ONLINE SAFETY ACT NETWORK

RESPONSE TO THE AUSTRALIAN GOVERNMENT'S STATUTORY REVIEW OF THE ONLINE SAFETY ACT 2021

Background

1. The Online Safety Act Network has been set up to support and coordinate UK civil society engagement during the implementation of the UK's Online Safety Act 2023. It is led by Maeve Walsh, with Prof Lorna Woods (Professor of Internet Law, University of Essex) as its principal adviser. Both Prof Woods and Maeve Walsh were integral to the work of Carnegie UK in advocating for a statutory duty of care for online harm reduction, an approach first proposed by Prof Woods with William Perrin, a Carnegie UK Trustee, in 2018[1] which went on to form the framework for the UK's Online Safety Act 2023.

2. We note that one of the main areas of focus in the Australian Government's call for input is on whether the Online Safety Act 2021 should be amended to impose on platforms a new duty of care towards their users.

3. A statutory duty of care approach, enforced by a regulator, returns the burden of safety to the company that is responsible for creating the harm in the first place. It forces companies to run safer systems. It is efficient in microeconomic terms and a competent, well-resourced regulator will prevent a company dodging its responsibilities. Australia moving to a duty of care model would then allow great synergies with the EU and UK regulation and thwart efforts by foreign companies to 'pick off' Australian regulators and enforce the will of the Australian parliament.

4. We set out below some thoughts (primarily in relation to questions in part 6 of the consultation) drawn from the Carnegie UK work between 2018 and 2023. While it is too early to talk definitively about how the UK Act will work in practice, we also draw some insights from the work we have done to assess the implementation plans of Ofcom (the UK's designated Online Safety regulator) through the OSA Network since autumn 2023.

[1] All the Carnegie UK work and a list of the main resources supporting the development of the UK Online Safety Act is here: https://carnegieuktrust.org.uk/carnegie-uk-online-safety-bill-resource-page/

5. We provide lots of links to further reference material throughout the submission - and the resource page for the Carnegie UK programme is archived here. We would be happy to provide more information if helpful - either in writing or via a video call at a convenient time - to those working on the Australian review.

Summary

6. This submission sets out detail on the following points.
   - A duty of care approach and safety by design are compatible but they are concepts with different scope (and how they are used may differ depending on who is speaking).
   - The sorts of mitigations may depend on whether someone is risk assessing an existing service rather than a new one.
   - Both fit into a systemic approach to regulation, rather than content-based regulation.
   - While a systemic approach is legitimate in terms of apportioning responsibility for actions, it may still be necessary to rely on content regulation and the enforcement aspect of services' gatekeeper functions.

The difference between content regulation and systems regulation

7. Historically, there has been a distinction between regulation of content (eg broadcasting regulation and, in the context of private individuals, defamation laws) and rules relating to the distribution network, the division relating roughly to knowledge of the content. The early approach to internet regulation could be said to adopt this approach, treating "intermediaries" as content-neutral distribution channels. While the boundary between content and distribution was never as clear cut as this description implies, the categorisation of all intermediaries as neutral became increasingly problematic. Intermediaries can carry out a range of functions.

8. In the initial work on the proposal for a statutory duty of care for online harm reduction by Professor Lorna Woods and William Perrin, these functions were divided into 4 categories:
   - infrastructure and hardware
   - the layer that makes the internet work (protocols and the domain name system)
   - application and e-commerce related services that are delivered to or interact with end-users (business or consumer) to enable the use of the internet – this includes search and social media platforms
   - cross-cutting sectoral services that support end-user services – eg payment services; advertising services and the interactive advertising infrastructure. [2]

9. On top of all this sits content, whether professionally produced/editorial content or user-generated content. While services might have specific primary functions, the ways that

[2] Adapted from Karine Perset, *The Economic and Social Role of Internet Intermediaries* (OECD, 2010), https://doi.org/10.1787/5kmh79zzs8vb-en

they are used can overlap.

10. Some intermediaries are not only in charge of the prioritisation of information users see, but they also encourage and reward certain types of content and behaviours. In both aspects they affect the information environment. Woods and Perrin noted that not only are services not neutral as to content, its impact arises as a result of choices that the services have made: about the terms of service, the software deployed and its design and the resources put into enforcing the terms of service.[3] Summarising their work, they argued that "[s]ocial media service providers should each be seen as responsible for a public space they have created, much as property owners or operators are in the physical world".[4] This responsibility is not directly for content, but for the choices the service makes in designing and operating its service, and the functionalities available to users. This is about the creation of the "synthetic landscapes"[5] – which has links to a "by design" approach (though much depends on what is meant by "by design") – but is not limited to that.

11. In looking to the system and services provided, rather than identifying particular instances of problem content, this proposal is congruent with the recognition of the role of software in constituting online environments and consequently on user choices found in the work of, for example, Lessig[6], Zuboff[7] and the like. Reidenberg argued for a *lex informatica* based on the idea that "[t]echnological capabilities and system design choices impose rules on participants".[8] While users are not just passive dots – and some users actively seek to manipulate systems – it is also important to remember the limits of human rationality[9] and that humans seem to suffer from a number of cognitive weaknesses that some services exploit, whether deliberately or not. Against this background, Perrin and Woods proposed systems-based regulation as a more legitimate and effective mechanism.

[3] Lorna Woods and William Perrin, An Updated Proposal, January 2019, https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/01/06085621/Internet-Harm-Reduction-final.pdf, para 6-7

[4] Woods and Perrin, Internet Harm Reduction: a Proposal, 30 January 2019, https://carnegieuktrust.org.uk/blog-posts/internet-harm-reduction-a-proposal/

[5] Woods and Perrin "Obliging Platforms to Accept a Duty of Care" in Martin Moore and Damian Tambini (eds) *Regulating Big Tech: Policy Responses to Digital Dominance* (Oxford: OUP, 2021)

[6] Lawrence Lessig, "The Law of the Horse: What Cyberlaw Might Teach", (1999) 113 *Harv. L. Rev*. 501

[7] Shosana Zuboff, "Big other: surveillance capitalism and the prospects of an information civilization" (2015) 30 *Journal of Information Technology* 75-89

[8] Joel R. Reidenberg, "Lex Informatica: The Formulation of Information Policy Rules through Technology", (1997-1998) 76 Tex. L. Rev. 553, p 544 available at: https://ir.lawnet.fordham.edu/faculty_scholarship/42

[9] Mark Leiser, "The Problem with 'Dots': questioning the role of rationality in the online environment" (2016) 30 *International Review of Law, Computers and Technology* 191-210

12. System here[10] has two aspects:
    - that the (software) system should be regulated rather than focussing on the content; and
    - that the service should comply with its duties by taking steps to identify hazards and risks and remove them or mitigate them.

    The vehicle that they proposed was a statutory duty of care, although they linked it to a by design approach[11], and noted that such a duty covers both harmful persuasive design and also careless service design that leads to harm.[12] They further noted that "[t]he statutory duty of care approach is not a one-off action but an ongoing, flexible and future-proofed responsibility."[13]

    In terms of deploying a systemic approach, a four-stage model was proposed of the flow of information across a service[14] which sits against and interacts with companies' governance processes and business model.[15]

[10] It is different from systemic risk where there is a possibility that a single event or development might trigger widespread failures and negative effects spanning multiple organisations, sectors, or nations - - see Christopher Wilson, Yamas Gaidosch, Frank Adelmann and Anastasia Morozova, "Cybersecurity Risk Supervision" International Monetary Fund, 2019, https://www.imf.org/-/media/Files/Publications/DP/2019/English/CRSEA.ashx

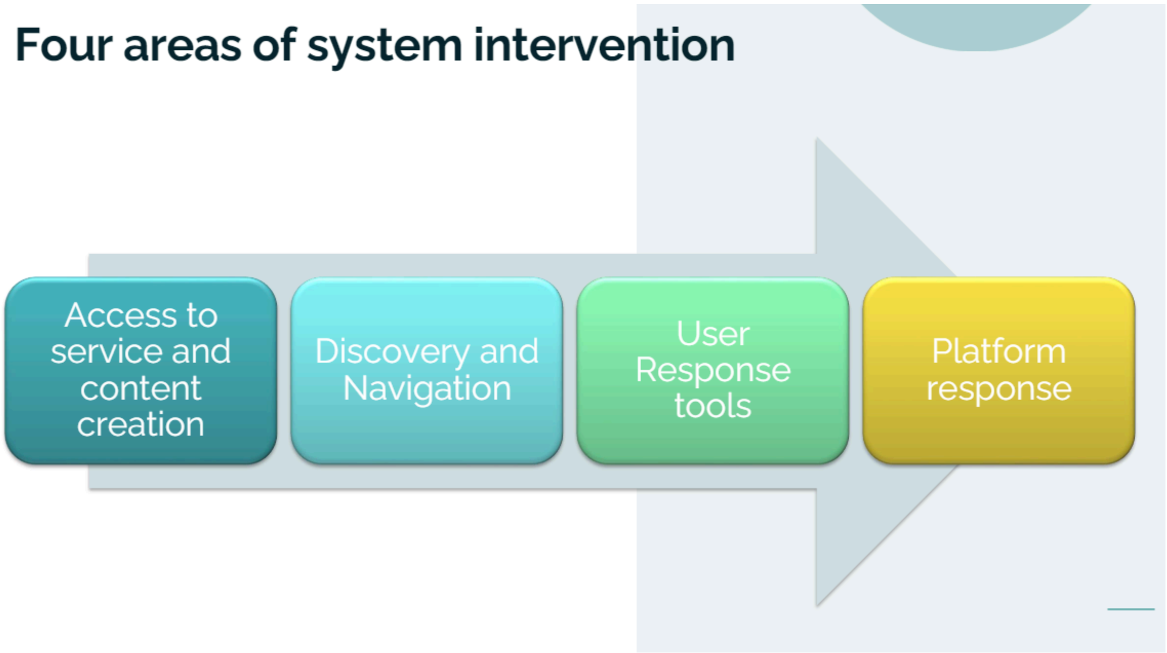
[11] Main report -p 12

[12] Updated Proposal para 43

[13] Main report p 13

[14] Lorna Woods, The Carnegie Statutory Duty of Care and Fundamental Freedoms, December 2019; https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/12/05125454/The-Carnegie-Statutory-Duty-of-Care-and-Fundamental-Freedoms.pdf See further, Ad Hoc Advice from Carnegie UK to United Nations Special Rapporteur on Minority Issues concerning guidelines on combatting hate speech targeting minorities in social media, November 2021; https://carnegieuktrust.org.uk/news-stories/ad-hoc-advice-to-the-united-nations-special-rapporteur-on-minority-issues/

[15] Lorna Woods, "The Duty of Care in the Online Harms White Paper" (2019) 11 *Journal of Media Law* 6, https://doi.org/10.1080/17577632.2019.1668605



Four areas of system intervention

Access to service and content creation

Discovery and Navigation

User Response tools

Platform response

(Image above taken from a presentation by Prof Lorna Woods and William Perrin to EPRA, available here)

13. While there has been much focus on the role of the recommender algorithm, and debate about whether there is a concern about filter bubbles, or rabbit holes for certain constituencies, the issue about design is not limited to this. It sits right the way across the four-stage model. So, design and business choices could include some of the following:
    - Metrics (and other hooks)
    - Frictionless communication (retweet/repost; like/upvote; share; ease of forwarding)
    - Content creation – eg livestreaming features
    - Content discovery systems (search engines/recommender systems/hash tags/trending today features/feeds from contacts)
    - Clickbait rewards
    - Targeted advertising
    - Limited investment in making complaints resolution easy/effective.

14. The four-stage model can also be used to show that there might be a distinction between general design choices (particularly at the point of service creation) and tools and techniques that apply once content is created (whether deployed by the user, or by the platform).

A Duty of Care and Safety by Design

15. A duty of care is well known as a tortious technique but has been used beyond that context – most notably it has been used in statutory contexts where the statute in effect adapts the basic principle to apply to a specific context or concern. This has shifted the duty from one embedded in private law claims to a regulatory tool.[16] In general, the obligation imposed is to exercise reasonable care and/or skill to avoid the risk of injury to relevant others, but in a regulatory context it reflects the idea that the risk creators should be the risk owners and allows for a forward looking and context appropriate approach based on general obligations that do not become outdated (or at least not too fast). The obligation is based on a case orientated to a particular objective or outcome, but because the focus is the level of care rather than the full achievement of the objective, the pressure to define the objectives in great detail is removed. The UK's Health and Safety at Work Act 1974 (HSWA), for example, uses the broad terms of harm. The context of the HSWA also illustrates a further point – the advantage of a general duty over lots of specific regimes. The HSWA was introduced following a disaster at a Welsh mining village, Aberfan. The safety legislation at the time was directed to the workers in the mines and not the people living in the village. Nearly a hundred and fifty people died yet, because no worker was hurt, the disaster did not fall within the existing legislation. The disaster, caused by 'ignorance, ineptitude and a failure in communications',[17] could have been prevented.

16. The duty of care could, in the view of Woods and Perrin, be satisfied where a provider adopted a risk assessment and risk mitigation process, supported by general risk governance structures. We have put forward an outline of some of the issues such an approach could cover.[18] Such an approach would incorporate an orientation to considering the consequences of design at the product development stages, and expressly consider the possible misuses of the technology as well as what happens when it scales. While this could be considered a "by design" approach, a duty of care could cover a wider terrain because it covers how the service is run and resourced, its business model as well as how it is designed. Moreover, the obligation relating to design might seem appropriate to the original development process, or when the service is updated but could also cover features that are not as integrated – so for example, standalone tools (including

[16] *Bourhill* v *Young* [1943] AC 92 (HL) per Lord Thankerton at 98 and per Lord Macmillan at 104

[17] Report of the tribunal appointed to inquire into the disaster at Aberfan on October 21st, 1966, (1966-67 (HC 553)), 19th July 1967, para 18, available: http://mineaccidents.com.au/uploads/aberfan-report-original.pdf [accessed 28 August 2019]. Note that there was no claim under negligence because the strict liability rule in Rylands v Fletcher 1868 LR 3 HL 330 applied.

[18] For example, see the application of this principle to online harm in Carnegie UK's work on a VAWG Code of Practice, developed with organisations from the VAWG sector, (https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2022/05/24163713/VAWG-Code-of-Practice-16.05.22-Final-1.pdf) and the advice submitted to the UN Special Rapporteur on Minority Issues on hate speech (https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2021/11/24164123/UN-Hate-Speech-draft-v.05a-1-1.pdf)

safety tech). The provision in these is also about the systems provided.

17. Significantly, and given the proposal is based on the duty of care, the measure of success is not wholly about output measures (though they may indicate whether an effective process is in place) but about the level of care found in outcome-orientated processes and choices. Assessment is about the features taken together and not just an individual item in isolation. Furthermore, assessment of appropriate mitigations will vary depending on when in the product development cycle the process sits; it may not be possible to design out or remove a core feature of a service. This does not mean that content-based solutions are the only options (eg requiring take down or down-ranking of specific items of content) – countervailing safeguards could potentially be designed in, or problematic features phased out over a period. In any event the provider should be clear about the identification of the risk and the mitigation response(s). Content moderation remains a necessary final point in any event because no design choice will be perfect or 100% effective.

18. Some of these changes may be content neutral (in the sense they are not identifying a topic of conversation) and preventative in that they may remove nudges or incentives towards certain potentially problematic types of content (eg misinformation – where clickbait headlines have generated significant revenues for junk news sites[19]). Some features can be tweaked *ex ante* (eg turning off autoplay as default) or *ex post* (deliberately choosing to demote certain speaker's content) and may consequently be assessed as being more systems-based in the former and content-focussed in the latter.

19. It is not possible entirely to separate systems from content and not all features will be content neutral. For example, the requirement to have efficient takedown processes sits on the boundary of system and content because it is necessarily linked to specific items of content. Perhaps the key point is that a systems approach would assess the process for takedown (what are the rules, how many people are there, how are they trained, what are the tolerances for false negatives and false positives, what are the appeals and oversight processes) rather than looking at the decisions on individual items of content. It would be possible for a system to get some decisions wrong and still satisfy the duty of care.

20. The "by design" approach is becoming ubiquitous in societal and regulatory responses to technological developments, particularly those involving digital technologies. Rather than waiting for problems to manifest and then deal with them, the developer of the service should seek to understand the risks and unintended side effects and seek to mitigate them. In the digital context, the "by design" terminology was first deployed in the data protection context by Ann Cavoukian when she was the Information and Privacy Commissioner of Ontario, Canada. She

[19] European Parliament, Automated tackling of disinformation, March 2019, p. 24: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf, accessed 20 March 2023

developed 7 foundational principles for privacy by design,[20] which could be adapted for the safety context more generally. Those foundational principles are:

1. Proactive not Reactive; Preventative not Remedial
2. Privacy as the Default
3. Privacy Embedded into Design
4. Full Functionality – Positive-Sum, not Zero-Sum
5. End-to-End Security – Lifecycle Protection
6. Visibility and Transparency
7. Respect for User Privacy

21. Note that the word “proactive” has developed some unfortunate connotations in the context of internet regulation, as it has been linked primarily to the use of upload filters. Proactive here is linked to design and is largely content neutral. We would emphasise the importance of the full functionality element; services should avoid unnecessary trade-offs and in particular avoid resorting to over-moderating to prevent having to think through more difficult problems.

22. A “by design” approach is seen in cyber security too. Here[21], we see five principles but which focus on the design of the system:
    - establish the context before designing a system
    - make the compromising of data or systems difficult for attackers
    - make disruption of the service difficult
    - make compromise detection easier
    - Design to naturally minimise the severity of any compromise

23. While account safety is not much discussed in the context of online harms, it is an important element of keeping users safe (for eg victims of domestic violence) or preventing misinformation. Understanding the context (and existing problems) before starting work is an important point that is often overlooked.

24. There are definitions of safety by design[22], notably including the approach of the Australian e-Safety Commissioner.[23] While these are focussed on the system broadly speaking, there is less focus on the design of the service and choice architecture; there is rather more emphasis on *ex pos*t responses to content (and the social contract embedded in terms of service). There seems

[20] Ann Cavoukian, Privacy by Design - The 7 Foundational Principles: Implementation and Mapping of Fair Information Practices, available: https://privacy.ucsc.edu/resources/privacy-by-design---foundational-principles.pdf

[21] NCSC, The Cyber Security Principles (v 1.0), 21 May 2019, https://www.ncsc.gov.uk/collection/cyber-security-design-principles

[22] See eg. UK Government’s Principles of Safer Online Platform Design: https://www.gov.uk/guidance/principles-of-safer-online-platform-design

[23] https://www.esafety.gov.au/industry/safety-by-design#safety-by-design-principles

less direct consideration of the earlier points of the information flow model, nor the business model. Nonetheless, the definition rightly notes that users' safety should be central to company processes and culture and that the responsibility to keep themselves safe should never fall entirely on the user.

Which services to target

25. In general terms, there has been a move toward technology neutral rules both in the interests of a level playing field for businesses and also in the interests of trying to ensure that regulation stays relevant for longer. While there may be difficulties in defining truly technology neutral rules in practice, in our view the best approach is to understand the sorts of functionalities that are relevant to the issues concerned and to try to describe those, rather than focussing on particular types of service as currently understood. If a broad duty approach is adopted, then the risk assessment which is context specific will likely vary depending on the distinct features of any particular service type.

Emerging issues from the Online Safety Act 2023 in the UK

26. The UK OSA[24] identifies types of content, and then specifies different duties – although the duties follow common themes of risk assessment and risk mitigation. It requires content to be defined before it is apparent which rules (if any) apply.[25] This adds complexity to implementing the regime, even before we think about the tension between an approach focussed on the design of the system (which happens before the content flows across it) and a regime which nonetheless depends on categorisation of content – especially if there is a focus on individual items of content rather than categories of content. Indeed, the focus around content tends to place the emphasis in discussions on take down and de-emphasises all the other points at which interventions could be made. It runs the risk of turning the regime into a more blunt instrument. It is also difficult to consider pure design harms – eg addictive design – in a model based on content.

27. As we noted at the start of this submission, it is still very early days in the UK regime and difficult to point to specific practical examples of where the hybrid approach of the UK legislation (part based on a systemic approach, involving risk assessment and upstream mitigation, and part based on a focus on types of content which require a moderation and/or takedown response) is problematic. The Act only came into force in November 2023 and won't be fully implemented - following multiple overlapping consultations by Ofcom - for at least another 18 months. (Our website contains links to all the relevant resources for this implementation phase, along with our

[24] https://www.legislation.gov.uk/ukpga/2023/50/enacted
[25] It is worth bearing in mind here that the general nature of the duty proposed by Woods and Perrin applied to providers taking care in relation to their own products, rather than following instructions about what specifically might be harmful from the legislature or regulator.)

analysis and responses to consultations to date.)

28. However, one particular example of note has emerged, which has wide ramifications for the effectiveness of the regime. In the Act, the illegal content safety duties are triggered by content linked to a criminal offence, not by a requirement that a criminal offence has taken place. However, Ofcom's proposals for the illegal content duties - on which it consulted at the end of 2023 - require that a criminal offence has taken place each time content is posted (rather than acknowledging that content which has been deemed illegal remains illegal when shared as it is still connected with the original offence). This leads to an unnecessarily limited view of relevant content being baked into the proposals. [26] A specific example of the implications of this is in the issue of reposting intimate images without consent - the repost is still the content linked to the original offence, it has not changed its nature.[27]

29. The approach set out by Ofcom in its proposals for the illegal content duties has formed the basis for its proposals for implementing the OSA's child protection duties. Like the illegal content duties, the Act sets out the content to which these duties apply - grouped into three categories (primary priority content, priority content and non-designated content). While the duties do not mandate the taking down of specific pieces of content (as they do in relation to illegal content), Ofcom's mirroring of the approach - identifying individual pieces of content and deciding what to do with them, rather than looking at the design of the service over which these designated pieces of content flow - is likely to lead to a similarly narrow mitigation response by regulated companies and a limitation in the protections for children envisaged by the Act.

Online Safety Act Network

June 2024

[redacted]

[26] More on the problems with the approach taken by Ofcom to the illegal content duties can be found in our public response, co-signed by 19 members of the Network: https://www.onlinesafetyact.net/analysis/osa-network-statement-on-illegal-harms-consultation/

[27] See recent campaign to bring in an amendment to the Criminal Justice Bill to address this gap: https://swgfl.org.uk/magazine/caroline-nokes-addresses-uk-parliament-to-prevent-intimate-image-abuse/