THE UNIVERSITY OF SYDNEY

Submission to the 2025 Review of the Australian Code of Practice on Disinformation and Misinformation

Centre for AI, Trust and Governance

Dr Francesco Bailo

Professor Terry Flew

Dr Rob Nicholls

Associate Professor Daniel Gozman

3 November 2025

Contents

Executive Summary

This submission proposes a reconceptualisation of the Australian Code of Practice on Disinformation and Misinformation. We argue that the current approach conflates two distinct types of harm requiring different policy responses: individual harm from exposure to dangerous content, and collective harm from systemic degradation of information quality across the digital ecosystem. We propose that platforms should be held accountable for preventing individual harm through content moderation, while contributing to ecosystem-wide monitoring of information disorder through a novel persona-based measurement system that protects epistemic rights.

1. Problem Definition: Two Distinct Harms

Disinformation and misinformation cause harm at two fundamentally different levels, each requiring distinct policy responses. Before examining these harms, we propose a foundational definition:

Misinformation: All content that, when exposed to, is hazardous for the enjoyment of epistemic rights. This approach is independent of both the intent or coordination behind content distribution (which is often difficult or impossible to assess) and the veracity of the content itself (which may be particularly difficult to ascertain during emerging or developing crises). For example, presenting factually-correct information without proper context or explanation can restrict epistemic rights by creating a misleading impression of scientific consensus or empirical reality. Cherry-picking of technical information—a tactic documented in both anti-climate change and anti-vaccination movements—exemplifies this harm. As Cook (2019) observes, cherry-picking involves selectively choosing data that leads to a conclusion different from the conclusion arising from all available data. By highlighting only data points that support a preferred narrative while ignoring contradictory evidence, cherry-picking distorts users'

understanding and impairs their ability to make informed decisions, thereby compromising their epistemic rights regardless of whether individual data points are factually accurate. Because transparency is an expected requirement for scientific studies—especially in highly contested domains—and critically because of the widespread accessibility of advanced technologies such as Generative AI based on large language models, this practice is expected to increase substantially (Asgary & Nayebi, 2025).

1.1 Individual-Level Harm

At the individual level, people may be influenced to adopt dangerous behaviors or make harmful decisions affecting themselves or their immediate social contacts. Examples include promotion of unproven health treatments, fraudulent financial schemes, or dangerous products. An example is the promotion of hydroxychloroquine as a COVID-19 treatment despite a randomised control study finding it ineffective in improving clinical outcomes (Haupt et al., 2021). This harm is direct, measurable, and amenable to content-level interventions.

Platform Responsibility: Platforms are fundamentally responsible for reducing individually harmful content distributed on their services. This includes both disinformation and misinformation where actors are coordinating dissemination for financial or political gain.

1.2 Societal-Level Harm

At the societal level, widespread exposure to false or misleading information erodes trust in both private institutions and public institutions, ultimately undermining democratic processes. This harm manifests as information disorder across the entire informational environment.

Shared Responsibility: This is a complex systemic problem requiring ecosystem-wide responses. No single platform controls the majority of public discourse. Users are active across multiple platforms (Thorson & Wells, 2016) with varying content moderation standards. Intermediaries move content across platforms, news media amplify narratives, and institutions shape the information

environment. Platforms should contribute data to monitor the problem's scale but cannot bear full responsibility alone.

2. Key Conceptual Framework

2.1 Epistemic Rights

Drawing on Nieminen (2024), we define epistemic rights as the individual right to:

- Have access to sufficient information to make informed decisions, or conversely, to accept the impossibility of making informed decisions when knowledge is unavailable
- Have sufficient competence in navigating information systems and reading relevant information about any issue

2.2 Information Disorder

Information disorder is a property of an information ecosystem measuring how difficult it is for average users to make sense of the information served to them. The more pronounced the disorder, the harder it becomes to exercise epistemic rights. Information disorder represents risk: the probability of encountering information that makes understanding an issue more difficult. Across society, different individuals have different susceptibilities to this risk.

3. Critical Limitations of Current Approaches

3.1 Focus on Volume Rather Than Epistemic Impact

A fundamental limitation of current approaches, including the existing ACPDM framework, is the focus on measuring the volume of misinformation and disinformation rather than assessing the real negative impact for society of this content's diffusion. The Code's emphasis on content removal, labelling, and deranking—whilst important for addressing individual harm—fails to capture the systemic degradation of information quality that restricts the enjoyment of epistemic rights. The compression of epistemic rights occurs not simply through exposure to individual pieces of false content, but through the cumulative effect of information disorder on users' capacity to make informed decisions at any given point in time. By focusing on counting instances of misinformation rather than measuring the affordances of the information ecosystem for informed decision-making, current transparency reporting mechanisms miss the more significant collective harm.

3.2 Fact-Checking Limitations

Fact-checking can only address a small fraction of problematic content and faces inherent limitations:

- Can only be conducted on a small fraction of content, usually hours or weeks after publication and mostly after peak attention
- Cannot be conducted in the presence of epistemic gaps where expert systems produce conflicting information
- Is increasingly perceived as politicised with potential for backlash and counterproductive effects

– Should be limited to interventions addressing personal harm rather than collective harm

3.3 Time Sensitivity and Crisis Contexts

Information quality observably deteriorates during times of crisis when “epistemic gaps or 'data voids'” (Golebiewski & Boyd, 2019) emerge and expand. Epistemic gaps refer to the widening distance between what is confidently known based on reliable evidence and what remains uncertain or unverified. During crises, these gaps grow rapidly as events unfold faster than verification processes can keep pace, creating information voids that are often filled with speculation, rumours, and misinformation (Bailo et al., 2024). This deterioration in information quality compounds the challenges of informed decision-making precisely when accurate information is most critical.

3.4 Context-Specific Vulnerabilities

Although information disorder is fundamentally a collective problem, it is experienced individually. As with every risk, the likelihood of negative effects is conditional on the characteristics and demographics of persons exposed, with overall risk compounded by interactions among hazardous content.

4. Proposed Solutions

4.1 Differentiated Intervention Approaches

Personal Harm: The current identification and containment approach through deranking, labeling, or removal is appropriate for minimising instances of individual harm where risk assessment is more straightforward.

Collective Harm: More challenging is addressing collective harm from aggregate degradation of information quality. The Discussion Paper acknowledges this challenge, noting that identifying quantitative metrics for useful platform comparisons has proven elusive: “We have as yet been unable

to identify a set of quantitative metrics that allow for useful comparisons of the diverse platforms that have made commitments under the Code" (p. 10).

This is where we propose industry should make its primary contribution.

4.2 Persona-Based Ecosystem Monitoring System

The Australian digital industry should contribute to establishing a monitoring and data reporting system providing real-time understanding of information quality throughout the Australian media ecosystem for average users and specific vulnerable profiles.

Key Principle: Information ecosystem quality should be measured in terms of users' capacity to make reasonable, timely, and informed decisions at any point in time. This represents an affordance of the information ecosystem: the contextual relationship between users and platforms. The fundamental question is whether the information ecosystem affords full protection of individual epistemic rights.

Implementation: DIGI should collaborate with platforms to define:

Personas of Interest:

For example,

- Average Australian, but also
- Female teenager (13-17)
- Male 30-45, no university degree
- Elderly (65+)

Topics of Interest:

For example,

- Deliberate exposure to ongoing election campaigns
- Deliberate exposure to health/medical information
- Deliberate exposure to financial products/investment schemes

Platforms should provide data about content these personas encounter through two distinct exposure modes: the *incidental* mode, where users stumble upon politically-relevant information while browsing for non-political content, and the *intentional* mode, where users actively seek political information (Nanz et al., 2022). An important precedent exists in the Social Science One partnership between academics and Facebook (King & Persily, 2020). DIGI should collaborate to define representative media diets for these personas. Collaborating platforms should provide likely content exposure based on platform likelihood estimates, with private content removed or deidentified. This content should be made accessible in near real-time to researchers across industry, academia, non-profit organisations, and journalism.

To complement the ecosystem perspective, Australia could also draw lessons from international models of API-based regulation. International experience demonstrates that carefully structured, tiered API access under regulatory oversight can strengthen transparency competition and accountability across digital ecosystems. Initiatives such as the UK's Open Banking framework, the EU's PSD2 directive, and the US 21st Century Cures Act on health data illustrate how regulated API access can reduce information asymmetries and improve systemic trust. Similarly, Australia's Consumer Data Right (CDR) mandates structured access in sectors including banking, energy, and telecommunications to promote competition and empower consumers.

These policy models reconfigure the boundaries of control between platform owners and citizens, enabling greater transparency, innovation, and accountability. Regulated access for researchers, journalists, and independent watchdogs would enhance visibility into the origins and spread of harmful content, particularly during elections or crises, while allowing independent assessment of platforms' self-regulatory practices. Such gateways could also provide citizens with real-time insight into how their personal data are used and shared, reinforcing trust in information systems. However, these approaches must be accompanied by robust privacy, security, and misuse safeguards to prevent unintended harms. By framing gateway access and oversight as matters of epistemic rights, these mechanisms would not only improve transparency but also strengthen the resilience of Australia's information ecosystem

5. Responses to Specific Review Questions

5.1 Should the scope of the ACPDM be reconsidered?

We recommend maintaining coverage of both disinformation and misinformation but implementing the differentiated approach outlined above. Disinformation and misinformation are intertwined in practice, and restricting scope would limit platforms' ability to address personal harm. However, the Code should explicitly distinguish between interventions addressing individual versus collective (i.e. social) harm.

5.2 Transparency Reporting

Rather than pursuing common quantitative metrics for content removal, transparency reporting should focus on:

- Platform contributions to the persona-based monitoring system
- Changes to policies and interventions in response to evolving environmental conditions
- Risk assessment processes for identifying individually harmful content
- Data quality and accessibility for independent researchers, including access to granular data via APIs that enables independent qualitative and quantitative analysis

5.3 Ecosystem Approach

We strongly support an ecosystem approach. The Code should explicitly acknowledge that platforms are but one actor in a complex system. The persona-based monitoring system would provide shared infrastructure for understanding ecosystem health, enabling coordinated responses from

government, educators, news media, and civil society alongside platform interventions.

We strongly support the strengthening of laws and regulations governing truth in political advertising and misleading advertising, while noting that both are outside of the scope of ACPDM.

In particular, political advertising laws need to be substantially updated to address the problem of increasing use of AI-generated deepfakes in election campaigns. The Australian Electoral Commission has been actively monitoring such activities and offering public advice on how to spot fake campaign materials, but legislative change is required to address the fast-changing technological landscape and the relative decline of traditional media channels as primary sources of political information.

There are several consumer protections that flow from the Australian Consumer Law (Schedule 2 to the Competition and Consumer Act 2010 (Cth)). Essentially, these relate to the prohibition of misleading or deceptive conduct or misrepresentation in trade or commerce. However, it is essential that the ecosystem minimises the harms that flow from “dark patterns” and other unfair trading practices.

The ACCC has identified several unfair trading practices that cause, or are likely to cause, consumer harm. The ACCC takes the view that these are not appropriately covered by current Australian Consumer Law protections. Examples of these practices include:

- (a) distortion or manipulation of consumer choice without necessarily being misleading or deceptive (for example, by creating an undue sense of urgency or scarcity);
- (b) making it difficult for consumers to cancel subscription services or leading to automatic rollovers from free trials;
- (c) pricing-related practices including drip pricing and dynamic pricing;
- (d) unreasonable post-sale practices that impose barriers to accessing customer support; and

- (e) utilising 'dark patterns' (deceptive design elements in user interfaces) that manipulate consumer decisions. These often exploit behavioural biases or information asymmetry.

Many of these unfair practices have a similar approach to the distribution of disordered information. A commitment from DIGI members to eliminate unfair trading practices in their ecosystems would assist in the reduction of misinformation and limit the introduction of disinformation into these ecosystems.

We support the recommendations of the Senate Select Committee on Adopting Artificial Intelligence that pertain to the impact of AI on democracy (The Senate, 2024).

6. Conclusion

The challenge of addressing misinformation and disinformation requires moving beyond content-level interventions toward ecosystem-wide measurement and protection of epistemic rights. By distinguishing individual from collective harm and implementing persona-based monitoring, the ACPDM can provide a foundation for evidence-based policy while respecting the legitimate concerns about freedom of expression that emerged during debate over the withdrawn legislation.

References

- Asgary, A., & Nayebi, M. (2025, August 19). *AI-generated misinformation can create confusion and hinder responses during emergencies*. The Conversation. https://doi.org/10.64628/AAM.a9yd5fqhh
- Bailo, F., Johns, A., & Rizoiu, M.-A. (2024). Riding information crises: The performance of far-right Twitter users in Australia during the 2019–2020 bushfires and the COVID-19 pandemic. *Information, Communication & Society*, *27*(2), 278–296. https://doi.org/10.1080/1369118X.2023.2205479

- Cook, J. (2019). Understanding and Countering Misinformation About Climate Change. In I. E. Chiluwa & S. A. Samoilenko (Eds), *Handbook of Research on Deception, Fake News, and Misinformation Online:* IGI Global. https://doi.org/10.4018/978-1-5225-8535-0
- Golebiewski, M., & Boyd, D. (2019). *Data voids: Where missing data can easily be exploited*.
- Haupt, M. R., Li, J., & Mackey, T. K. (2021). Identifying and characterizing scientific authority-related misinformation discourse about hydroxychloroquine on twitter using unsupervised machine learning. *Big Data & Society*, *8*(1), 20539517211013843. https://doi.org/10.1177/20539517211013843
- King, G., & Persily, N. (2020). A New Model for Industry–Academic Partnerships. *PS: Political Science & Politics*, *53*(4), 703–709. https://doi.org/10.1017/S1049096519001021
- Nanz, A., Heiss, R., & Matthes, J. (2022). Antecedents of intentional and incidental exposure modes on social media and consequences for political participation: A panel study. *Acta Politica*, *57*(2), 235–253. https://doi.org/10.1057/s41269-020-00182-4
- Nieminen, H. (2024). Why we need epistemic rights. In M. Aslama Horowitz, H. Nieminen, K. Lehtisaari, & A. D'Arma (Eds), *Epistemic rights in the era of digital disruption* (pp. 11–28). Springer International Publishing. https://doi.org/10.1007/978-3-031-45976-4_2
- The Senate. (2024). *Select Committee on Adopting Artificial Intelligence (AI)* [Final Report]. Parliament of Australia.
- Thorson, K., & Wells, C. (2016). Curated Flows: A Framework for Mapping Media Exposure in the Digital Age. *Communication Theory*, *26*(3), 309–328. https://doi.org/10.1111/comt.12087

Contact

Centre for AI, Trust and Governance

Francesco Bailo

+61 (0)2 8627 6895

francesco.bailo@sydney.edu.au

sydney.edu.au

CRICOS 00026A

THE UNIVERSITY OF SYDNEY