Title: Participant classification framework and category level confidence appendix for "Monitoring misinformation, building discernment: A crowdsourcing study during the 2025 Australian federal election"
Authors: Ariel Kruger (1), Piers Howe (2), Morgan Saletta (3)
Date: August 27th, 2026
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix A: Participant classification framework and category level confidence

Before assigning a formal classification, participants were instructed to apply an initial evaluative screen referred to as the "sniff test." This step was designed to prompt reflective judgement and orient participants toward core norms of ethical democratic discourse. The sniff test did not generate a recorded variable; rather, it functioned as a heuristic guide to determine whether a post appeared to fall afoul of basic standards of truthful and constructive public communication.

The sniff test asked participants to consider whether the content violated any of the following norms of ethical democratic discourse:

- **Truthfulness:** Does the content appear truthful, or does it make claims that seem clearly or verifiably false, exaggerated, or unsupported?
- **Quantity:** Does the content provide sufficient relevant information, or does it appear vague, incomplete, or selectively detailed in a way that obscures key facts?
- **Relevance:** Does the content address the issue it claims to address, or does it distract, deflect, or misdirect?
- **Clarity:** Is the message clear and straightforward, or is it confusing, ambiguous, or unnecessarily complex?
- **Authenticity:** Does the contribution appear genuine and contextually appropriate (e.g., originating from a legitimate participant in Australian political discourse), or does it raise concerns about impersonation or deceptive identity?
- **Helpfulness:** Does the content appear intended to contribute constructively to democratic discourse, or does it seem designed to disrupt, inflame, or derail discussion?

If a post was judged to fall afoul of one or more of these norms, participants were required to assign it a primary classification category from a predefined framework of 17 misinformation types (with the option to select a secondary category where relevant). The framework was designed to distinguish among different mechanisms by which content may mislead, rather than simply reducing posts to a binary true/false judgement. The framework was developed pragmatically rather than derived from a single theoretical model, with the priority being categories that non-expert participants could readily understand and apply. We began from categories common to existing public-facing typologies of misinformation (e.g., Bounegru et al., 2018; Wardle & Derakhshan, 2017), producing an initial list of overlapping types. Three researchers then independently coded a development set of 24 social media posts using this list, comparing classifications and resolving disagreements through discussion. Where posts were not adequately captured by the initial categories, additional categories were added inductively to reflect the forms of misinformation present in the material. This process produced the final framework of 17 categories.

***Table A1**. Misinformation classification categories and descriptions.*

Category	Description
False content	Factually incorrect information, either wholly or partially false.
Out of context	Accurate or partly accurate information that misleads due to missing or distorted context.
Unproven/conspiratorial	Claims made without evidence (or using only anecdotal evidence), explanation, or conspiracies, e.g., baseless accusations or attributing events to secretive or deceptive forces without credible evidence.
Disconnection	Misleading headlines, visuals, or captions that do not accurately represent the content.
Altered media	Deepfakes, images, videos, or audio that have been altered, edited, or AI-generated to mislead.
Rage bait	Content designed to provoke strong emotional reactions for engagement—such as anger, outrage, or division—often by using inflammatory language, misleading framing, or exaggeration.
Impersonation	Misleading use of names, images, or branding to mimic real people, organizations, or media.
Bogus authority	Content that falsely claims or distorts authority, credentials, or expertise to mislead audiences. E.g., pretending to be an expert, or claiming expertise in an area they don't have.
Objectivity masquerade	Egregious examples of unlabeled partisan opinion, editorial, or analysis masquerading as factual, fair, or objective news, misleading audiences about its neutrality. Most likely to be from fringe media outlets.
Political defamation	False or defamatory claims intended to harm the reputation of political candidates, parties, or election officials.
Misunderstood satire	Satirical or parody content mistaken for real news, often due to unclear labelling or users spreading posts from parody accounts as factual reports.
Cherry-picking	Content that selectively highlights specific facts, quotes, or data while ignoring relevant context or contradictory information.
Unlabeled ads	Political ads lacking proper authorization or disclosure, potentially misleading viewers.
Political scams	Fraudulent schemes that exploit political events, candidates, or causes to deceive individuals into giving money.
Out of scope/not a problem	Not related to the Australian election or politics generally OR passes the sniff test.
True/partly true	The content is partly or wholly accurate.
Other	Fails the sniff test, but doesn't fit any of the categories.

Category level confidence

Across assessments, raters recorded their confidence (0–10) in each classification. Differences in mean confidence across categories were tested using a linear mixed-effects model, with classification category as a fixed effect and a random intercept for rater to account for the uneven number of assessments contributed by each participant (N = 3,037 assessments). Category significantly predicted confidence, $\chi^2(16) = 144.44$, $p < .001$, and a non-parametric Kruskal–Wallis test returned the same conclusion, $\chi^2(16) = 116.89$, $p < .001$. The test indicates that confidence varied across categories but does not identify which categories differed from one another; pairwise comparisons were not conducted, given the number of categories and the small number of assessments in several of them.

Table A2. *Mean confidence rating across categories.*

Category	**Mean confidence**	***N***
Political scams	8.36	11
False content	7.6	664
Misunderstood satire	7.56	39
Altered media	7.5	147
Disconnection	7.42	139
Unproven/Conspiratorial	7.38	362
Political defamation	7.3	92
Out of context	7.23	477
Impersonation	6.94	77
True/partly true	6.86	158
Out of scope/not a problem	6.84	321
Objectivity masquerade	6.81	53
Rage bait	6.71	260
Cherry-picking	6.68	174
Unlabeled ads	6.58	12
Bogus authority	5.14	7
Other	4.36	45