Policy submission to Digi on the Australian Code of Practice on Disinformation

24 November 2020

Introduction

The Australian Muslim Advocacy Network (AMAN) welcomes the opportunity to input on the proposed Australian Code of Practice on Disinformation. We look forward to engaging with DIGI, digital platforms, ACMA and relevant government departments and regulators to enable considered regulation that protects Australian institutions, citizens and democracy.

We welcome the recognition in the Code's preamble of the need for collaboration and cooperation among relevant stakeholders, including civil society.

Given the complexity of this fast-evolving field, the Australian Government looked to benefit from the expertise of the tech industry in developing this voluntary code before taking further advice from ACMA on next steps, including whether further regulation is needed.

However, this draft needs considerable work to operate as a judicable Code. Also as it stands, this Code has the effect of diluting existing obligations of social media companies under international law.

AMAN strongly suggests that Digi, and ACMA, reflect seriously on amendments suggested in this submission, Reset Australia's submission and through this consultation process.

Improvements to the Code

AMAN proposes the following amendments.

The definition of disinformation

'Inauthentic behaviour' has been made a threshold component of disinformation. According to the Code, it includes 'spam and other forms of deceptive behaviours (including via automated systems) which encourages users of Digital Platforms to propagate content which may cause Harm. We would like to verify whether actors that use non-automated deceptive behaviours, such as producing and disseminating 'pseudo-news', even if they don't rely on bots or fake accounts, would fall within the remit of this definition.

The Code also specifies that 'Disinformation does not include… misleading advertising, reporting errors, satire and parody, or clearly identified partisan news and commentary'.
Misleading political advertising should not be excluded along with misleading advertising.
We also seek clarity on the definition of 'clearly identified partisan news and commentary'. In particular, does this refer to commentary from politicians? Does it exclude far right news platforms on the basis that such conspiracy theories are merely their political opinion?

The proposed definition also excludes content that would otherwise be unlawful. Hate speech is unlawful in Australia on the basis of race (*Racial Discrimination Act 1975 (Cth*) s18C) at the national level, and on various protected bases in state jurisdictions, including on the basis of a person's faith. However, this definition of hate speech is not adopted by social media companies, which by and large are based in the United States. Ideally, the forthcoming Online Safety Bill will help to bridge the gap between Australian law and platform policy on hate speech, to stop this grey area from enabling unaccountability. However, if it doesn't bridge this gap, and this Code also excludes unlawful content from its ambit – that will leave groups targeted by hateful conspiracy theory with little hope for systemic improvements.

Recommendations

1. Insert 'and non-automated' after 'automated' in 'spam and other forms of deceptive behaviours (including via automated systems) in the definition of Inauthentic behaviour

2. Insert '(excluding misleading political advertising)' after 'misleading advertising' in the list of items that Disinformation does not include.

3. Insert 'A site cannot be identified as partisan news and commentary if it has no enforceable editorial standards that align with Australian editorial standards for print and broadcasting news and opinion.'

4. Clarify the distinction between conspiracy theory about a group or member of a group on the basis of race, religion, ethnicity, sexuality, national origin (or other characteristic) and "partisan news or commentary".

5. That the definition of disinformation be considered in conjunction with the forthcoming Online Safety Bill, including its proposed Basic Online Safety expectations, to make sure there is no gap in regard to requirements for stopping hateful conspiracy theories and disinformation about groups on the basis of protected characteristics.

T

The definition of harm

The Code defines Harm as 'imminent and serious threat' to 'democratic political and policymaking processes'; or 'public goods such as the protection of citizen's health, the environment and security'.

The policy rationale for the word 'imminent' is not explained in the discussion paper or the Code. That rationale cannot possibly connect to the aims of safeguarding Australians' health and democracy.

It is suggested that the word 'imminent' be deleted as while some threats posed by conspiracy theory are imminent (for example acts of violence or extremism), this ignores the 'creeping threat'[1] to

[1] 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.

democracy posed by disinformation propagated about 'out-groups' via extremist networks. Research into the nature of 'dangerous speech' that has laid the foundations for historical atrocities and genocide also reveals the critical long lead-up role of disinformation[2]. It also has consequences for longer-term crises, such as climate change.

An imminent threat means a threat that is happening soon, or at any moment. Opacity about the measurements and data used by platforms to identify which threats are imminent is a concern. Given the very poor rates of hate crime reporting and inconsistent classifications in Australia, and the subsequent lack of data, it is reasonably foreseeable that platforms won't ever prioritise conspiracy theories targeting minority groups in a consistent manner. Even the occurrence of the Christchurch massacre has not prompted platform action on anti-Islam conspiracy theories.

The international legal principles, the *Siracusa Principles*, provide that any derogation or departure from fundamental rights and freedoms under the ICCPR, like freedom of speech, be the least intrusive but necessary option to securing those human rights. However this does not mean that only imminent threats should be considered. The code's current definition of harm therefore dilutes platform obligations under international law, which include scrutinising all serious threats to fundamental human rights enabled or facilitated by their services.

Recommendations

6. Delete the word 'imminent' from the definition of harm.
7. Require that consistent metrics of harm be transparently developed, including with input from civil society representing groups most marginalised and endangered by disinformation, and that implicated data be made available for public scrutiny by researchers.

Establish Standards

This Code should aim to provide *standards* which can act as a benchmark for adjudication. The guiding principles contained in this Code should be reframed as standards, with the opportunity for platforms to apply to 'opt-out' of certain standards from an independent code administrator where those standards are clearly not relevant or applicable to their service.

Recommendations

8. Reframe the guiding principles as standards that can act as a benchmark for adjudication.
9. Reverse the 'opt-in to specific measures' process outlined in this Code so that platforms sign up to Code, but then apply to the independent Code administrator to 'opt-out' from certain standards on the basis of relevancy.

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

Safety by design as a judicable standard:

By and large the guiding principles included in the Code are welcomed, especially the way it has articulated 'freedom of expression' to include 'the importance of enabling diverse perspectives and voices to be heard in political debate'. Using the words of HOPE not hate, 'This regulation instead should aim to *maximise* freedom of speech online for more people, including those from minority backgrounds whose speech is consistently marginalised online and elsewhere.'

However, there is no clear principle articulated in relation to safety and security of person, despite this being a major threat posed by disinformation.

It is suggested that the Code include 'safety by design' as a principle and measure, as it is clearly and appropriately emphasised by Australia's e-Safety Commissioner and Australian Government. This principle provides that platforms should not focus on achieving users' safety by 'bolting on' extra protections, but "address online harms, alongside user safety and rights, in the product development lifecycle so that safety is systematically embedded into organisations' culture and operations" (e-Safety Commissioner).

The submission from Reset Australia, an independent, non-partisan organisation committed to countering digital threats to democracy, has conveyed that there should be 'ongoing and proactive auditing of the content that algorithms amplify to users, focusing on the spread of disinformation and its impacts', and pushed for 'an industry-wide framework for assessing the risk for harm of disinformation and formalise guidance on mapping specific measures to these risks'.

As Evelyn Douek, an Australian expert researcher based in the US who recently presented an Australian Parliamentary Inquiry on Foreign Interference in Social Media said:

> It may be that we need far more radical reforms than individual CIB-hunting operations (reforms centered on transparency and changing algorithmic amplification) to make sure public discourse isn't exploited and manipulated in corrosive ways.[3]

Recommendations

10. Introduce an ad judicable standard in relation to taking all reasonable measures to safeguard communities from disinformation in the design of their platforms.

Not profiting from disinformation

While the Guiding principles point to the need to scrutinise advertising placements, this is only one way that platforms profit from disinformation campaigns. It is recommended that signatories to the Code articulate a principle of not profiting from disinformation more generally. Reset Australia has also recommended that the measures include more specific pathways to 'effectively disrupt the economic drivers of disinformation'.

[3] Evelyn Douek, 'What Does "Coordinated Inauthentic Behavior" Actually Mean? There's no clear definition, and that's worrisome', *Slate*, 2 July 2020.

Recommendations

11. Introduce a judicable standard in relation to taking all reasonable measures to not profit from disinformation.

Public disclosure about intervention requests

Whether it is the Australian Government, or another nation state, or segment of the community that is making a request of a digital platform to intervene in particular disinformation, and carrying out this request would adversely affect a group or member of a group on the basis of a protected characteristic, platforms should have a duty to make this request public before intervening. Essentially, platforms should be given to who can make requests to intervene in disinformation, and the degree of public disclosure that goes with it.

Recommendations

12. Introduce a judicable standard regarding public disclosure about intervention requests that platforms consider or plan to act upon that could adversely affect a group or member of a group on the basis of a protected characteristic.
13. Review the standards with regard to what actions or plans by digital platforms should be afforded public disclosure to balance power between powerful and less powerful actors, afford natural justice and uphold the integrity of the Code administrator and Code.

Excluded Services and Products

Under the 'scope, application and commencement of the code', the Code says that the following are not Services and Products subject to this Code:

> 'D) content including e-books, videos, films, television, radio broadcasts or podcasts that is provided for the purpose of entertainment of education'

The necessity of this exclusion is not justified in the documents provided, which is unusual given the risk posed by disinformation contained in videos, films, radio and podcasts, even if they are self-labelled as 'entertainment' or 'education'. For example, the impact of this exclusion on Youtube's responsibilities under the Code or on any platform hosting video or podcast application content needs clarification.

A significant amount of white nationalist/supremacist and ethno-nationalist content is present on Youtube and other video hosting platforms. Social media companies are used by hate organisations to amplify their content to new audiences, and there are ways in which malicious or dangerous third party URLs can be excluded by platforms. Twitter introduced a policy specific to this in mid 2020.

Recommendations

14. Clarify exclusion (D) in relation to 'excluded services and products' so that it is clear that video and audio content shared by digital platforms, including that hosted by external third party sites, are not excluded by this Code.

Code administration

We agree with Reset Australia's view that 'the third-party organisation chosen to be the Administrator of this Code must be independent, objective and prioritise public interest' and their recommendation 'that the sub-committee established by the Code Administrator to monitor and review the actions and committees of Signatories be made up of diverse range of representatives'.

We also agree that the Code would be greatly strengthened by working with its subcommittee, relevant stakeholders and ACMA to develop a 'common reporting structure and shared KPIs that will be able to adequately monitor the implementation of Signatories commitments and measures'. Lessons need to be learned and applied from the European experience, and thought should also be put into what streamlining can occur with data transparency reports mandated under the forthcoming Online Safety Bill.

Recommendations

15. The third-party organisation chosen to be the Administrator of this Code must be independent, objective and prioritise public interest.
16. Mandated transparency reports under the forthcoming Online Safety Bill should include data connected to harm metrics referred to in Recommendation (7) of this paper and data on the algorithmic amplification of disinformation.

Complaints handling

A complaints adjudication process should be spelled out in this Code, rather than signposted for introduction within 6 months.

Methods of adjudication and consequences for failure to comply with the Code are essential to ACMA being able to develop meaningful advice to the Australian Government on the likely effectiveness of the Code, which has been requested by mid 2021. According to the timeframe, ACMA may have no detail on a complaints mechanism to consider. As a matter of good drafting, the standards need to be considered in the context of what remedies complainants can achieve through lodging a complaint.

Recommendations

17. Articulate a complaints process in the Code in its first iteration.

Who is AMAN?

AMAN is a national policy development and advocacy body dedicated to securing the physical and psychological welfare of Australian Muslims.

Our objective to create conditions for the safe exercise of our faith and preservation of faith-based identity, both of which are under persistent pressure from vilification, discrimination and disinformation.

Online harms experienced by Muslims and their broader implications

In the last federal election, there were approximately 12 fringe parties running with a discriminatory anti-Muslim policy – this is the largest number of groups that we have recorded. We remain very concerned about the exportation of RWE rhetoric from the UK, Europe, Canada and USA to Australia through disinformation and conspiracy theory campaigns on social media platforms, and its potentially devastating impacts for Australia's democracy, social cohesion and national security.

These conspiracy theories were used by Tarrant to justify his terror attack on two Christchurch mosques, resulting in the deaths of 51 Muslim men, women and children. Breivik cited the same conspiracy theory about Islamic demographic invasion and replacement when he murdered 77 non-Muslim people, mostly teenagers, in Oslo Norway in 2011. There have been recent examples in Australia where right-wing extremists have been radicalised by these theories to the point of threatening, and planning terror attacks on mosques or 'left-wing' premises of the general population. Conspiracy theories about groups on the basis of religion or race, is a more implicit but highly effectual form of attack, especially in terms of dehumanisation. However it tends to not be detected under hate speech protections by digital platforms.

The host sites of disinformation that falsely contextualises or connects contemporary events with overarching conspiracy theory appear to not be prioritised for or fall within the ambit of fact-checking as they are often 'pseudo-news' or blogs. Platforms will focus on false stories that achieve volumetric high rates of sharing, at the expense of considering cumulative disinformation aimed at comparatively smaller captured audiences, in order to prime them to accept extreme worldviews that deny the moral worth of designated 'out-groups'.

It also remains unclear as to how far platforms' definition of 'opinion' is protecting these external sites from fact-checking, or from flagging as an unreliable or harmful source. For example, platforms are treating 'counter jihad' ideology and conspiracy theories – including propagating that

(a) personal religiosity in Islam leads to sub-human behaviour and extremism.
(b) Muslims/Islam are invading the West to take over through immigration and high fertility rates.
(c) Islam/Muslims are waging violent war with the West/clash of civilisations.

as a form of political or partisan discourse – whereas based on CVE-field research, it is clear that these theories dehumanise Muslims, are inaccurate[4] and mislead the public. Proponents of these theories build support for far right nationalist activism and violent extremism. A high degree of volatility in moving towards violence has been observed in the far right milieu[5], with slippage between offline 'anti-Islamisation' events and online white supremacy also recorded in Australia[6]. It's connection to extremism by ideologically motivated Muslims is also just beginning to be understood.[7]

Facebook has taken action on conspiracy theories where they've passed a certain threshold (measured by Facebook) to be 'violence-inducing'. For example in October 2020, Facebook amended their platform to ban violence inducing conspiracy networks such as Q-Anon. It has also acknowledged in August policy amendments that 'harmful stereotypes' can lead to real world violence, and listed one pertaining to the Jewish community that has fuelled white supremacist networks. However, it continues to remain silent on conspiracy theories targeting and harming the Muslim community, even in the aftermath of Christchurch, other acts of terror in Europe, and numerous examples of anti-Muslim hate crime (Europe, North America, Australia), mob violence (India), alleged cultural genocide (China) and violent genocide (Myanmar).

A recent report by the Institute of Strategic Dialogue found that anti-Muslim hate organisations based in the US were able to organise and fundraise on mainstream social media platforms, along with anti-LQGBTI and anti-immigrant organisations[8]. White supremacist and nationalist organisations were far less able to engage these benefits. The role of anti-Islam conspiracy theory as a gateway for audiences to become socialised to more fringe extremist discourses is established in research. Therefore its role within both the ideological information eco-system and revenue generation for extremist movements demands urgent and concerted attention from digital platforms.

For any clarification or comment, please contact:

Rita Jabri Markwell
Policy advisor
Australian Muslim Advocacy Network
advocacy@aman.net.au

4 Johannes Beller and Christoph Kröger, 'Religiosity, religious fundamentalism, and perceived threat as predictors of Muslim support for extremist violence' (2018) 10(4) *Psychology of Religion and Spirituality* 345; Anne Aly & Jason-Leigh Striegher, 'Examining the Role of Religion in Radicalization to Violent Islamist Extremism' (2012) 35 *Studies in Conflict & Terrorism 12;* Anthony Cordesman, 'Islam and the Patterns in Terrorism and Violent Extremism' (Center for Strategic & International Studies, 17 October 2017).

5 Mario Peucker, "Should we stop referring to some extremists as right-wing?", *ABC Religion and Ethics*, 20 October 2020.

6 Mario Peucker, Debra Smith and Muhammad Iqbal, 'Mapping Networks and Narratives of Far-Right Movements in Victoria' (Project Report, Institute for Sustainable Industries and Liveable Cities, Victoria University, November 2018), 11.

7 Tahir Abbas, 'Far Right and Islamist Radicalisation in an Age of Austerity: A Review of Sociological Trends and Implications for Policy', ICCT Policy Brief (International Centre for Counter Terrorism – The Hague, January 2020).

8 Institute for Strategic Dialogue and Global Disinformation Index (2020) *Bankrolling Bigotry: An overview of the online funding strategies of American hate groups.*