28 June 2024

Statutory Review of the Online Safety Act 2021

Submissions were prepared by Rita Jabri Markwell on behalf of the AMAN Foundation Ltd and the Australian Muslim Advocacy Network (AMAN) with support from

- Professor Nicole L Asquith, University of Tasmania and Convenor of the Australian Hate Crime Network,
- The Human Rights Law Centre,
- Jewish Council of Australia and
- Alliance Against Islamophobia.

Contents

1. ACKNOWLEDGEMENT

We acknowledge the lands of the Jagera, Toorbul, Wurundjeri, Bunurong, Gadigal, Ngunnawal, Darug, and Wadawurrung people, on which we work and live. We pay our respect to Elders of those lands, both past and present. This land always was and always will be Aboriginal and Torres Strait Islander land because sovereignty has never been ceded. We recognise the role of the colonial legal system in establishing, entrenching, and continuing the oppression and injustice experienced by First Nations peoples. We have a responsibility to work in solidarity with Aboriginal and Torres Strait Islander people to undo this.

2. INTRODUCTION

The **AMAN Foundation Ltd (the Foundation)** works to prevent the harms of systemic racism, online hatred and Islamophobia through policy engagement and law reform.

The **Australian Muslim Advocacy Network Ltd (AMAN)** created the Foundation for this harm prevention work.

AMAN has brought a legal complaint against Facebook/Meta under the Racial Discrimination Act 1975 (Cth) and a complaint against Twitter/X under the Queensland Anti-Discrimination Act 1991 (Qld).

AMAN works with a range of civil society involved in considering online harms, including Reset Australia, Purpose, and the Human Rights Law Centre. AMAN also works collaboratively with a range of anti-racism stakeholders, including the Islamophobia Register Australia, the Jewish Council of Australia, the Alliance Against Islamophobia, and individual lawyers and legal scholars from the First Nations community.

Contributors and Reviewers

Nicole L. Asquith is a Professor of Policing in the School of Social Sciences in the College of Arts, Law, and Education. Nicole has worked with and for policing services for over 25 years, primarily in relation to vulnerable victims. Before returning to the University of Tasmania, Nicole was the Associate Professor of Policing and Criminal Justice at Western Sydney University, and Senior Lecturer at Deakin University. In addition to her academic roles at UTas, Nicole is the Convenor of the Australian Hate Crime Network, and has published widely on most forms of hate crime and targeted violence, and contributed to policy and practice development within

and outside policing, including advising the Special Commission of Inquiry into LGBTQ Hate Crimes.

Human Rights Law Centre uses strategic legal action, policy solutions and advocacy to support people and communities in eliminating inequality and injustice and building a fairer, more compassionate Australia.

The **Jewish Council of Australia** is a coalition of Jewish academics, lawyers, writers and experts on antisemitism and racism.

Alliance Against Islamophobia promotes policies and practices that support the rights and well-being of Muslim communities. This includes advocating for the implementation of anti-discrimination legislation and engaging in dialogue with government, media, and other key stakeholders to promote a more inclusive and equitable society.

3. PRINCIPLES

a) A proactive and resilient architecture must

i. adopt an 'atrocity-prevention' approach focused on maintaining collective social barriers to dehumanisation. This will help shift the burden further upstream to digital platforms and away from communities most affected by downstream carceral, securitised and heavy-handed policing approaches.

ii. capture content that socialises people towards violence yet limits the aperture of regulator pro-active intervention to serial or systematic vectors of hate that lower an audience's barriers to violence.

iii. Ensure that dehumanisation groups based on protected characteristics are treated as a public harm rather than a private problem.

iv. Aim for a public information environment that supports diversity of opinion, veracity and accuracy of information is vital to Australia's obligations under various international instruments, including the ICCPR (freedom of expression, freedom of opinion, the right to non-discrimination, no advocacy of hatred), IESCR (the right to health). Preventing and moderating the advocacy of hatred enables greater freedom of expression by groups targeted by hatred. It also supports their fulfilment of the right to health by reducing exposure to a social atmosphere that denies their human qualities.

v. Ensure the overall design of the Act supports immediate powers for e-Safety to proactively address well-defined harms effectively and, secondly, to handle community complaints more effectively. **Figure 1 below** outlines the underlying rationale for this approach.

vi. Ensure definitions for online hatred against groups:

1. encompass hateful material, whether communicated through speech or words; the curation or packaging of information; images; and insignia.

2. are universally applicable and resilient to cyclical changes in targeted groups.

3. Are capable of securing public support and consensus about what constitutes hatred across different contexts, and being applied evenly.

4. Methodically link to one of the most dangerous forms of hatred – hatred that positions a group outside the human family, making them an easier and more deserving target for violence.

5. Provide a framework for education that supports critical thinking by applying logic and universal principles that people would like to see applied to them.

6. Provide clear guidance that opinions about countries, nation-states, governments or militaries do not constitute hate speech or vilification. Individuals or groups raising such examples should be referred to human rights-based judicial complaint processes to allow nuanced assessment.

b) Process matters.

i. E-Safety must adopt an approach grounded in multistakeholderism rather than individually and separately consulting each community group.

ii. E-Safety must not platform groups that engage in public acts of racism or racist nationalism in these consultations. Having clear anti-racism principles will assist e-Safety in governing its approach to consultation.

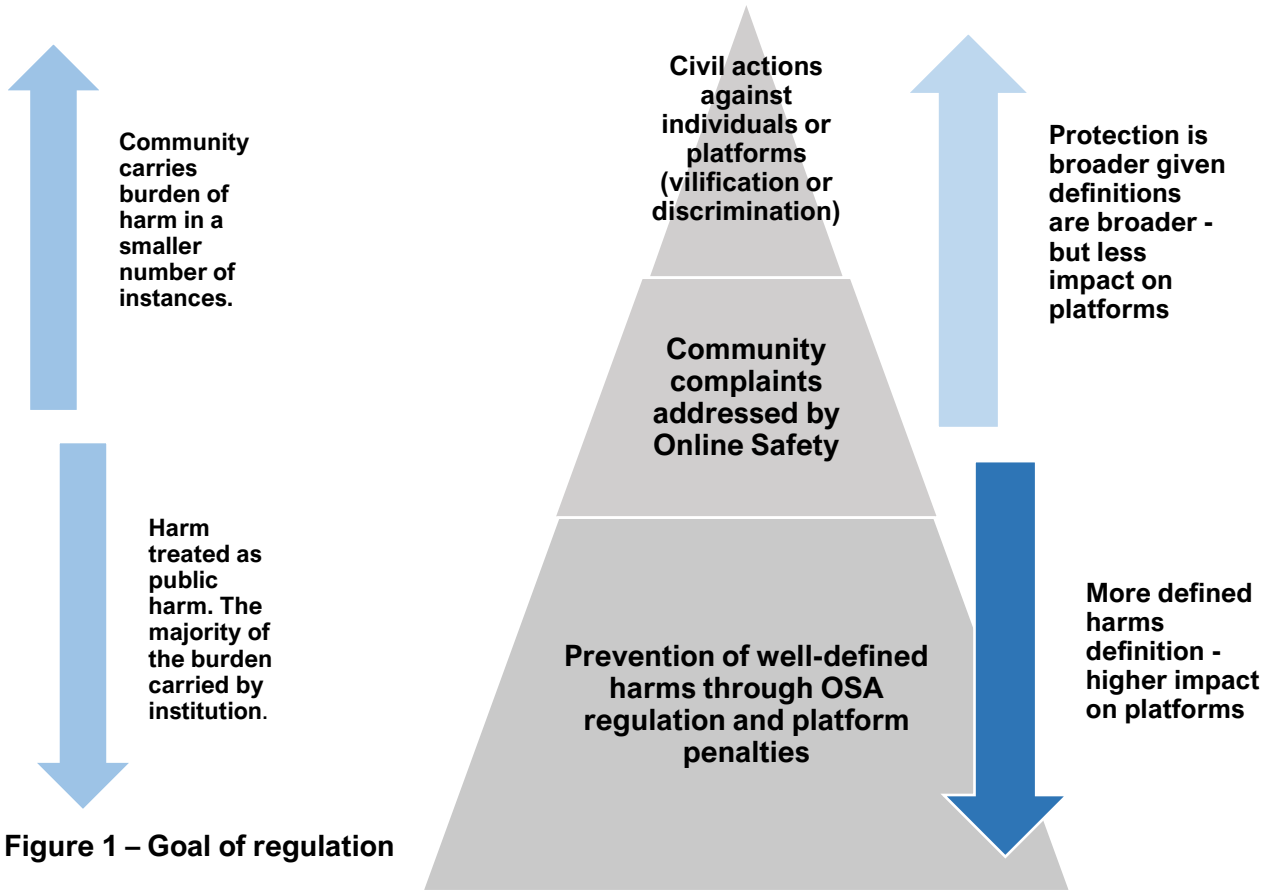

Figure 1 – Goal of regulation

4. RECOMMENDATIONS

4.1 Improve governance arrangements and capability by

a) introducing public reporting by e-Safety about the representativeness of its staff in relation to the natural diversity of Australia's population;

b) expanding the mechanisms for monitoring and assessment to include researchers and civil society, who are more equipped to identify emerging trends and patterns in misinformation and disinformation. As a starting point, consider Article 40 of the European Digital Services Act and the 'crowdtangle provision' supporting immediate access to aggregations of public data.

4.2 Amend the *Online Safety Act 2021* (Cth) to prohibit the serial or systematic publication of dehumanising material.

a) The e-Safety Commissioner consults with Australian communities on definitions of dehumanising material about protected groups, with a view to including it as a particularly visible, egregious and harmful form of hate speech in the online content scheme. AMAN provides proposed wording as a starting point for discussion in **Schedule 1**. This means that the e-Safety Commissioner would have notice and takedown powers in relation to this content and the power to impose proportionate penalties on serial or systematic actors and the platforms that enable them. There would be a need for both proactive monitoring/action by the Commissioner and a mechanism for responding to complaints. Transparency reporting requirements on platforms would also apply.

4.3 By a further miscellaneous amendments bill, clarify that

a) section 18C of the *Racial Discrimination Act 1975* (Cth) has extraterritorial application to foreign-based digital platforms;

b) discrimination provisions of various federal discrimination laws have extraterritorial application to foreign-based digital platforms; and

c) relevant entities can bring *discrimination* complaints on behalf of groups or communities based on protected characteristics.

4.4 Require Social Media Companies to establish anti-racism or anti-dehumanisation units ("the unit"). The e-Safety Commissioner should publish terms of reference for the compliance unit ("the unit") that includes the following:

a) The unit will operate in Australia with Australian staff;

b) The purpose of the unit is to demonstrate the social media company's commitment to being antiracist by maintaining compliance with section 18C of the Racial Discrimination Act 1975 (Cth) and other vilification laws at the state and territory level on its platform for content viewed by Australian audiences.

c) The unit will proactively identify, assess, de-monetise and deplatform actors that use the platform to engage in dehumanisation, systematically or serially, through text, imagery or the curation and packaging of stories about a group on the basis of **a protected characteristic.** In that regard, the social media companies will review:

i. the material posted on the page or group over six months since the date of the last post or for the time that the page or group has been open if less than six months ("the Posts");

ii. the content of the Posts, including text, files, video and images;

iii. the imagery and headlines that appear in the 'preview box' accompanying the post of a third-party link;

iv. the content hosted at third-party sites referred to by third-party links in the Posts; and

v. the comments on the Posts;

d) Criticism or even hatred of nation-states, governments or militaries will continue not to constitute hate speech for the purposes of content moderation;

e) The unit publishes a yearly transparency report that provides disaggregated data and qualitative information about the contraventions it identifies and takes action on. Specifically, this report will provide a breakdown by the protected group targeted by racial hatred ("type of racial hatred") and, within such breakdown, a further breakdown of what type of moderation action was taken;

f) The unit must notify users about the moderation of their content that the company finds to contravene Australian racial hatred and vilification laws. This notification should outline the type of action taken and the reason for the action;

g) The unit will notify the Australian public on its website when any law enforcement or Government requests that the company moderate content to maintain compliance with section 18C. That notification will include the date of the request, the source of the request, the type of racial hatred covered by the request, and

h) Per Australian privacy principles, the unit will grant independent researchers access to data to conduct methodological and evaluative reviews of its work and transparency reports.

4.5 Do not allow exemptions for 'professional news content' in relation to online hatred against groups.

a) If an exemption is allowed, ensure the definition is strong enough, unlike that proposed in the News Bargaining Code of Exposure draft of the Misinformation and Disinformation Bill.

b) It is in the public interest for this Bill to not allow well-resourced and far-reaching news outlets to continue inciting hatred online and failing to moderate their comment threads. At the very least, the Bill must increase their requirements for transparency and accountability to benefit from that exemption.

c) AMAN recommends that the Australian Government work with Australian researchers, anti-racist civil society and the Global Disinformation Index to formulate these requirements. AMAN provides proposed wording as a starting point for discussion in **Schedule 2**.

4.6 Improve the function of existing cyberabuse and cyberbullying provisions in situations involving volumetric attacks on an individual's protected attributes.

5. DISCUSSION

a) We seek to ensure that the responsibility for monitoring and acting in relation to actors that engage in serial or systematic dehumanisation on social media platforms be owned and discharged by the platforms, rather than placing that burden on targeted communities, including their community members and organisations.

b) We have identified technology-based experts who will point to the ease with which the platforms can identify actors engaged in serial or systematic racial dehumanisation of groups based on a protected characteristic and take preventative steps (which they have failed to do).

c) In the 2019 federal election, approximately 12 fringe parties were running with a discriminatory anti-Muslim policy – this is the most significant number of groups we have recorded. There was an open license to dehumanise and denigrate Muslims as part of their online activity to recruit members and gather votes.

d) In the 2022 federal election, there was a substantial contraction in fringe parties running explicit anti-Islam policies. This reflected an overall re-orientation of far-right groups toward electoral misinformation (originally, US Politics based), dehumanisation based on gender diversity (especially cissexism/transphobia), and medical misinformation (Covid vaccines and the narrative that Covid was overstated/a hoax), and climate misinformation (portraying climate science as part of a reset global conspiracy).

e) However, it is predicted that the 2025 federal election will invigorate race-based campaigns, especially if we also see a transition to Donald Trump's Presidency in the United States.

5.2 Current regulatory environment

a) The *Online Safety Act 2021* (Cth) and *Broadcasting Services Act 1992* (Cth) do not address dehumanising disinformation operations platformed and profited from by international digital platforms. As such, regulators like e-Safety and ACMA are not positioned to act.

b) The current Australian Code of Practice on Disinformation and Misinformation applies to international digital platforms. However, it has no effective enforcement mechanism and is self-regulatory.

c) The Broadcasting Services Act contains some safeguards against vilification by media, but they are rarely enforced and do not apply in relation to online content.[1]

5.3 Regulating International Digital Platforms

a) While systems that promote safety by design are critical, we cannot escape the need for definitional clarity on harms. For example, seeking transparency on algorithms or hate speech data won't help if our framework is ambiguous on how we define harm.

b) The UK online safety bill developments in December 2022 underscore the pitfalls of not providing definitional clarity, with previous efforts to address online hate and misinformation erased from the bill.

5.4 The effect of dehumanisation

a) Referring to the Australian terrorist who carried out the Christchurch attack, Lentini (2019, 43) explains that,

Tarrant's solution to the crisis – indeed one on which he felt compelled to enact – was to annihilate his enemies (read Muslim migrants). This included targeting non-combatants. In one point in his 'manifesto', he indicates that they constitute a much greater threat to the future of Western societies than terrorists and combatants. Thus, he argues that it is also necessary to kill children to ensure that the enemy line will not continue…Tarrant indicated that, when trying to remove a nest of snakes, the young ones had to be eradicated. Regrettably, children were among those whom he allegedly shot and killed.2

b) A similar narrative inspired Anders Breivik, the Oslo terrorist who murdered 77 people in 2011. Breivik cited the author of JihadWatch, one of the information operations cited in Australian research.[3] The historical links between these two attacks, in terms of their relationship to 'counter jihad' dehumanising information operations considerable.[4] With respect to dehumanisation, Kaldor (2021) notes,

Breivik also refers to Muslims as "wild animals," who he argues are freely bringing about European "genocide" because "traitors... allowed these animals to enter our lands, and

[1] Refer to supplementary submission from AMAN containing evidence of unmoderated hate speech on Sky News Australia Facebook page.

[2] Lentini, Peter. 2019. "The Australian Far-Right: An International Comparison of Fringe and Conventional Politics" in Mario Peucker and Debra Smith, eds. *The Far-Right in Contemporary Australia*. Singapore, 43.
[3] Abdalla, Mohamad, Mustafa Ally, Rita Jabri-Markwell. 2021. "Dehumanisation of Out-Groups on Facebook and Twitter: Towards an Assessment Framework for Online Hate Actors and Organisations." *SN Social Sciences* (1) 9; Peucker et al (2022), op cit.

[4] Rita Jabri Markwell, "The online dehumanisation of Muslims made the Christchurch massacre possible" *ABC Religion and Ethics*, 31 August 2020, https://www.abc.net.au/religion/the-online-dehumanisation-of-muslims/12614148

continue to facilitate them." In keeping with the naturalistic theme, Tarrant's text is also rife with mixed metaphors describing how individuals such as himself can no longer escape Western civilisation's contamination: "there is no sheltered meadow... there is not a single place left where the tendrils of replacement migration have not touched." Comparing immigrants to a "vipers [sic] nest", he implores followers to "burn the nest and kill the vipers, no matter their age." Crusius similarly bewails how those without the means to "repel the millions of invaders" "have no choice but to sit by and watch their countries burn." The repetition of animalistic metaphors is no accident: the perpetrators intentionally dehumanise immigrants by depicting them as beastly, thereby making their complaint about Western society's perceived decline more justifiable to their readers.5

c) 'Dangerous speech', a category expounded in detail by Maynard and Benesch (2016), is speech that constructs an 'outgroup' as an existential threat to the 'in-group,' whether this threat is real or otherwise (81).[6] Dehumanisation and another technique called 'threat construction' are two techniques used in dangerous speech. They are often inextricably linked: 'where dehumanization makes atrocities seem acceptable, threat construction takes the crucial next step of making them seem necessary' (82).

d) Researchers from Macquarie and Victoria Universities published the first study mapping the online activity of right-wing extremists (RWE) in New South Wales, Australia.[7] The study identified the dehumanisation of out-groups to in-group audiences as a core component of their online socialisation. [8]

e) Significantly, their research found that dehumanisation existed on 'low-risk' platforms like Facebook and Twitter 'without violating platform moderation policies.'

f) An investigation by Guardian news revealed an overseas commercial enterprise that was

   i. using its 21-page network to churn out more than 1,000 coordinated faked news posts per week to more than 1 million followers, funnelling audiences to a cluster of 10 ad-heavy websites and milking the traffic for profit.

   ii. The posts stoke deep hatred of Islam across the Western world and influence politics in Australia, Canada, the UK and the US by amplifying far-right parties

[5] Sophie Kaldor, 'Far-Right Violent Extremism as a Failure of Status: A New Approach to Extremist Manifestos through the Lens of Ressentiment' (Research Paper, International Centre for Counter-Terrorism – The Hague, May 2021) 17 https://icct.nl/app/uploads/2021/05/Far-Right-Violent-Extremism-as-a-Failure-of-Status.pdf.

[6] Maynard, Jonathan Leader and Susan Benesch. 2016. "Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention." *Genocide Studies and Prevention: An International Journal* 9(3): 70.

[7] Department of Security Studies and Criminology. (2020, October 9). Mapping Networks and Narratives of Online Right-Wing Extremists in New South Wales (Version 1.0.1). Sydney: Macquarie University.

[8] Maynard and Benesch, op cit.

such as Australia's One Nation and vilifying Muslim politicians such as the London mayor, Sadiq Khan,, and the US congresswoman Ilhan Omar.

5.5 Algorithmic impact

a) There are three main categories of algorithms[9]

Algorithms for content recommendations: Platforms may use algorithms to recommend content in ways that personalise recommendations for individual users based on their past behaviors (as well as inferred characteristics) and optimize expected value to the company by maximizing individual users' expected engagement with recommended content. When platforms use algorithms to maximize engagement, they cannot fully prevent harmful third-party contents from being recommended to users if those users have consumed similar contents in the past.

Algorithms for content moderation and safety: Platforms use algorithms to prevent and reduce harms by semi-automating the process of flagging, removing, and re-ranking third-party contents likely to violate platform policies or laws. When this process is performed at scale, the algorithms cannot perform perfectly and are continuously optimized to balance between precision and accuracy.If a platform prioritizes accuracy over precision in using algorithms for content moderation, its process would have a high false positive rate. Most large platforms therefore choose to prioritize precision over accuracy, which allows most users to post contents but can sometimes lead to extensive harm when false negatives are shared widely.

Algorithms for advertising and commerce: Platforms use algorithms to serve targeted ads to individuals through "retargeting," which relies on expressed and inferred information about those individuals that the platforms had already compiled. Algorithms that are used in techniques like "retargeting" primarily benefit companies, and this encourages companies to collect more and more data about users.

(b) The above is important to note as it applies to state and non-state-sponsored information operations.

(c) Research into algorithmic impact still requires well-defined definitions from the Online Safety Act to make assessments.

5.6 The formula of dehumanising information operations

a) Abdalla et al. (2021) studied the operation of five (5) online information operations located within the extreme right 'counter jihad' movement. The leading actor studied conveyed a similar demographic invasion narrative in Tarrant's manifesto. They found that Facebook and Twitter's automatic detection tools could not detect explicit dehumanising slurs or violent fantasy in the fantasies threads, meaning that online

[9] Integrity Institute, *Summary: Integrity Institute's Amicus Brief on Gonzalez v. Google,* 9 December 2022, https://integrityinstitute.org/amicus-brief-summary-sheet

communities could react together to information towards a targeted group without disruption.[10] In this phenomenon, we see a combination of the cognitive, behavioural and social license granted to participants in the comment threads to erase the humanity of the target group and indulge in violent fantasy. It is contended by Abdalla et al. that,

i. the marshalling of stories to create an overwhelming sense of crisis and disgust does not always rely on explicit dehumanising descriptors, verbs, or coded language in the headlines. Where an audience had been primed over time, implied properties in text are capable of triggering entire sub-texts.

ii. Further, actors often attributed subhuman actions to Muslims in the headlines to dehumanise all Muslims overtime. Platforms could not detect this technique as they were focused exclusively on dehumanising comparisons, synonyms and adjectives (e.g., disease, filth, cancer, weeds, insects).

b) Abdalla et al's research points to several predictors that could be used to make competent and consistent assessments of hate actors running purposed information operations.

5.7 Volumetric 'pile on' attacks

a) We ask e-Safety to acknowledge that coordinated 'pile-ons' on people, causing a person to receive a large volume of hatred about their protected attribute(s), is inherently harmful. A person reporting this experience should not need to demonstrate psychological harm arising from it. The evidence of hatred based on a protected attribute should be sufficient. Online material initiating such activity should be quickly actionable under existing cyberbullying and cyber abuse provisions.

b) Adopting definitions of dehumanising material may also help e-Safety in analysing the impact of particular complaints.

5.8 Access to data for research

a) The Digital Services Act to Article 40 of the digital services act (Europe) enables platform-to-researcher data sharing, with guardrails for privacy protection.

b) The Digital Services Act requires very large online platforms or very large online search engines to provide access to data to vetted researchers for the sole purpose

of conducting research.[11] Researchers become vetted upon application if their application fulfils the following requirements:[12]

i. Affiliation with a scientific research organisation;

ii. Independence from commercial interests;

iii. Disclosure of the funding of their research;

iv. Capability of fulfilling specific confidentiality and data security requirements in relation to protecting personal data and a description of their specific technical and organisational measures;

v. Their access to the data is necessary for the purposes of their research;

vi. Their research is for the purpose of the detection, identification, and understanding of specific risks to the EU or the assessment of the adequacy, efficiency and impacts of the risk mitigation measures of very large online platforms and very large online search engines; and

vii. They make their research results public, free of charge, within a reasonable period after their research is completed, subject to the rights and interests of the recipients of the service concerned.

c) The procedure to underpin this law is being developed in subordinate legislation.

d) The specific risks to the EU to which the purpose of research may relate are:[13]

i. The dissemination of illegal content;

ii. Actual or foreseeable negative effects on fundamental rights enshrined in the Charter of Fundamental Rights of the European Union;

iii. Actual or foreseeable negative effects on civic discourse and electoral processes, and public security; and

iv. Actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors, and serious negative consequences to the person's physical and mental well-being.

[11] Single Market For Digital Services and amending Directive 2000/31/EC, Regulation (EU) 2022/2065 of the European Parliament and of the Council, 19 October 2022 art 40(4).

[12] Ibid art 40(8).

[13] Ibid art 34(1).

e) We note these categories are broad and not well-defined because the DSA, as European legislation, leaves that definitional work to member states. Australia cannot escape from the definition question if it is looking to empower the research sector with such a mechanism. This is again, where definitions such as the dehumanising material definition in **Schedule 1** become important.

f) Providers of very large online platforms and very large online search engines are required to give vetted researchers access without undue delay to data (known as the "crowdtangle provision".[14]

[14] Ibid art 40(12).

SCHEDULE 1

AMAN's working definition of dehumanising material, updated 15 July 2023

Note this definition is subject to ongoing revision until it is formally published.

(1) Dehumanising material is the material produced or published, which an ordinary person would conclude, portrays the class of persons identified on the basis of a protected characteristic ("class of persons") as not deserving to be treated equally to other humans because they lack qualities intrinsic to humans. Dehumanising material includes portraying the class of persons:

(a) to be or have the appearance, qualities, or behaviour of

(i) an animal, insect, filth, form of disease or bacteria;

(ii) inanimate or mechanical objects; or

(iii) a supernatural alien or demon.

(b) are polluting, despoiling, or debilitating an ingroup or society as a whole;

(c) have a diminished capacity for human warmth and feeling or to make up their own mind, reason or form their own individual thoughts;

(d) homogeneously pose a powerful threat or menace to an in-group or society, posing overtly or deceptively;

(e) are to be held responsible for and deserving of collective punishment for the specific crimes, or alleged crimes of some of their "members";

(f) are inherently criminal, dangerous, violent or evil by nature;

(g) do not love or care for their children;

(h) prey upon children, the aged, and the vulnerable;

(i) was subject as a group to past tragedy or persecution that should now be trivialised, ridiculed, glorified or celebrated;

(j) are inherently primitive, coarse, savage, intellectually inferior or incapable of achievement on a par with other humans;

(k) must be categorised and denigrated according to skin colour or concepts of racial purity or blood quantum; or

(l) must be excised or exiled from public space, neighbourhood or nation.

(2) Without limiting how the material in section (1) is presented, forms of presentation may include,

(a) speech or words;

(b) the curation or packaging of information;

(c) images; and

(d) insignia.

Intention component

If the above definition was used as a standalone civil penalty, it should be complemented by an intention component:

in circumstances in which a reasonable person would conclude that the material was intended to portray the class of persons as not deserving to be treated equally to other humans or to incite hatred, serious contempt or severe ridicule toward the class of persons.

Adding an intention element may make enforcement more difficult and may not be necessary, especially if the definition is used as part of a legal framework where there are already intention components or exceptions available.

How did we develop this working definition?

AMAN developed this working definition after spearheading a study of five information operations online (Abdalla, Ally and Jabri-Markwell, 2021). The first iteration of this definition was published in a joint paper with UQ researchers (Risius et al, 2021). It continues to be developed with input received from researchers, lawyers and civil society.

Possible dehumanising conceptions are surfaced through research and then tested against Haslam's frame of whether it deprives a group of qualities that are intrinsic to humans. If a subject is dehumanised as a mechanistic form, they are portrayed as 'lacking in emotionality, warmth, cognitive openness, individual agency, and, because [human nature] is essentialized, depth.' A subject that is dehumanised as animalistic, is portrayed as 'coarse, uncultured, lacking in self-control, and unintelligent' and 'immoral or amoral' (258).

Some conceptions are found to fall outside the frame of dehumanisation but could still qualify as vilification or discrimination, for example, using anti-discrimination laws.

The three categories of dehumanising comparisons or metaphors in Clause (a) are drawn from Maynard and Benesch (80), and fleshed out with further examples from tech company policies (refer to Meta for example).

Clause (b) is derived from Maynard and Benesch (80).

Clause (c) is derived from Haslam (258).

Clauses (d) and (e) are elements of dangerous speech that Maynard and Benesch refer to as 'threat construction' and 'guilt attribution' respectively (81). However, Abdalla, Ally and Jabri-Markwell's work shows how such conceptions are also dehumanising, as they assume a group operates with a single mindset, lacking independent thought or human depth (using Haslam's definition), and combine with ideas that Muslims are inherently violent, barbaric, savage, or plan to infiltrate, flood, reproduce and replace (like disease, vermin)(15). The same study found that the melding and flattening of Muslim identities behind a threat narrative through headlines over time was a dehumanisation technique (17). Demographic invasion theory-based memes (9) or headlines that provided 'proof' for such theory (20) elicited explicit dehumanising speech from audiences.

Maynard and Benesch write, 'Like guilt attribution and threat construction, dehumanization moves out-group members into a social category in which conventional moral restraints on how people can be treated do not seem to apply' (80).

Clauses (f), (h), (i) are drawn from the ‘‘Hallmarks of Hate’, which were endorsed by the Supreme Court of Canada in *Saskatchewan (Human Rights Commission) v. Whatcott* 2013 SCC 11, [2013] 1 S.C.R. 467. These Hallmarks of Hate were developed after reviewing a series of successful judgements involving incitement of hatred to a range of protected groups. These clauses were tested using Haslam’s definitional frame for the denial of intrinsic human qualities.

Clauses (f) (‘criminal’) and (g) are drawn from harmful characterisations cited in the Uluru Statement of the Heart.

Clauses (j) and (k) were updated following AMAN’s observations of online information operations generating disgust toward First Nations Peoples. Disgust is a common effect of dehumanising discourse. These clauses were tested using Haslam’s definitional frame for the denial of intrinsic human qualities.

Clause (l) was drawn from Nicole Asquith’s Verbal and Textual Hostility Framework. (Asquith, N. L. (2013). The role of verbal-textual hostility in hate crime regulation (2003, 2007). Violent Crime Directorate, London Metropolitan Police Service.) The data and process used to formulate this Framework is exceptional. Reassuringly, this research had surfaced examples that were already captured by this Working Definition of Dehumanising Material.

This working definition is a work in progress. AMAN welcomes feedback as it continues to be developed.

Updated 15 July 2023

SCHEDULE 2

Possible improved wording for 'professional news content' definition

(i) Professional news content produced by a news source who

(a) Is subject to

i. The rules of the Commercial Television Industry Code of Practice, the Commercial Radio Code of Practice or the Subscription Broadcast Television Codes of Practice; or

ii. Rules of code of practice mentioned in paragraph 8(1)(e) of the Australian Broadcasting Corporations Act 1983 or paragraph 10(1)(j) of the Special Broadcasting Services Act 1991; and

(b) Is subject to internal editorial standards that

i. Relate to the provision of quality journalism;

ii. Ensure that factual information is reported without bias;

iii. Implement labels that assist readers and audiences in distinguishing between news and opinion content;

iv. Require diversity of opinion on controversial issues;

v. Require pre-publication fact-checking and post-publication corrections that are adequately and transparently disseminated;

vi. Prohibit material that is hateful or incites hatred against individuals or groups on the basis of protected characteristics;

vii. Are published on its website and easily accessible; and

viii. Provide an electronic email address and postal address for complaints.

(c) Publishes current information on their website that

i. Provides full transparency as to its sources of funding; and

ii. Provides full transparency as to the number of executive or board-level financial and editorial decision-makers.

(d) Has editorial independence from the subjects of the news source's news coverage.