AMAN
AUSTRALIAN MUSLIM ADVOCACY NETWORK
Law • Policy • Research • Voice
www.aman.net.au

Policy proposals to combat online dehumanisation of minorities

10 November 2021

AUSTRALIAN MUSLIM ADVOCACY NETWORK

[redacted]

Website: www.aman.net.au

Director, Online Safety Reform and Research Section

Department of Infrastructure, Transport, Regional Development, and Communications

GPO Box 2154

Canberra ACT 2601

Dear Director,

Please accept this submission concerning the draft Online Safety (Basic Online Safety Expectations) Determination.

The Australian Muslim Advocacy Network (AMAN) seeks to create a safer and more inclusive Australia by safeguarding Australian Muslims' equal rights and protections. We are a civil society organisation that researches and monitors online discourse, research, and dialogues on policy.

Our engagement has been direct with platforms and industry bodies like the GIFCT and civil society networks, like the Christchurch Call to Action. We have engaged with researchers and NGOs across Australia and internationally. We participated in the *Australian Code of Practice on Misinformation and Disinformation* consultation process. This submission builds on AMAN's earlier proposal to the Australian Government on the Online Safety Act from 12 February 2021, which can be accessed here.

We have also practised in the field of applying existing vilification laws to online hate speech. Through our experience of the Fraser Anning case, we have specific advice about our anti-discrimination framework may be connected to the online sphere.

However, recognising that hateful echo chambers are a public harm, not a private one, we have also proposed civil penalties through a notice and action model for actors that serially publish dehumanising language or discourse – creating a consequence for both the individuals and the platforms that allow and amplify this practice.

Further, we have proposals to restore transparency and the market conditions for platform accountability.

In respect to the draft Determination, it appears that the most significant scope for the arguments made in this paper is under the first core expectation – that platforms provide a safe environment for users. We are not in a position to evaluate how our proposals could be incorporated in this Determination, given the broader constraints of the Act. We encourage the Australian Government to engage with us.

[redacted]

Yours faithfully,

[redacted]

AUSTRALIAN MUSLIM ADVOCACY NETWORK (AMAN)

Proposal No.	What we are proposing	Who is this proposal for	Why will this help
A	Defining dehumanizing language and discourse in policy	Government, Community awareness, Media, Law enforcement, Platforms, Regulators, Political parties, GIFCT	Increase understanding of the harm and community resilience to it
B	Creating civil penalties for serial dehumanizing content (for individuals and the platforms that enable it	Federal Government, regulators	Disincentivize making money this way
C	A hate actor assessment framework - to measure aggregate harm of borderline content that dehumanizes over time	Platforms, A.I.models, regulators	Identify and address disinformation not currently picked up
D	Increase access to justice for victim communities against online hate actors	Federal Government	Close the gap between Australia's standards on vilification and discrimination and the online sphere.
E	Mandate transparency on a range of matters	Federal Government, regulators,	Address the discriminatory and harmful effects of algorithms and provide key information to consumers/users and advertisers
F	Antitrust legislation	Federal Government, ACCC	Enable advertisers to have a choice about where to advertise, therefore restoring market forces that pressure social media companies to uphold human rights

Table of Contents

ACKNOWLEDGEMENT

AMAN acknowledges the traditional custodians of the land we work upon, the Elders and Ancestors who have walked this land before us. We pay our respect to Aboriginal and Torres Strait Islander peoples both past, present and emerging. We fully support the self-determination of Aboriginal and Torres Strait Islander peoples.

INTRODUCTION

Social media plays a significant role in priming and socialising people towards violence. The current response is to expand the national security apparatus continually. However, this comes with substantial costs to our collective freedoms. Currently, we are dealing with **a disinformation and internet governance problem by ramping up surveillance and police**. Despite the challenges in confronting disinformation and internet governance, there is a worse cost in accepting the idea that it is too complex or too much of a slippery slope to act. Ethnic and religious minorities in Australia are being asked to battle the erosion of their collective safety and security alone.

Confronting the challenges of internet governance and harms like disinformation must include civil society at the table. The human rights at stake are too great for government officials to make these decisions alone.

On regulation, one of the first challenges is how to define what is unsafe or unharmful in a way that does not give rise to substantial ambiguity. Defining extremist material or activity at law is more fraught. This ambiguity creates anxiety about state or tech intrusions on freedom of speech. Thus, instead of defining extremist material or activity, the Australian Government should consider targeting a technique that many violent extremist movements rely on: dehumanisation of outgroups.

Designing proportional levers that both disrupt the most potent vectors of harm and restore market forces to pressure social media companies to uphold human rights must be considered. Our proposals below respond to these issues.

PROPOSALS A-C

A. Define the act of dehumanisation in policy

(1) We propose that the Australian Government define the act of dehumanisation in policy and enable community education and discussion about its meaning. We suggest the definition below.

An actor that serially or systematically produces or publishes material, which an ordinary person would conclude,

(a) presents the class of persons identified on the basis of a protected characteristic (e.g., race or religious belief) to have the appearance, qualities, or behaviour of an animal, insect, filth, form of disease or bacteria, inanimate or mechanical objects, or a supernatural threat. This material would include words, images, and/or insignia. ("***Dehumanising language**")*

(b) curates information to a specific audience to cumulatively portray that the class of persons identified on the basis of a protected characteristic (e.g., race or religious belief)

(i) are polluting, despoiling, or debilitating society;

(ii) have a diminished capacity for human warmth and feeling or independent thought;

(iii) act in concert to cause mortal harm; or

(iv) are to be held responsible for and deserving of collective punishment for the specific crimes, or alleged crimes of some of their "members" ("***Dehumanising discourse"***)[1]

(2) We recommend teaching school children about the role of dehumanisation in historical atrocities and how to spot discourse that may be trying to dehumanise a minority group.

(3) We recommend that this education be offered to law enforcement, the media industry, media regulators, social media companies based in Australia, social media regulators, the advertising industry, elected representatives, and their staff.

[1] A similar, earlier definition was also outlined in Risius et al (2021).

(4) We recommend that all political parties change their governing documents to define dehumanisation as above and require candidates and elected representatives not to publish or promote it.

B. Introduce civil penalties in the Online Safety Act

(1) It is proposed that Australia's Online Safety Framework be expanded to actors who serially or systematically publish materials from a website or organisation that, over time, creates an aggregate harm of dehumanising an outgroup to an ingroup audience.

(2) It is proposed that the civil penalties would mirror the definitions for dehumanising language and discourse provided above.

(3) The e-Safety Commissioner should identify an actor that meets the standard of aggregate harm. We support the principle that judicial functions should not be delegated to platforms. Therefore, it is proposed that the e-Safety Commissioner's office decide whether there is a breach of the proposed civil penalties and issue a notice to the platform and the actor who published the content. A platform or individual's failure to take action should incur penalties for both. There would be the option for judicial review.

(4) The *Actor Indicators* (below), identified in AMAN's 2020 study, could be used by e-Safety administrators to assess dehumanising discourse.

(5) The e-Safety Commissioner should also consider the context in making a determination. The Rabat Plan also emphasises context: of the speaker's power, their intent, the content and form and spread.

The rationale for how it meets international legal guidelines on good internet governance law is below.

C. Introduce industry standard for assessing dehumanising discourse

We propose that the e-Safety Commissioner develop an industry standard for assessing aggregate harm of dehumanisation. That industry standard could build upon the Actor indicators below. In 2020, we studied five actors producing significant amounts of blog or pseudo-news content that triggered explicitly dehumanizing and violent responses by users on Facebook and Twitter. The findings of that research were published in a peer-review journal in September 2021 (Abdalla, Ally and Jabri-Markwell, 2021).

That study found the following markers were common to all five actors' information operations (*"Actor indicators"*):

(a) Dehumanizing conceptions or conspiracy theories on the actor's website (where applicable) about an identified group ("the outgroup") based on a protected characteristic;

(b) Repeated features of the **headlines and images** that are curated for a specific audience, including:

(i) Essentializing the target identity through implicating a wide net of identities connected to the protected group (e.g., “Niqab-clad Muslima,” “boat migrants,” “Muslim professor,” “Muslim leader,” “Iran-backed jihadis,” “Ilhan Omar,” “Muslim father”);

(ii) A high degree of hostile verbs or actions (e.g., stabs, sets fire) attributed to those subjects;

(iii) A primary proportion of actor’s material acting as “factual proofs” to dehumanizing conceptions about outgroup;

(iv) Potential use of explicitly dehumanizing descriptive language (e.g., frothing-at-the-mouth) or coded extremist movement language with dehumanizing meaning (e.g., invader, a term used in RWE propaganda to refer to Muslims as a mechanically inhuman and barbaric force). However, for the most successful actors, dehumanizing slurs were avoided to maintain legitimacy and avoid detection; and

(v) Where there was no dehumanizing language, there was a presence of “baiting” through rhetorical techniques like irony to provoke ingroup reactions; and

(c) Evidence in the user comment threads of a pattern of hate speech against the outgroup.2

[2] This summary can also be found in Riisus et al (2021).

Illustrative snapshot of one actor's headlines

India: Muslim poses as Hindu to trap minor Hindu girl, abducts and rapes her, pressures her to convert

Bangladesh: Muslims threaten Christian family, force them to leave their home, steal their land

Poland vows to 'defend Europe' from 'migrant invasion' unleashed by Belarusian dictator Lukashenko

While World Focuses on 'Islamophobia,' Christians Live Precarious Existence in Muslim Lands

France: Muslim migrant showers racist insults on black Africans, screams 'Allahu akbar,' steals knife

Germany: 63-year-old Afghan Muslim migrant slits 21-year-old wife's throat in front of their children

Audio: Robert Spencer on mass Muslim migration in Europe and Biden's nominee for Religious Freedom ambassador

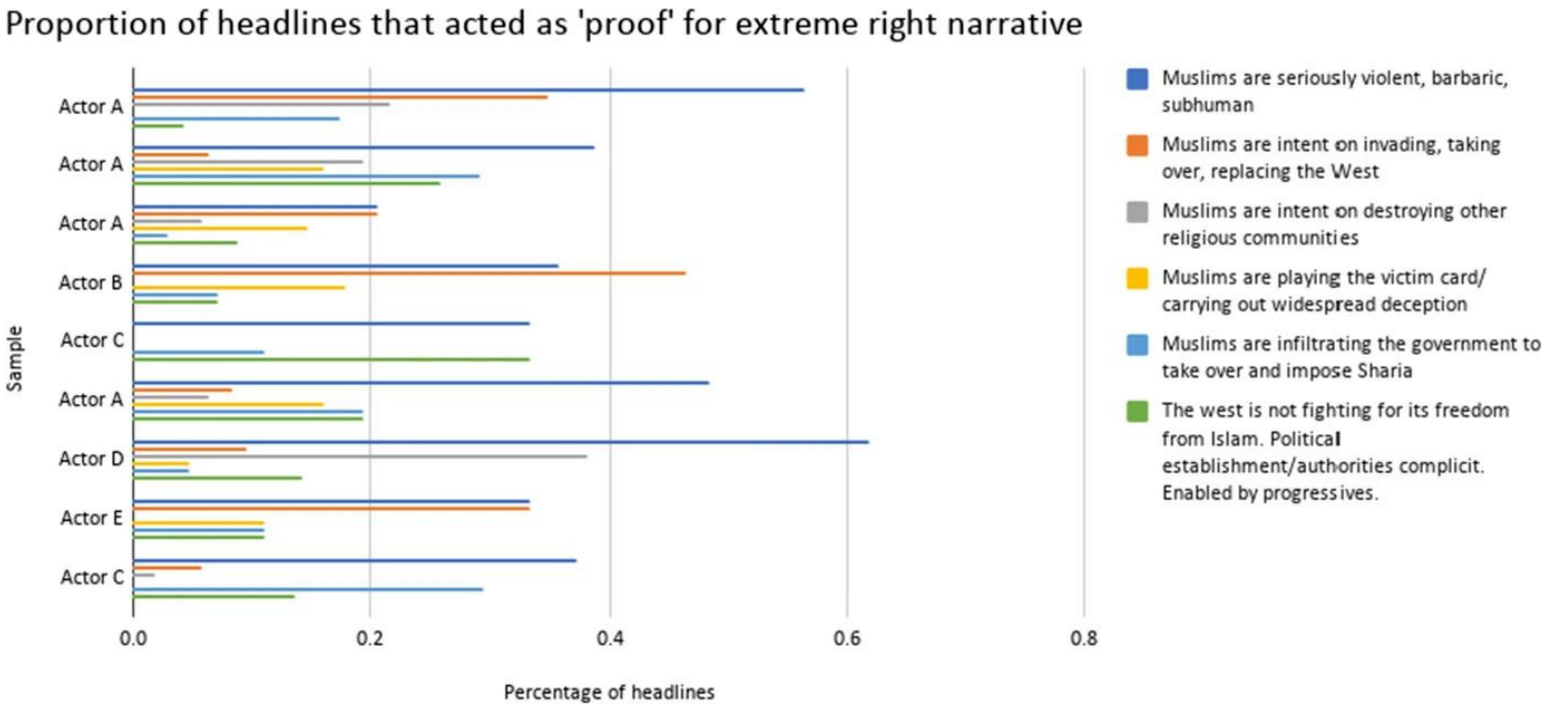
Cyprus: Six Muslims, including at least four migrants, plot jihad murders of five Israelis

Austria: Muslim migrant father, officially 21, attacks his son, 10, and two bystanders with a knife

Robert Spencer Video: Jihadis Are Entering the US Across the Southern Border

Norway: Muslim migrant screaming 'Allahu akbar' and quoting Qur'an attacks people with knife, 'motive unclear'

Figure 8 from Abdalla, M., Ally, M. & Jabri-Markwell, R. Dehumanisation of 'Outgroups' on Facebook and Twitter: towards a framework for assessing online hate organisations and actors. SN Soc Sci 1, 238 (2021). https://doi.org/10.1007/s43545-021-00240-4



RATIONALE

It is a challenge to consider the dispersed social harm that stems from dehumanisation into an individualistic frame. Our recommendations have focused on the vectors of this harm, that being individuals who serially post dehumanising material; and through setting an industry standard for digital platforms when making detailed and contextualised assessments about individual accounts, pages, groups, and channels. As civil provisions, this would create a consequence for both individuals serially engaged in this practice, along with platforms that disregard it. As civil provisions, it is also possible to set aside the requirement often put forward in criminal contexts that there be evidence of foreseeable or imminent physical harm.

The Rabat Plan also emphasises context: of the speaker's power, their intent, the content and form, spread, and likelihood and imminence of harm. While imminence of harm would not be a necessary threshold requirement for the civil penalty we have proposed, the other contextual factors would be considered. It is also vital that targeted communities are consulted on their particular contexts. Otherwise, decision-makers will fail to make fully competent judgements.

The Rabat Plan of Action noted the importance of distinguishing not just criminal and civil prohibitions but on a broader class that will "still raise concerns in terms of tolerance, civility, and respect for the convictions of others." If we limit civil prohibitions to the most severe end of the spectrum (serial and clear-cut examples) and invoke the Act's Basic Online Safety Expectations and an Industry Standard as levers to engender platform accountability on a broader range of dehumanising speech or discourse, this will go a long way to satisfy Australia's obligations under international human rights law in terms of protecting freedom of expression.

The connection between violent extremism and dehumanisation

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.

Anders Breivik, the Oslo terrorist who murdered 77 people in 2011, was inspired by a similar Anti-Islam demographic invasion narrative. The links between these two attacks in ideology and other aspects are considered in the literature. On 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 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.

A Victoria University study in 2018, the year before the Christchurch attack, found that the narratives expounded by Tarrant were prevalent on Facebook (Peucker, Smith and Iqbal, 2018),.

Right-wing extremism in the NSW context has been defined as 'individuals, groups, and ideologies that reject the principles of democracy for all and demand a commitment to *dehumanising and/or hostile actions against*

outgroups' (Department of Security Studies and Criminology, 2020, p. 1) [emphasis added].

Policy gap on purposed information operations to dehumanise 'outgroups'

At least Facebook, Instagram, Youtube, LinkedIn, and Twitter ban dehumanising speech or content. However, the presence of dehumanising language is unnecessary to propagate dehumanising discourse. Social media companies primarily rely upon the specificity of dehumanizing language to detect dehumanisation. Explicit dehumanising adjectives and comparisons are often made in the comment threads in response to dehumanising disinformation, but comment thread violations are poorly detected by automated tools and rarely reported by users in hateful echo chambers.

The Australian Government's only policy addressing disinformation is a Code of Practice recently designed and instituted by the social media industry body Digi. This self-regulatory code applies a definition of disinformation where it must cause serious and imminent harm. This is an impractical definition that voids the very function of disinformation, which is a cumulative and creeping threat. AMAN has outlined its concerns with this Code (AMAN, 2021).

How research defines dehumanisation

Dehumanisation offers an enduring, internationally accepted, and well-defined concept, grounded in genocide prevention studies and increasingly in the literature on countering violent extremism.

Dangerous speech', a category that has been expounded in detail by Maynard and Benesch (2016)[3], is speech that constructs an 'outgroup' as an existential threat to the 'ingroup,' whether this threat is real or otherwise (81). Maynard and Benesch identify the range of techniques commonly used in dangerous speech. Dehumanisation and another technique called 'threat construction' are often inextricably linked, given that 'where dehumanization makes atrocities seem acceptable, threat construction takes the crucial next step of making them seem necessary' (82).

[3] Maynard JL, Benesch S (2016) Dangerous speech and dangerous ideology: an integrated model for monitoring and prevention. Genocide Stud Prev 9(3):70

According to the authors, dehumanisation is the most frequently employed technique in dangerous speech, where '[t]argets … are described in a variety of ways that deny or diminish their humanity, reducing the moral significance of their future deaths or the duties owed to them by potential perpetrators' (80). Dehumanisation is often achieved by 'describing them as either biologically subhuman ("cockroaches," "microbes," "parasites," "yellow ants"), mechanically inhuman ("logs," "packages," "enemy morale"), or supernaturally alien ("devils," "Satan," "demons")'—and has been used historically to represent a minority as an existential threat to the majority(80). Maynard and Benesch also find that dehumanisation can be carried out without 'hatred or blatant exclusionary discourse' (70).

Haslam's (2006) model that proposes links between conceptions of humanness and corresponding forms of dehumanization provided further detail for a theoretical base of this study's discourse analysis. Like Maynard and Benesch, he refers to 'animalistic' and 'mechanistic' forms of dehumanisation but details the characteristics that underpin both. 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).

Distinguishing disinformation from news commentary and partisan talk

Information campaigns acting as vehicles for widespread dissemination of dehumanizing conceptions and discourse will need to be distinguished from news commentary, partisan talk, or fringe discourse (Risius et al, 2021, 58). This is possible using the Actor indicators above.

A policy proposal from University of Queensland and AMAN researchers to the Global Internet Forum to Counter Terrorism (Risius et al, 2021) suggested that serial or systematic dehumanization of an outgroup should be used as a definitory factor to distinguish violent extremist content from fringe discourse.

The evaluative framework for good law

Freedom of expression is a fundamental human right, but so is the right to life, the security of person, equality, and non-discrimination. Restrictions of the right to freedom of expression must be clearly prescribed by law, pursue a legitimate aim, be necessary for a democratic society, and be proportionate to the aim pursued. Avoiding discrimination and personal endangerment on the basis of one's race or religion are legitimate aims that are necessary to democracy. But achieving these aims must be done through laws that are proportionate and clearly prescribed by law.

Distinguishing state and platform responsibilities

Where online speech is concerned, an appropriate legal framework clearly establishes and distinguishes states' obligations and intermediaries' responsibilities to protect the human rights of online users (AccessNow, 2020). Currently, there is no such clarity or legal framework in Australia, where it concerns vilification, discrimination, and disinformation. The Online Safety Act is silent on those aspects of the law. This lack of clarity doubles the load upon affected, marginalised segments of the community and has a discriminatory effect (section 9 *Racial Discrimination Act 1975* – indirect discrimination).

There is also an emerging view from digital rights defenders that state regulatory models (imposing penalties upon platforms) should focus specifically on manifestly illegal content (like child abuse material) and "avoid regulation regarding ever-evolving definitions of online societal phenomena, such as disinformation, hate speech, or terrorist content" (AccessNow, 2020, 42). Where content decisions involve restricting access to context-dependent illegal content, the legitimate purpose should always be determined by an independent judicial authority or other independent administrative body whose decisions are subject to judicial review.

Good practice for notice-and-action procedures

AccessNow argues that the following notice-and-action procedures should be considered as adequate, determined by the type of infringement at stake as well as the category of content:

(1) Notice-and-notice

(2) A notice-wait-and-takedown mechanism that enables a content provider to file a counterclaim

(3) Notice-and-judicial takedown, where courts review the legitimacy of content removals, should always be available to all users, regardless of the type of content

(4) Private notice-and-takedown should only be used in limited content cases that are legally defined as manifestly illegal.

Further AccessNow advocates for Basic minimum requirements of a valid notice.:

(1) Reason for complaint

(2) Location of the content

(3) Evidence for the claim

(4) Consideration of limitations, exceptions, and defense available to the content provider

(5) Declaration of good faith. Notices submitted by states should be based on their assessment of the illegality of the notified content, following international standards. Language for content restrictions should provide for notice of such restriction being given to the content producer/issuer as early as possible unless this interferes with ongoing law enforcement activities. Information should also be made available to users seeking access to the content according to applicable data protection laws. Users should not be forced to identify themselves when submitting the notice, and they should provide their contact details only voluntarily

The content of the law: legality and definitional clarity

The United Nations Human Rights Committee has found that the International Covenant on Civil and Political Rights (ICCPR) applies to the online sphere.

Article 19(3) of the ICCPR sets out a framework describing the limited circumstances in which states may legitimately restrict freedom of expression (UN Human Rights Committee, 2011). The Global Network Initiative's Report on Content Regulation and Human Rights (2020) explains that this framework consists of three interrelated principles: legality, legitimacy, and necessity.

The principle of legality establishes two requirements for the regulation of expression. First, it requires that restrictions on freedom of expression must be provided with public laws "formulated with enough precision to enable an individual to regulate his or her conduct accordingly" (UN Human Rights Committee, 2011, para 25). These laws must be validly enacted and publicly available so that individuals are effectively put on notice as to what conduct and content are prohibited. Second, they must "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not." This latter concern is significant in-laws that outsource speech regulation enforcement to private actors of varying sizes, business models, and capacities.

The legality requirement is essential to mitigating the chilling effect of ambiguous laws on online expression. Any vagueness or ambiguity is likely to cause individuals to refrain from exercising their rights and lead intermediaries to be overly aggressive in censoring expression for fear of being held in violation of the law.

The problem with defining extremist material

One of the Australian Online Safety framework goals is to prevent violent extremists and terrorists from exploiting digital platforms. While this is an important aim, it does not capture the actors and online echo chambers that work to socialise individuals towards violence.

Online hateful echo chambers that socialise individuals towards violence include a significant amount of violent fantasy and incitement. Platforms rarely detect materials as they are buried within comment threads and lack organisational labels that platforms rely on.

Proposals to expand designation and proscription lists have struggled with the political and legal difficulty of defining 'extremist ideology' or 'extremist rhetoric' where there are no explicit or imminent calls to violence. The scope for 'terror-scaping' ideas, organisations, or individuals, merely because they present as extreme, unpopular, or fringe, is a genuine concern, especially for marginalised communities that are already subject to over-policing and may have legitimate grievances with nation-states.

A review by the U.K Independent Commission for Countering Extremism recommended establishing a legal framework to counter hateful extremism, which it has defined as:

> activity or material directed at an outgroup" (e.g., Muslims) who are perceived as a threat to an ingroup (e.g., a Far-Right group) "motivated by or intending to advance a political, religious, or racial supremacist ideology: a. To create a climate conducive to hate crime, terrorism, or other violence; or b. Attempt to erode or destroy the fundamental rights and freedoms of our democratic society as protected under Article 17 of Schedule 1 to the Human Rights Act 1998 ('HRA').

Their report emphasises that this is a working definition, not a legal one. It also recommended treating hateful extremism with as much priority as terrorism.

Establishing that social media companies are liable for the publication of vilification (s124A Antidiscrimination Act (QLD) or hate speech (s18C Racial Discrimination Act) would be an essential step towards lifting the stakes for these companies. Currently, there is no incentive for those companies to meet Australian standards. The harm of hateful echo chambers that dehumanise minority groups to ingroup audiences is public harm not covered by the current OSA and one that the Australian Government cannot leave to victim communities to battle alone.

Previous attempts of policymaking in this area tend to oscillate between very general approaches (e.g., U.K.'sfailed Bill to ban extremist speech in 2015) and specific guidelines, often adopted by platforms, to list the types of hate speech or incitement that will not be accepted. The latter approach misses organisations or websites that serially attempt to socialise individuals towards extremist violence, especially when they skirt beneath the threshold of hate speech or criminal incitement (for example, through disinformation).

Platforms are motivated to assess one piece of material at a time rather than patterns of behaviour over time of hateful online echo chambers. The material relied upon can dehumanise in aggregate over time in ways that are not apparent if assessing each piece individually.

Targeting incitement to violence won't be enough

While it may be tempting to set the threshold higher at incitement to violence, incitement to violence is a problematic and inappropriate threshold here given:

(1) Platforms often demand it poses an imminent threat – creating an impractical evidentiary burden for whole communities targeted by the material. Measuring the 'tipping point' for danger appears only to be workable where extremist violence or genocide has already occurred, and incitement can be retrospectively measured. Imminent harm is more useful in criminal contexts involving threats against individuals.

(2) The most prevalent and pernicuous threats to community safety are not organisations or websites openly inciting, threatening, or glorifying violence but inducing it indirectly through dehumanising materials about outgroups to ingroup audiences.

PROPOSAL D

D. Improving access to justice for complaints against platforms under Anti-Discrimination laws

(1) We recommend that you consider legislating a liability for social media companies who publish vilifying material (as per state anti-discrimination frameworks) and material that contravenes section 18C of the Race Discrimination Act 1975. Platform failure to meet Australian standards on vilification is a failure to maintain user safety.

(2) In online vilification matters, it would be helpful to confer the relevant human rights body power to automatically alert the appropriate digital platform once the complaint is accepted. Suppose the complaint is later upheld and the platform did not remove the content initially. In that case, this could contribute to evidence used by the e-Safety Commissioner to issue penalties to the digital platform, or perhaps the platform would need to share in the costs order against the complainant. This may be a way to accelerate platform accountability and possibly deliver a quicker outcome.

(3) The Committee may also consider whether any criminal or civil standard includes a corporate liability component for platforms that recklessly allow the material to remain online. If this was introduced, the e-Safety Commissioner could be conferred with powers to issue a warning notice to platforms. The notice may be unenforceable, but non-compliance could be used as evidence of corporate recklessness. The threat of prosecution lifts the performance of platforms in managing violent and hateful echo chambers. This idea is based on the approach to managing Abhorrent Violent Material between the e-Safety Commissioner office and AFP.

Qualification

Currently, companies are not incentivised to dedicate resources to monitor their platforms and remove actors that are serial offenders. As bad actors can move from platform to platform, the proposals above are not enough to shift the discriminatory burden currently experienced by racial and religious minorities in combatting a public harm. Further anti-discrimination framework does not lend itself to the size of financial penalties that would be required to prompt systemic change by a digital platform.

RATIONALE

AMAN & ICQ v Fraser Anning case study

AMAN has been successful in the Queensland Civil and Administrative Tribunal (QCAT) in a vilification complaint against former Australian politician Fraser Anning.[4] The Tribunal found Mr. Anning vilified the Muslim community in the wake of the Christchurch Massacre and 141 times in total, citing breaches in the Queensland Anti-Discrimination Act 1991.

[4] In 2016, politician Pauline Hanson and her party experienced a resurgence to Australian politics, focused on Muslims as the targeted outgroup'. Later, she also brought Fraser Anning to the Australian Parliament, who unashamedly socialised white replacement extremist theories, arguing all Muslims, "including so-called moderates" were attempting to conquer western countries through immigration and high fertility rates. After the Christchurch terror attack by an Australian white supremacist, he argued 'the real cause of the bloodshed' was the immigration program in New Zealand that allowed 'Muslim fanatics.' In the 2019 election, Anning made a video outside a Brisbane mosque calling 'Islamification' a 'huge threat' to Australia. That very same mosque endured a vandalism incident within months of this video. 'Remove kebab,' a term calling for the expulsion and murder of Muslims, along with St Tarrant, was graffitied across its front wall. Queensland vilification laws were the only tool we had to resist dehumanising conspiracy theory about our community being propagated to significant public audiences – putting us in real danger.

Facebook removed at least 80 public posts; however, they refused to disband his two public pages. Mr. Anning appears to be living in the United States. Mr. Anning did not participate at all in the proceedings despite repeated attempts to make contact. QCAT also ordered Mr. Anning to stop vilifying our community further. However, these public pages had continued to vilify Muslims from when we catalogued the offending material. We have no indication that Mr. Anning will respect the court orders.

It would be against the public interest to require a vilified community to lodge a court action to compel these platforms to remove content every time further vilification occurs. Hate speech standards enforced by Facebook generally Facebook's position suggests that they have not found Mr. Anning's pages to violate Facebook policy, as their Terms of Service gives them the ability to disband accounts that serially violate their Community Standards.

We submit it is in the public interest for Facebook to apply Australian legal standards and for Australian authorities to regulate their performance under a duty of care model, rather than leaving it to the community to litigate every artifact of hate speech. Combatting white nationalist and far-right racist pages and groups is a burden for our community that we are facing alone. Vilification laws tend to treat a public harm as a private one (Gelber and McNamara, 2014). It is psychologically exhausting, damaging, and unsustainable. We operate with the help of volunteers.

Case study – Great replacement proponent

A Queensland person was propagating the same ideology as Brenton Tarrant but indirectly by falsely contextualising events and supplying a steady stream of disinformation to a cultivated online audience. He was able to exponentially increase his audience through a Facebook page that he administered. The intense disgust demonstrated in these user reactions shows that this actor has successfully incited hatred, including explicit dehumanising slurs and violent fantasy.

In June 2019, the Facebook page shared a poster with a picture of a white family with two children and a Muslim family with 4 wives and 12 children. It had the same title as Tarrant's manifesto: "The Great Replacement". The meme was accompanied by similar derogatory statements implying that Muslims plan to conquer countries like Australia through higher fertility rates. The intense reactions to this poster were revealed in the extensive comments, with a significantly high proportion employing explicit dehumanising language, as well as expressions of wanting to kill or see Muslims dead. Responses included:

> 'Shoot the ', 'Islam is a cancer on global society for which there is no cure', 'You import the 3rd world you become the 3rd world. And when they become the majority then what next? They won't have whitey to leech off. Just

like locusts, infest & strip everything until there is nothing left', 'Deport the PEDO crap', 'They breed like rats','Drown em at birth', 'Fun those scumbags.muslums....reminds me of aids', 'Society should start culling the Muslims', 'I think I now understand why during the serbian / croat the serbs culled the women', 'I'm going out tonight to do as much as i can to solve this problem'.

However, after collecting evidence, we decided against lodging a vilification complaint. We were deterred by the costs (time and expense) and the likelihood that he may use this action, over the year or two it takes to resolve, as a platform to present himself as a martyr and gain more followers.

The Queensland Human Rights Commissioner has also recognised that the Anti-Discrimination conciliation-based framework cannot deliver the safest or most appropriate process (or outcomes) in many cases where the respondent is unwilling to engage or conciliate.

Instead, we decided to test the federal criminal law for using a carriage service to menace, harass or cause offence (s474.17 Criminal Code 1995 (Cth). This law has been used to protect individuals who are the victims of online racist hatred when individually targeted, but not to protect a person who is a member of a targeted community (for example, Muslims in general). The Australian Federal Police advised this law was not appropriate for the problem.

Despite the Christchurch massacre, Oslo massacre, Myanmar genocide, and other atrocities, Facebook does not treat demographic invasion conspiracy theories about Muslims as violence-inducing (like it has for Q-Anon), nor has it instituted a hateful stereotypes policy to pick up these theories (compare to hateful stereotypes policy it has introduced to protect Jewish communities and Black communities).

Muslims are one of the most targeted outgroups by white nationalist and far right groups online. Anti-Muslim movements were the predominant force behind the growth of organised white supremacy and Neo-Nazi movements in Australia. The threat to safety does not only affect Australian Muslims but all Australians. Australians are looking to the Australian Government to connect our anti-discrimination framework to the online sphere without burdening communities with the task of taking on social media giants.

Making and running a complaint takes great resources, time, and courage for a community advocate. The fear of repercussions can be strong and prohibitive, leading to vulnerable groups further retreating and not exercising their rights.

PROPOSAL E

E. Transparency

(1) Social media companies do not provide adequate insight into how their News Feed curation algorithms work and how efforts to demote harmful content have impacted the distribution of such content on the service.

(a) Platforms must clearly define terms such as “demote” and “amplify,” which are often used when discussing the platform’s “reduce” approach to tackling harmful content. Additionally, platforms must also clarify what they mean when they says they “promote” certain types of content, such as information from authoritative sources (Singh, 2021).

(b) Platforms must outline the types of content that their News Feeds prioritise. As whistleblower Frances Hougan outlined in the last month, in 2018, Facebook changed the configuration of its News Feed algorithms to prioritize “engaging” content. This change has resulted in some sensationalist and harmful content appearing higher in and consumed more in the News Feed. While it may not be feasible—or even helpful—for the company to disclose a complete list of the signals it considers when ranking content in the

News Feed, a company should provide transparency around which of these signals have the most influence over how content is presented to users on the News Feed. Ranking and recommendation algorithms can significantly impact the content users see and engage with, and therefore how they see the world (Singh, 2021).

(c) Companies must provide more transparency around the impact of its content demotion efforts. The platform regularly promotes the "reduce" approach when discussing how it combats the spread of COVID-19 misinformation and election disinformation. However, the company has provided very little data to demonstrate that its efforts to reduce the distribution of such content on the News Feed have succeeded in preventing the consumption of this content and decreasing the harmful effects on their services. This data is critical for demonstrating accountability and for justifying the use of the "reduce" approach (Singh, 2021).

(d) Recent leaks of Facebook internal documents (Bennett and Nguyen, 2021) and Facebook's oversight board confirmed that Facebook has a list of actors that it excises from its policies due to their profile, newsworthiness, or engagement ratings. We suspect this is why a number of key anti-Muslim hate actors propagating

disinformation and demographic invasion conspiracy theory are supported by the platform. Platforms should be required to be transparent about the names of the actors or organisations they excise from their policies and the reasons for the exception.

(e) We support the views of AccessNow that any use of automated tools has to be based on clear and transparent policies, including transparency mechanisms for the independent assessment of their creation, functioning, and evaluation. Platforms should abstain from practices aimed at "nudging," influencing, or manipulating users without their knowledge or consent. The use of content curation technology, such as news feed hierarchization or recommendation algorithms, should be made as clear and transparent as possible. In both automated moderation and content curation, platforms should:

(i) Make automated systems as transparent as possible.

(ii) Publish information about how these systems are used and the procedures behind their application.

(iii)Make the systems available for independent auditing.

(f) In addition to the AccessNow points above, we propose that it should be transparent whether each platform allows

(a) human choice and control over recommendation and ranking algorithms[5] and

(b) user choice and control over recommendation and ranking algorithms.

(2) However, we also support the view that measures to provide algorithmic transparency will only benefit the public if we know how platforms monetise negative amplification. The Global Disinformation Index writes,

> There is no need to audit the details of the recommender system algorithms, for we already know what they are designed to do: The purpose of the algorithm is to maximise the chances of being able to sell the largest number of ad spots by putting the most engaging content in front of us at all times. And much research has shown that negative, hate filled, fear inducing content is much better at keeping us hooked than straight news or even kitten pics. So while the business model of technology companies remains primarily ad funding, the algorithms they design will "solve for engagement," and the content of choice will be toxic, divisive and disinformation…
>
> The final lever, monetisation, is a powerful one. Without the reward of advertising or other forms of monetisation as the end result, algorithms may well be trained in less damaging ways for the human brain. Advertisers have a right to choose which content their adverts fund. Currently in both the open web and closed social platforms, they have limited control over where their ads end up. Efforts underway across the online advertising system to improve

[5] A recommendation of Twitter. See Twitter, *Protecting the Open Internet: Regulatory Principles for Policy Makers.*

> both transparency and choice for the advertiser while also protecting privacy for the citizen, are welcome... (Melford and Rogers, 2021)

Thus, we propose that transparency reports include the degree of control and visibility that advertisers have in determining where their ads are placed. We also recommend that the Australian Government include the Global Disinformation Index in ongoing policy dialogue.

(3) We propose that social media companies also provide full disclosure of advertisers who have used their platform to advertise to Australian audiences, and the country of origin of those advertisers. Requiring information on their advertising spend would also help the Australian Government to understand the market share of social media companies. Further, not easily knowing who these advertisers are is a significant obstacle to important consumer led conversations.

(4) Further, we propose that social media companies should be required to keep a register of proposals or requests for policy changes that

(a) have been received by Australian civil society or Government, including identifying the organisation or agency it has come from
(b) have been made following requests from Australian civil society or Government to their policies.

It will aid civil society to know who else is engaging with social media companies. Changing policies is extremely challenging because of the power imbalance. Stakeholders are rarely connected or know what else is being proposed. There is a public interest in the policies that social media companies apply. It will also give the Australian Government important information about how responsive social media companies are to the community.

(5) We propose that social media companies be required to report on their measures to

(a) Advance racial justice, including through reducing the burden on communities that most experience racism.

(b) Advance environmental justice, including through assisting society to move towards net zero carbon emissions.[6]

This would enable a more straightforward comparison across platforms of the strength of their efforts for consumers and advertisers. However, addressing market monopolies is also a critical part of enabling market forces. Without this, this recommendation won't take us far.

[6] For example actions that could be taken: https://adassoc.org.uk/ad-net-zero/

(6) Minority community experts on the discriminatory effects of disinformation, hate speech, and algorithmic bias must be included in the governance arrangements to draft Transparency Report templates and the e-Safety Commissioner's review of the Platforms' transparency reports. As the Global Network Initiative (2020) Report on Content Regulation and Human Rights has stated,

> civil society actors continue to provide constructive and often prescient advice drawn from the real-world experiences of the most vulnerable and marginalized communities. Processes for legislative deliberation should therefore be open and non-adversarial, drawing on broad expertise to ensure results are well thought out and evidence-based. Unelected regulatory or oversight bodies should also prioritize transparency and consultation with diverse constituencies.

We acknowledge there are different ideas from civil society, like those outlined by the Tony Blair Institute for Global Change (Bennett and Nguyen, 2021), the views put across by the Christchurch Call Advisory Network members, and the GIFCT working group members. However, internally within Australia, policy discussion between civil society and government is scarce.

PROPOSAL F

F. Anti-trust Legislation

The Australian Competition and Consumer Commission (ACCC) must address social media company monopolies for advertising space. It has been reported that ACCC chair Rod Sims says he is considering asking the government for extra powers to help tackle the dominance of tech giants such as Google, Apple, and Facebook.

We propose that the ACCC take an active role in disrupting market dominance on advertising for all social media companies. The problem is that ordinary market forces that would drive advertisers to choose a platform that does a better job of upholding human rights have been eliminated. Consumers have high expectations of big brands regarding corporate social justice, but this does not translate into pressure on the platforms they use to advertise.

CONCLUSION

We want to help the Australian Government contend with hate speech, disinformation, and dehumanisation of minorities in the Online Safety Act (OSA). Currently, the OSA empowers the e-Safety Commissioner to act on abuse where it targets an individual. Still, it does not contend with hateful echo chambers that endanger segments of the community and Australia as a whole.

Regulation is required to counter bad actors online who nurture hateful echo chambers against minority communities. Potent vectors of harm are purposed information operations or actors that serially publish dehumanising language or discourse. This type of material is readily apparent and assessable by an administrative body like the e-Safety Commissioner through a notice and action model, with the protection of judicial review. This proposal recognises that platforms, Government, and civil society mutually benefit from precise and clearly defined public laws to reduce the risk of being overused or weaponised.

Enforceable standards assure Australians that their place in the community *matters*. The Australian Government must do this to protect freedom of expression for all.

The Government's failure to regulate

- emboldens perpetrators to carry out hateful abuse, harassment, threats, assault, and vandalism in public places.

- emboldens perpetrators to target online users from those communities in public threads and private messages.

- encourages far-right networks and mainstreams and legitimises their standing to broader audiences, posing a risk to Australians.

- places a discriminatory burden on minority communities who must 'defend' that they are human, litigate and battle a public harm alone.

Further proposals contained in this document are designed to create the conditions for platform accountability using different and proportionate levers. Moving forward, we ask you to include civil society like AMAN alongside law enforcement and researchers in policy dialogue.

References

Abdalla, M., Ally, M. & Jabri-Markwell, R. Dehumanisation of 'Outgroups' on Facebook and Twitter: towards a framework for assessing online hate organisations and actors. *SN Soc Sci* **1,** 238 (2021). https://doi.org/10.1007/s43545-021-00240-4. Available at: **https://rdcu.be/czgAo**

AccessNow (2020) *26 Recommendations for on Content Governance: A Guide for Lawmakers, Regulators, and Company Policy Makers.* p 42. Available at: https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf

AMAN Statement, *Australian Code of Practice on Disinformation and Misinformation*, 22 February 2021, Available at: http://www.aman.net.au/?page_id=653

Bennett, A., Nguyen, M. (2021) "Two Ideas to Make the Facebook Papers a Turning Point for Accountability in Tech", 27 October 2021, Tony Blair Institute for Global Change. Available at: https://institute.global/policy/two-ideas-make-facebook-papers-turning-point-accountability-tech

Global Network Intitiative (2020) *Content Regulation and Human Rights: Analysis and Recommendations.* Available at: https://globalnetworkinitiative.org/wp-content/uploads/2020/10/GNI-Content-Regulation-HR-Policy-Brief.pdf

Kaldor, S., '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.

Gelber, K., and McNamara, L. (2014) 'Private Litigation to Address a Public Wrong: A Study of Australia's Regulatory Response to "Hate Speech"', Civil Justice Quarterly 33 (3): 307-334.

Melford, C., and Rogers, D., (2021) "Want Less Awful Content? Stop Focusing on Content Moderation". Available at: https://disinformationindex.org/2021/07/want-less-awful-content-stop-focusing-on-content-moderation/

Haslam, N. (2006) Dehumanization: an integrative review. Personal Soc Psychol Rev 10:257

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.

Peucker, M., Smith, D., Iqbal,M. 'Mapping Networks and Narratives of Far-Right Movements in Victoria' (Project Report, Institute for Sustainable Industries and Liveable Cities, Victoria University, November 2018.

Risius, M, Blasiak K, Wibisino S, Jabri-Markwell R, Louis W (2021) *Dynamic Matrix of Extremisms and Terrorism (DMET): a continuum approach towards identifying different degrees of extremisms. Report to the Global Internet Forum to Counter Terrorism.*

Singh, S (2021) "Facebook Releases Information on Algorithmic Content Ranking, but More Transparency Is Needed", *New America,* 19 October 2021. Available at: https://www.newamerica.org/oti/blog/facebook-releases-information-on-algorithmic-content-ranking-but-more-transparency-is-needed/

United Nations Human Rights Committee, General Comment No 34, CCPR/C/GC/34, 12 September 2011.