 16(1)(b).

truthful. Therefore, any rebuttal would be regarded as "misinformation" even if the rebuttal itself is correct. This is especially a concern as the news media is often perceived as more authoritative. Thus, if the news media misleads the population, more people will believe it, increasing the harm that such misinformation would cause. This implies that the news media should especially be subject to anti-misinformation provisions if they were to exist.

Excluding news media from anti-misinformation rules is also based on the faulty assumption that the news media is inherently reliable. News organisations are subject to professional standards. However, it does not follow that they always follow them. An example is the reporting of Heston Russell, who successfully sued the ABC for defamation and was awarded $390 000.[51] Subsequently, allegations emerged that the ABC "altered" or "doctored" audio in their clips to make Heston Russell appear worse. This includes allegations that they edited the sound of gunshots in audio that they played.[52] Despite these concerns, ABC managing director David Anderson maintained that the reporting was in the public interest. Defamation litigation is not infrequent. However, the incidence of such litigation likely understates the magnitude of the problem given that defamation cases are expensive to launch; and thus, unavailable to most Australians. However, defamation is at least a remedy when the news media lies about an individual. If the news media lies about an event, and that lie becomes the *de facto* truth, then exempting the media from anti-misinformation rules exacerbates that entrenchment, removes accountability, and increases the harm that mistruth might do.

The net result of this situation is that anti-misinformation rules must contain appropriate exemptions. If anti-misinformation rules are to exist, exemptions should *not* apply to large, established organisations that have the resources to ensure that they are truthful and whose mistruths could create significant harm. To do otherwise would further entrench those established interests.

Problem 3: How do you enforce the rules?

Enforcement is often imperfect. Any attempt to classify content as "misinformation" or "disinformation" runs the risk of inaccurately classifying such content. Truthful or innocuous content might be wrongly classified as "misinformation" or "disinformation". Alternatively, "misinformation" or "disinformation" might be wrongly classified as truthful. The foregoing sections highlight the problems with how to define the words "misinformation" and "disinformation", and the issue with deeming some people's content inherently innocuous. However, there are also issues with how the rules are enforced and implemented.

Anti-misinformation rules require a way to enforce them. However, the enforcement mechanisms can have undesirable consequences. Enforcement can involve government agencies monitoring content. However, the optics of government directly removing or suppressing content would deter such an approach. Direct government involvement would also raise the risk of political censorship. Alternatively, it can involve deputising media platforms to police the content that appears on that platform. This latter option is less resource-intensive for the government. This is the approach that Australia's anti-misinformation rules take.

Australia's anti-misinformation rules focus on deputising media platforms to suppress misinformation and disinformation. Australia's anti-misinformation rules require media platforms to develop a code that outlines how they detect, and reduce, misinformation and disinformation.[53] But, as indicated, the meaning of the words "misinformation" and "disinformation" is broad, forcing media platforms to interpret and apply those words under their codes. The relevant legislation enumerates several example-factors that the anti-misinformation codes could, or should, consider in relation to

[51] *Russell v Australian Broadcasting Corporation (No 3)* [2023] FCA 1223, [509]; E Dudley and S Elsworth, "ABC Ordered to Pay Ex-Commando Heston Russell's Legal Fees" *The Australian* (online, 24 October 2023) <https://www.theaustralian.com.au/business/media/abc-ordered-to-pay-excommando-heston-russells-legal-fees/news-story/3cb9790d3c3d7aed1340a7e24da5ff5b> .

[52] S Elsworth, "ABC Boss Admits the Broadcaster Ignored Concerns about Altered Audio Involving the November Platoon" *The Australian* (online, 20 September 2024) <https://www.theaustralian.com.au/business/media/abc-boss-admits-the-broadcaster-ignored-concerns-about-altered-audio-involving-the-november-platoon/news-story/013b5689c2e5e0a377529774f7b4d1fc>.

[53] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth), s 48.

suppressing "misinformation".[54] The enumerated factors all pertain to suppressing "misinformation", but none pertain to avoiding suppressing innocuous content.

The regulator oversees the social media companies, their anti-misinformation codes, and the compliance therewith. If the regulator finds the code to be inadequate, it can take remedial actions to bring it to the regulator's standards.[55] Similarly, the regulator ensures compliance with the code to identify whether the media platform has suppressed misinformation. The regulator monitors and analyses whether the media platform has failed to suppress misinformation or disinformation.[56] The regulator is not tasked with analysing whether the media companies have suppressed innocuous content. Such false suppressions are neither reported nor analysed; and thus, go undetected. Thus, given that there are penalties for failing to suppress misinformation but not for suppressing innocuous content, the rules incentivise media platforms to "err on the side of caution" and risk suppressing too much content rather than too little.

Social media companies have struggled with how to address misinformation and disinformation, even aside from the definition thereof. There have been two overarching approaches: fact-checkers and "community notes". Neither approach is perfect and both approaches face challenges identifying "misinformation".

Fact checkers have challenges identifying "misinformation" and raise the risk of ideological bias. Australia's Voice Referendum highlights the challenge of relying on fact checking entities. One key example pertains to a debate over the length of the "Uluru Statement".[57] The "Uluru Statement" catalysed the Voice referendum. There was debate over precisely how long the Uluru Statement was and what it contained. This became important due to concerns about whether the Uluru Statement, which grounded the Voice, was misleading through omission and whether there were clues as to what the Voice body was intended to achieve. Some argued that what the Voice was intended to achieve would, or should, influence how people vote as it would suggest the consequences of that vote.

The debate over the statement's length fell into two major camps. On the one hand, the widely disseminated Uluru Statement was one page. On the other hand, several words in the statement required definition, such as "makarrata" or "agreement making" (i.e., "treaty making"). This led to investigative work, which revealed background documents underlying the one-page Uluru Statement.[58] Within these records was a longer document also entitled the "Uluru Statement from the Heart".[59] This longer document outlined that the Voice would precede, and originate, a treaty process. This treaty process was envisaged to involve (inter alia) "a financial settlement (such as seeking a percentage of GDP)" and "recognition of authority and customary law".[60] At one point, leading "Voice Architects" had argued that the Uluru Statement encompassed this longer document.[61] However, they subsequently asserted that it was only one page and did not include the documents pertaining to treaty-making.[62]

The notion of "misinformation" arose when a journalist argued that the Uluru Statement was more than merely one page, it revealed what the Voice would do, and that this was an argument against

[54] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth), s 44(3).

[55] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth), ss 50, 51.

[56] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth), s 52.

[57] Uluru Statement from the Heart, ″Uluru Statement from the Heart″ (26 May 2017) <https://ulurustatement.org/the-statement/view-the-statement/> .

[58] National Indigenous Australians Agency, "Record of Meetings: FOI-2223-016" (2023) <https://www.niaa.gov.au/sites/default/files/documents/foi-log/foi-2223-016.pdf>.

[59] National Indigenous Australians Agency, n 58, 87–112.

[60] National Indigenous Australians Agency, n 58, 105.

[61] For example, the assertion that "The Statement was drafted and overwhelmingly endorsed by the Convention's delegates. It is 15 pages long" in M Davis and P Anderson, *Our Voices from the Heart* (HarperCollins, 2023) 146.

[62] See for e.g., J Cross, "Megan Davis Slams 'Misinformation' on Breadth of Uluru Statement from the Heart", *National Indigenous Times* (online, 9 August 2023) <https://nit.com.au/09-08-2023/7150/megan-davis-slams-misinformation-on-breadth-of-uluru-statement-from-the-heart>.

voting for the Voice.[63] The ensuing backlash resulted in claims that the journalist was spreading misinformation in order to undermine the pro-Voice campaign. Subsequently, a fact-checker – the RMIT FactLab – purported to "fact check" the claims, arguing that indeed the Uluru Statement was just "one page".[64] The fact-check initially occurred on Meta's Facebook platform, which prompted Sky News to commence legal action against RMIT Fact Lab.[65] But, the fact check raised the issue of how to reconcile the finding with the authors' original comments that the Uluru Statement was more than one page, the existence of the other 26-page document titled "Uluru Statement from the Heart", and the fact that vague words require definition and meaning, which must derive from context. To this end, an alternative fact-checker – ABC's Media Watch – disputed the original fact check.[66] The net result is that identifying what is "misleading" is nontrivial, it is not clear that fact-checkers always decide correctly, and value judgments necessarily are involved. Furthermore, there are concerns that fact checkers might themselves be ideologically biased.

Meta – the owner of Facebook and Instagram – subsequently moved to drop fact checkers. As indicated, the original fact-check sparked allegations of bias. This resulted in Meta suspending its partnership with RMIT FactLab[67] for two months.[68] However, Meta subsequently announced that it would stop using third-party fact-checkers in the United States, arguing that "experts, like everyone else, have their own biases and perspectives. This showed up in the choices some made about what to fact check and how".[69] That is, fact-checkers might selectively fact check ideological opponents, and be more prone to treat claims they disagree with as "misinformation" even if they are merely disputed.

The "community notes" approach to fact checking has advantages and disadvantages and is itself imperfect. X (formerly Twitter) has focused on using "community notes" for fact-checking.[70] The "community notes" approach involves individual users posting a proposed "note" attached to content if they believe it to be misleading or false. The proposed notes are then rated by other users. Ideally, users with different ideological backgrounds would rate the proposed note. A "community note" would be rated as reliable if users with different rating histories, or backgrounds, agreed as such. In that case, the fact-checking community note would remain attached to the original content, alerting other readers to such.

The "community notes" model has several claimed advantages. First, "community notes" can arguably be designed to be less biased than relying on individual fact-checkers. This itself turns on notes on content only being accepted when appropriately rated by people of different backgrounds. This stands in contrast to the reliance on individual fact-checking organisations, which can become ideological

[63] P Credlin, "Peta Credlin Explains 26-Page Uluru Statement Yes Campaign Doesn't Want Aussies to See" *Sky News Australia* (3 August 2023) <https://www.youtube.com/watch?v=3qAWIzOk_0c>.

[64] RMIT Lookout, "The Uluru Statement from the Heart Is One Page Long, Not 26" (10 August 2023) <https://www.rmit.edu.au/news/factlab-meta/uluru-statement-from-the-heart-is-one-page> .

[65] J Hyland, "Sky News Australia Takes Legal Action against Facebook Fact Checker RMIT after They Censored Peta Credlin's Uluru Statement from the Heart Claim" *Daily Mail* (online, 11 September 2023) <https://www.dailymail.co.uk/news/article-12502913/Sky-News-legal-RMIT-FactLab.html>.

[66] P Barry, "Sky Hosts at War", *Media Watch* (ABC, 21 August 2023) <https://www.abc.net.au/mediawatch/episodes/credlin/102755280>.

[67] P Collins, "Facebook Ends Partnership with RMIT's FactLab after Receiving Complaints about Bias over The Voice" *Daily Mail* (online, 29 August 2023) <https://www.dailymail.co.uk/news/article-12455459/Facebook-ends-partnership-RMITs-FactLab-receiving-complaints-bias-Voice.html>.

[68] A Meade, "RMIT's Fact Check Reinstated by Facebook Two Months after Suspension over News Corp Voice Complaints" *The Guardian* (online, 7 November 2023) <https://www.theguardian.com/australia-news/2023/nov/08/rmit-factlab-fact-check-service-reinstated-facebook-suspension-voice-referendum> .

[69] J Kaplan, "More Speech and Fewer Mistakes" (Meta, 7 January 2025) <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/> accessed 8 January 2025.

[70] X, "Community Notes: A Collaborative Way to Add Helpful Context to Posts and Keep People Better Informed" (2025) <https://communitynotes.x.com/guide/en/about/introduction> .

silos.[71] Second, when implemented as such, community notes can increase broader trust in fact-checking,[72] which itself reduces the spread of false or misleading content by disabusing people of the original content's veracity. Third, a "community notes" approach can address a larger amount of content than can "expert" fact-checkers. This is because companies are inherently resource-constrained and thus have limits on how many fact-checkers they can hire. Thus, the fact-checkers must focus on a subset of content, which itself could be ideologically determined.

The "community notes" system has some disadvantages. First, community notes rely on the "general public" rather than "experts". In technical areas, non-experts might err or might "guess" whether content (or a note) is accurate. Second, community notes are vulnerable to a cascade effect or to organised attempts to discredit content for ideological reasons. The community notes systems can reduce this risk by requiring people of different backgrounds to have approved a "note" before the system accepts it. However, this depends on appropriately identifying different raters' backgrounds or biases. Third, community notes might be inaccurate because notes on content might reinforce prior biases. For example, an erroneous proposed note might trigger confirmation bias in someone who sees the proposed note to rate it. These factors mean that community notes are an imperfect way to evaluate content.

The net result is that enforcing anti-misinformation rules is challenging. Anti-misinformation rules confront the issue of how to identify, and suppress, offending content. Government can either directly suppress misinformation, creating the risk of government censorship or the appearance thereof. Alternatively, if they deputise media platforms to suppress misinformation, they run the risk of how to balance the desire to suppress misinformation with the need to avoid over-suppressing content. Furthermore, it raises the issue of how to identify such content, including whether to use fact-checkers or community crowdsourcing.

CONCLUSIONS

Addressing "misinformation" and "disinformation" is fraught. The problems with doing so are severalfold, and Australia has confronted them in drafting anti-misinformation rules. These issues mean that legislation to suppress "misinformation" and "disinformation" should be tightly written, and limited in scope, if such legislation is to exist at all. Further, the legislation should avoid per se deeming some parties' content to be innocuous, giving it the appearance of being truthful. Otherwise, such parties can themselves spread misinformation without being held accountable.

The first fundamental issue with anti-misinformation rules is how to define "misinformation" or "disinformation". Australia's anti-misinformation rules define "misinformation" as content that is "reasonably verifiable as false, misleading or deceptive" and that can cause or contribute to serious harms of various kinds, which itself is defined broadly. However, the issue is that all such terms require additional definition and context. This is commonplace in legislation. However, in the context of ascertaining "truth", what is believed to be true changes over time as assertions begin, evidence is gathered, and content-accuracy is improved. This is akin to how the scientific method operates. Thus, by suppressing content ex ante, such rules effectively deem the status quo to be inherently accurate, preventing firmly believed, but inherently false, narratives from being corrected.

The second major problem is in relation to excluding parties from anti-misinformation rules. Anti-misinformation rules might exclude groups, such as the government or media, from their ambit. Such exclusions might be out of self-interest, lobbying pressure, or a belief that such parties are inherently accurate and reliable. However, if the motivation is self-interest, then the exclusion is undesirable and not in the public interest. If the exclusion is based on a perception that the party is reliable, then that assumption must itself be tested – and well justified – and that party must be open to scrutiny and discipline through other avenues.

[71] D Walsh, "Bosses, Check Your Political Bias When Hiring" *Chicago Booth Review* (online, 23 January 2023) <https://www.chicagobooth.edu/review/bosses-check-your-political-bias-when-hiring>; KL Swigartet al, "Working While Liberal/Conservative: A Review of Political Ideology in Organizations" (2020) 46 *Journal of Management* 1063.

[72] CP Drolsbach, K Solovev and N Prollochs, "Community Notes Increase Trust in Fact-Checking on Social Media" (2024) 3 *PNAS Nexus* 217.

The third issue discussed in this article is enforcement. Given that the words "misinformation" and "disinformation" are broad, enforcement must be well calibrated. Thus, the enforcement mechanism must aim to avoid wrongly suppressing innocuous content. However, where – as in Australia – the rules task social media companies with suppressing content, they can result in excessive suppression. This is because the rules penalise companies for failing to suppress "misinformation" but not for erroneously suppressing innocuous content.

These issues add together to create a further fundamental problem: the rules implicitly undermine free speech. The broad definitions capture significant sets of content and deter challenges to the status quo. Exemptions help to entrench that status quo or powerful interests. Enforcement mechanisms can incentivise broad content suppression. Anti-misinformation rules thus pose a significant challenge to free speech.