ONLINE SAFETY ACT NETWORK

Submission to the Australian Department of Infrastructure, Transport, Regional Development, Communications and the Arts consultation regarding the

Online Safety (Basic Online Safety Expectations) Determinations

February 2024

We are grateful for the opportunity to submit to this consultation. We should be happy to provide more evidence would that be helpful. This submission draws on previous work submitted by Carnegie UK to the Australian House Select Committee on Social Media and Online Safety in 2022.[1]

Summary

- Australia's very early lead in online safety regulation set an example for other countries. As others have caught up with Australia we now know far more about online harms, how social media companies work and the tools available to governments to protect citizens.
- The Online Safety Act Network brings together civil society advocates, researchers and campaigners with an interest in the effective implementation of the UK legislation. The Network builds on the work previously undertaken by Carnegie UK. Our experience in the UK suggests there is an opportunity for Australia to protect more citizens, more effectively by imposing a statutory duty of care on tech companies to keep people safe. This would build on the strong foundations of the eSafety commissioner, the *Online Safety Act* and the ACCC.
- A statutory duty of care would require companies to design safer platforms and implement safer processes and systems to run them. Much like any other hazardous industry, social media companies would have to perform risk assessments under regulatory supervision.
- This approach requires a well-resourced, informed and steely regulator and a mechanism to ensure that companies do not only write codes of practice that suit them and do not really achieve policy objectives.

[1]See https://www.aph.gov.au/Parliamentary_Business/Committees/House/Former_Committees/Social_Media_and_Online_Safety/SocialMediaandSafety/Submissions

Online Safety Act Network

The Online Safety Act Network (the 'Network') is committed to keeping advocates, researchers and campaigners informed and connected during the UK's *Online Safety Act*'s implementation. The Network builds on the expert advisory and convening power established by Carnegie UK during the development and Parliamentary passage of the Online Safety Bill. The Network is led by Maeve Walsh and Professor Lorna Woods, University of Essex, and supported by Reset.Tech.

In this submission we highlight how requiring safety expectations over 'selected' designated systems will not fully protect Australians, and explain the duty of care approach. We understand that this comprehensive duty of care approach falls within the remit of the terms of reference for the Australia *Online Safety Act* review,[2] and seek to connect it here to the role of Basic Online Safety Expectations ('BOSE'). Each nation must find its own path to tackling online harm, reflecting its local incidence but, where there is common ground, there may be strength in acting together. We explain the evolution of our work at the end of this paper.

We also note the proposals to address Hate Speech within the BOSE review. We refer to a draft social media code on hate speech that we created for the United Nations Special Rapporteur on Minorities. This draft code is rooted in the ICCPR which underpins much law and practice on freedom of expression in Australia and takes into account the Ruggie Principles on corporate social responsibility.

Reducing harm through better design, systems and processes

The Carnegie UK approach is "systemic". It requires companies to design for safety and run less risky systems and processes—similar to product safety or health and safety requirements for workplaces. It focuses on the systems that make up the social media platform and not directly on the content posted by users. This approach is flexible and more likely to be future proof; as it does not mandate specific solutions or link to particular technologies (either in terms of identifying problems or solutions), there is a reduced risk that the regime will become outdated.

This approach recognises that the platforms are synthetic environments created by platform operators and that they are not neutral as to how people discover and create content. Choices made by the platforms about how they design their services affect the content seen (e.g. default to autoplay, curated playlists, data voids[3] and algorithmic promotion) and even the content produced (e.g. through financial

[2]DITRDCA *Terms of Reference – Statutory Review of the Online Safety Act 2021* https://www.infrastructure.gov.au/sites/default/files/documents/tor-statutory-review-online-safety-act-2021-8Feb.pdf [accessed 14 February 2024]

[3]A data void – a search term for which there is no content can be exploited by disinformation actors by encouraging people to search for a formerly void term and then placing disinformation there. See Michael Golebiewski and danah boyd for the role in radicalising Dylann Roof. https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

incentives for content creators, or the feedback loop created through metrification; platform-designed emojis can create a new shorthand for communication).[4]

Focussing on platform systems and processes allows a greater range of possible interventions that are human rights compliant. In general, the systems-based approach is neutral as to the topics of content. Under a systems approach most interventions allow speech to continue, but could:

- affect its visibility (e.g through changes to a recommender algorithm that stop some content being aggressively promoted, switching off autoplay),
- limit the speed or extent to which material spreads (e.g. through limiting the number of people to whom one message may be forwarded), and
- even influence the manner in which the message is expressed (e.g. through 'did you mean to send that' prompts or delayed sending allowing retrieval or regular reminders as to rules relating to harassment and hate speech). So, United Nations Freedom of Expression Rapporteur Irene Khan suggested that it may be appropriate to use systems based measures such as downranking, demonetizing, friction, warnings, geo-blocking and counter-messaging than simply blocking things.[5] Systems-based interventions may allow potentially conflicting human rights of the many platform users to be more optimally balanced than would be the case in a regime in which the only response is to take content down.[6]

A systems-based approach 'system' has a double meaning. First, it refers to the software and business systems, and the fact that they are the focus of attention under this approach. While questions of content inevitably arise, they are dealt with indirectly. Such an approach does not, however, displace content rules. There are systems concerns here too. A service provider may have a policy prohibiting hate speech, but it might choose to run the platform in such a way that the policy is not enforced effectively: a weak system undermines the policy.

We note that the proposals for reforms to the BOSE lists a number of systems that would become subject to basic safety expectations if the proposals are adopted. This is welcome, however listing designated systems will inevitably create gaps. All systems must be subject to duties of care.

---

[4]Anne Wagner, Sarah Marusek and Wei Yu 'Sarcasm, the smiling poop, and E-discourse aggressiveness: getting far too emotional with emojis' (2020) 30 *Social Semiotics* 305 DOI: https://doi.org/10.1080/10350330.2020.1731151; there are additional issues around differential understanding of emojis potentially exacerbated by different 'fonts' used by different platforms.

[5]Irene Khan, Public Comment by UN Special Rapporteur on Freedom of Opinion and Expression Irene Khan on Facebook Oversight Board Case no. 2021-009, 9 September 2021, available: https://www.ohchr.org/Documents/Issues/Opinion/Legislation/Case_2021_009-FB-UA.pdf [accessed 21 September 2021].

[6]L. Woods, The Carnegie Statutory Duty of Care and Fundamental freedoms, December 2019, available: https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/12/05125454/The-Carnegie-Statutory-Duty-of-Care-and-Fundamental-Freedoms.pdf [accessed 21 September 2021]; Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, (A/74/486), 19 October 2019, para 51, available: https://www.undocs.org/A/74/486 [Accessed 22 July 2021].

Secondly, the approach requires each business to introduce a system for risk assessment, risk mitigation and reparation. This challenges companies which seek to operate on the basis of 'naive innovation' or wilful blindness. The recent Wall Street Journal reporting reveals documents demonstrating that senior management seemingly chose to ignore issues flagged by employees; this reporting supports earlier claims by civil society actors.[7]

We note that the proposals for reforms to the BOSE do not strictly require risk assessment, risk mitigation and reparation. This is a missed opportunity to ensure systems will be effectively made safe.

Making all systems and processes safer – a duty of care

The pace of change in both technology and behaviour on social media is such that detailed rules tackling specific harm are likely to become outdated or ineffective very quickly. The Carnegie UK approach draws from experience of other areas of safety regulation such as workplace health and safety which in the UK, as in Australia is determined by a duty on the people who control and are responsible for the hazardous environment.

Note, it is not expected that the duty of care will lead to a perfect environment – it cannot solve all problems on the Internet. It may improve the general environment so as to allow more targeted, content focused measures if needed; it can therefore be seen as working in tandem with rules aimed at improving notice and action requirements in relation to specific categories of speech.

The obligation has, in essence, four aspects:

- the overarching obligation to exercise care in relation to user harm;
- risk assessment process;
- establishment of mitigating measures; and
- ongoing assessment of the effectiveness of the measures.

While we propose a general duty, the existence of such a duty does not mean that statute cannot specify specific obligations within the general duty – for example, the need to have an effective complaints mechanism, obligations of transparency for particular issues, the need to take particular steps with regard to specific types of content (e.g. child sexual abuse and exploitation material). A general duty is therefore not incompatible with the existing obligations and current role of the eSafety Commissioner.

The European Union's *Digital Services Act* has taken an approach with similar effect: the DSA requires 'very large online platforms' to show 'due diligence' that its systems and processes do not cause harm.

[7]See e.g. Center for Countering Digital Hate, Malgorithm: how Instagram's Algorithm Publishes Misinformation and Hate to Millions during a Pandemic, available: https://www.counterhate.com/malgorithm [accessed 21 September 2021].

Risk assessment

Assessment of risk to an external, rather than shareholder-led standard is central to reducing harm. Ideally, companies would be required to assess risk continually and then put in place mitigation to reduce harm. This breaks down into a number of aspects:

- define risk (including identification of hazards and likely harms) and understand the consequences;
- evaluate the likelihood;
- identify how the organisation could eliminate, mitigate, control or react to the risk;
- test and evaluate control measures;
- identify where improvement is needed.

When identifying risk and control measures the differential impact on sub-sets of the user group should be taken properly into account.

Risk assessment, management and mitigation to local standards set by democratic governments is accepted practice for global multinationals in hazardous industries. As parliaments determine social media to be a hazardous industry similar methods can be employed, adjusting for the importance of freedom of expression.

An effective, neutral regulator to enforce the duty of care

The UK *Online Safety* Act model involves enforcement of the duty of care to responsibilities of the UK media regulator OFCOM. OFCOM is an independent regulator at arms-length from the Executive. OFCOM received many of their powers at the point of Royal Assent in October 2023 and published their first consultation—on the illegal harms duties —a couple of weeks later. A number of their additional powers have since come into force under a series of commencement orders; one of which is on information-gathering, which OFCOM will start to use to require further evidence and information from regulated services to inform further iterations of the codes of practice.

Under the Act, OFCOM has the powers to levy substantial fines on companies that breach their duties and has a strong track record of defending its work in the courts against global corporations with large legal departments.

If Australia were to choose elements of a duty of care regime then the regulatory ‘type’ required to enforce would be more like the ACCC than the eSafety Commissioner, suggesting a need to grow and restructure the latter.

There will be great strength in regulators around the world working together, as we can already see competition regulators doing in respect of large technology companies, and via the Global Online Safety Regulators network,[8] of which OFCOM and eSafety Commissioner are members.

Hate Speech – draft guidelines for United Nations

We note that the BOSE review is also proposing safety expectations around hate speech on social media. We also note that much Australian law and practice on freedom of expression is derived from International Covenant on Civil and Political Rights.[9] We draw the committee's attention to Carnegie UK's work on social media hate speech working within the norms of international human rights law. Carnegie UK submitted draft guidelines for social media companies on combatting hate speech to the United Nations Special Rapporteur on Minority Issues.[10] The guidelines are a generalised 'systems and processes' approach to the issue designed to be applicable in many jurisdictions where there may not be functioning regulatory systems. The guidelines are based on work done with groups representing victims of hate speech in the UK. The Special Rapporteur has produced draft guidelines for the UN Human Rights Council.[11]

The draft hate speech guidelines provide a practical approach to for social media companies to combatting hate speech compliant with international human rights law. The guidelines could inform thinking on regulation in almost any democracy and in relation to many problem areas (not just hate speech) – we raise them in this consultation for consideration.

Background to Carnegie UK's work

In 2016 Woods and Perrin carried out work with an MP (on the private members bill '*Malicious Communications (Social Media) Bill'*) to try to ensure that social media platforms gave adequate tools to users to help them defend themselves from online abuse. This focus on design features and tools formed the basis for a larger project that Woods and Perrin commenced in early 2018 after the UK Government's Internet Safety Strategy Green Paper in Autumn 2017 detailed extensive harms but few

[8]Office of the eSafety Commissioner 2024 The Global Online Safety Regulators Network https://www.esafety.gov.au/about-us/who-we-are/international-engagement/the-global-online-safety-regulators-network [accessed 15 February 2024]

[9]See for instance Australian Government Attorney General guidance on right to freedom of opinion and expression. https://www.ag.gov.au/rights-and-protections/human-rights-and-anti-discrimination/human-rights-scrutiny/public-sector-guidance-sheets/right-freedom-opinion-and-expression [accessed 15 February 2024]

[10]Published at https://www.carnegieuktrust.org.uk/news-stories/ad-hoc-advice-to-the-united-nations-special-rapporteur-on-minority-issues/ [accessed 15 February 2024]

[11]Published at https://www.ohchr.org/sites/default/files/2022-06/Draft-Effective-Guidelines-Hate-Speech-SR-Minorities.pdf [accessed 15 February 2024]

solutions. Initially published as a series of blogs, the work developed into a public policy proposal to improve the safety of users of internet services through a statutory duty of care, enforced by a regulator.[12] A full reference paper[13] drawing together their work on a statutory duty of care was published in April 2019, just prior to the publication of the UK Online Harms White Paper.[14]

The UK government published both its interim[15] and full[16] responses to the White Paper, with significant shifts in each iteration towards a more systemic approach to regulation of harm that is closer to our model than the initial White Paper version, which was framed around a series of content-based codes of practice. The UK government has now passed an *Online Safety Act* with a strong emphasis on systems and processes regulation and overlapping duties of care on companies to protect users.[17]

Carnegie UK has now completed its programme of work on online harms and no longer has an ongoing interest in this area. Prof Lorna Woods is now an adviser to the Online Safety Act Network.

Maeve Walsh, DIrector, Online Safety Network.

[12]See https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/ [accessed 15 February 2024]
[13]See https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf [accessed 15 February 2024]
[14]See https://www.gov.uk/government/consultations/online-harms-white-paper [accessed 15 February 2024]
[15]See https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response [accessed 15 February 2024]
[16]See https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response [accessed 15 February 2024]
[17]See https://www.legislation.gov.uk/ukpga/2023/50/enacted [accessed 15 February 2024]