Submission to the House Select Committee on Social Media and Online Safety

1. We are grateful to the Clerk for inviting us to submit. We should be happy to give evidence by video link if doing so would be helpful.

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 their citizens.

- Carnegie UK has been at the forefront of work on effective regulation of social media. 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 Australian Competition and Consumer Commission (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.

- Australians, via Parliaments and regulators, should ultimately be in charge of setting rules, not companies.

Carnegie UK

2. The Committee has heard evidence about reducing online safety through a statutory duty of care[1]. Carnegie UK was the first to describe a regulatory system based on a statutory duty of care for social media in 2018[2]. We have worked with civil society, government, civil servants and parliamentarians in many countries to develop this work. The UK government policy is rooted in our approach[3], as is the similar EU Digital Services

---

[1] JABRI-MARKWELL, Ms Rita, Adviser, Australian Muslim Advocacy Network Oral Evidence 22 December 2021
[2] See our full reference paper (2019) which brings together all our work on the subject to that point: "Online Harm Reduction: a statutory duty of care and a regulator", available here: https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/04/06084627/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf
[3] The UK government published its Draft Online Safety Bill in May 2021: https://www.gov.uk/government/publications/draft-online-safety-bill

Carnegie United Kingdom Trust

Registered Charity No: SC 012799 operating in the UK   Registered Charity No: 20142957 operating in Ireland   Incorporated by Royal Charter 1917

Act (specifically the provisions relating to 'very large online platforms').[4] We explain the evolution of our work at the end of this paper.

3. In this submission we explain the duty of care approach to inform the Committee's deliberations. 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.

4. We also refer to a draft social media code on hate speech that we created for the United Nations Special Rapporteur on Minorities[5]. This draft code is rooted in the International Covenant on Civil and Political Rights (ICCPR) which underpins much law and practice on freedom of expression in Australia and takes into account the Ruggie Principles on corporate social responsibility[6].

Reducing harm through better design, systems and processes

5. The Carnegie UK approach is what we have termed "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 focusses 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.

6. 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[7] and algorithmic promotion) and even the content produced (e.g. through financial incentives for content creators, or the feedback loop created through metrification; platform-designed emojis can create a new shorthand for communication[8]).

7. In Carnegie's 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

---

[4]   https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

[5]   See ad hoc advice to UN Special Rapporteur here:
https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2021/07/25105219/UN-Hate-Speech-draft-v.05a-1.pdf

[6]   https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

[7]   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 Dylan Roof. https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

[8]   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.

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.

8. 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.[9]

9. 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-based 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.[10] Such 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.[11]

10. A co-regulatory approach should let the companies find solution to the problems they have created rather than government diktat. But the regulator on behalf of the public needs to be able to set standards for companies or rapidly correct inadequate behaviour.

---

9   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].

10   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].

11   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].

11. Moreover, making the service provider responsible for implementing better systems is economically efficient, consistent with the "polluter pays" principle[12] returning external costs to society into the service provider's production decision. Social media platforms are all consciously designed and run, mainly by people at publicly owned companies. People at those companies make choices and trade-offs in how services are designed and run. Companies should and indeed must favour the shareholder interest within the law. Facebook whistle-blower Frances Haugen made many allegations of senior management being presented with evidence about harm arising from the operation of the platform. But management then made decisions that favoured the shareholder rather than the public interest – even on matters as difficult as child safety or modern slavery. So an external mechanism is required to modify behaviour.

Making systems and processes safer – a duty of care

12. 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. Requiring operators to identify hazards and risk of harm avoids this problem. Carnegie's 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.

13. 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.

14. 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.

15. 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 wholly compatible with the existing obligations and the current role of the e-Safety Commissioner.

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

---

[12] OECD, "Recommendation of the Council on the Implementation of the Polluter Pays Principle", 1974, available: https://legalinstruments.oecd.org/en/instruments/11.

Risk assessment

17. Assessment of risk to an external, rather than shareholder-led, standard is central to reducing harm. In the Carnegie regime 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), 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.

18. 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

19. The Carnegie model in the UK suggested adding the enforcement of the duty of care to responsibilities of the UK media regulator OFCOM. OFCOM has a track record of finding an acceptable balance on difficult social issues relating to media. OFCOM already has broad information gathering powers which would be extended to social media companies. OFCOM is an independent regulator at arms-length from the Executive. It is covered in the UK by human rights legislation that requires it to consider the impact of its work on rights such as speech, the rights of children, freedom from threats of violence, more general personal physical and psychological integrity etc. OFCOM is also bound by a duty to be proportionate in its actions – proportionate to a company size/capability and the risk its activities pose. This guards against over regulation, especially of new entrants. OFCOM also has a substantial research function that allows the regulator to take strongly evidence-based decisions.

20. OFCOM should have powers to levy substantial fines on companies that breach the duty (including by carrying out an inadequate risk assessment, notably by being wilfully blind), the ability to direct companies to take corrective action and in the worst case take measures to block their services from the UK. Many have proposed some form of liability for Directors of social media companies as is found in health and safety legislation and financial services regulation – this is germane when companies are so big they can absorb even very large fines with ease.

21. We are conscious that the regulator will be dealing with companies with very substantial legal resources. OFCOM has a long track record of defending its work in the courts against global corporations with a large, well established legal department. OFCOM's current turnover is £130 million per annum which will grow substantially when it assumes online safety responsibilities.

22. 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 or the Australian Communications and

Media Authority (which has a track record of regulating national and commercial broadcasters and telecommunications companies) rather than the eSafety Commissioner, suggesting a need to grow or restructure the latter.

23. 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.

Hate Speech – draft code of practice for United Nations

24. We note that the Committee has taken evidence from victims of hate speech on social media. We note also that much Australian law and practice on freedom of expression is derived from International Covenant on Civil and Political Rights[13]. We draw the Committee's attention to our 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.[14] 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 will launch his guidelines later this year.

25. The draft hate speech guidelines provide a practical approach for social media companies to combat 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 submit them to the Committee for consideration.

Lorna Woods OBE FRSA (Professor of Internet Law, University of Essex)
William Perrin OBE FRSA (Trustee, Carnegie UK)
Maeve Walsh FRSA (Associate, Carnegie UK)

January 2022

---

[13]   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

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

ANNEX: 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 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[15]. A full reference paper[16] 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[17].

The UK government has since published both its interim[18] and full[19] 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 produced a draft online safety bill[20] with a strong emphasis on systems and processes regulation with duties on companies to protect children and from illegal content. Parliament has just produced a detailed scrutiny report on the draft Bill[21] which draws heavily on Carnegie's work to improve the Bill.[22]

All our work can be found here: https://www.carnegieuktrust.org.uk/programmes/tackling-online-harm/

[ENDS]

---

[15]   https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/

[16]   https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

[17]   https://www.gov.uk/government/consultations/online-harms-white-paper

[18]   https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response

[19]   https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response

[20]   https://www.gov.uk/government/publications/draft-online-safety-bill

[21]   https://publications.parliament.uk/pa/jt5802/jtselect/jtonlinesafety/129/12902.htm

[22]   https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2021/11/10133722/Amendments-Explanatory-Notes-Carnegie-UK-Revised-Online-Safety-Bill-1.pdf