Submission to the Online Safety Charter Consultation Paper, 12 April 2019

Alice Witt,[1] Rosalie Gillett,[2] Nicolas Suzor[3]

QUT Faculty of Law and Digital Media Research Centre

Executive Summary

We are a group of Australian academic researchers working in digital rights issues and the social implications and regulation of technology. This submission outlines what we believe should be key priorities for any attempt to shape the direction of future reform of online safety policy and legislative arrangements in Australia, and specifically, responds to discussion questions 14 and 33. In summary, we:

- Stress the importance of clear, certain rules for content hosts and other internet intermediaries. Any content regulation or intermediary liability scheme must be proportionate and include strong safeguards for due process. We are extremely concerned about recent developments that seek to hold technology firms criminally liable for user-generated content, as this could have a number of unintended and potentially detrimental impacts;
- Highlight the importance of recognising that a major potential pitfall and risk of content removal is the takedown of critical expression and valuable counter-speech. This risk can be mitigated through multi stakeholder engagement and collaboration; training moderators (or regulators) to identify and understand content that counteracts hate speech; and ensuring due process mechanisms are in place (Question 14); and
- Emphasise the importance of transparency and accountability measures. In particular, we believe that in any system of co-regulation, it is critical that the technology firms that regulate online content provide access to granular, disaggregated data that can facilitate research on public interest questions (Question 33).

I. Regulatory approaches should be decentred, and necessary and proportionate

The last two decades of attempts to regulate certain content and behaviour in the online environment show us that 'top-down,' 'command and control' regulatory responses are often

[1] PhD Candidate, QUT Faculty of Law and Digital Media Research Centre <rosalie.gillett@qut.edu.au>
[2] PhD Candidate, QUT Faculty of Law and Digital Media Research Centre <ae.witt@qut.edu.au>
[3] Associate Professor, QUT School of Law and Digital Media Research Centre <n.suzor@qut.edu.au>.

ineffective and insufficient.[4] In order for regulation to be effective in 'decentred' contexts,[5] like the internet, it must be 'hybrid (combining governmental and non-governmental actors), multi-faceted (using a number of different strategies simultaneously or sequentially), and indirect'.[6] Addressing the complex issues around emerging technologies requires nuanced rather than blunt and heavy-handed regulatory measures.

Specifically, we suggest that it is not desirable to hold internet intermediaries liable, criminally or otherwise, for content posted by their users, without a clear takedown regime that provides adequate safeguards for due process and speech interests. We are extremely concerned about the recent introduction of the *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth), which creates criminal liability based on an unclear standard of imputed knowledge (recklessness) of abhorrent content. This is an example of rushed legislation that creates great risk to legitimate expression and the use of general purpose technologies by failing to adequately set out the obligations of technology companies.

Effective internet content regulation schemes must be carefully designed. Generally speaking, Australian intermediary liability law is a mess of confusing, unclear, and sometimes conflicting principles.[7] As a guide to good regulatory design, we particularly recommend the *Manila Principles on Intermediary Liability*, a set of standards developed by a group of key civil society organisations.[8] We note particularly that any internet content regulation scheme should have clear standards and detailed procedures for identifying prohibited content, challenging determinations, and reinstating material wrongfully removed. Crucially, it is important to avoid laws that create lopsided incentives, where internet intermediaries are only liable for failing to remove content, because they introduce undue legal risk and uncertainty and encourage intermediaries to remove content when they are in doubt.[9]

More generally, many of the challenges of regulating harmful content are often highly contextually and culturally specific.[10] These are not issues that can be safely left to the

[4] Nicolas Suzor et al, 'Submission to Human Rights and Technology Issues Paper' (2018) <https://eprints.qut.edu.au/122520/1/2018%20AHRC%20Human%20Rights%20and%20Tech%20Issues%20Paper.pdf> 13.
[5] See, eg, Nicolas Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4(3) *Social Media + Society* 1 <https://journals.sagepub.com/doi/10.1177/2056305118787812>.
[6] Julia Black, 'Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a "Post-Regulatory" World' (2001) 54(1) *Current Legal Problems* 103, 111.
[7] Kylie M Pappalardo and Nicolas P Suzor, 'The Liability of Australian Online Intermediaries' (2018) 40 *Sydney Law Review* 469 <http://www.austlii.edu.au/au/journals/SydLawRw/2018/19.html>
[8] Electronic Frontier Foundation et al, *Manila Principles on Intermediary Liability* (24 March 2015) <https://www.manilaprinciples.org/>.
[9] Nicolas P. Suzor, *Lawless: The Secret Rules That Govern Our Digital Lives* (Cambridge University Press, 2019) <https://eprints.qut.edu.au/123199/>.
[10] See, eg, Alice Witt, Nicolas Suzor and Anna Huggins, 'The Rule of Law on Instagram: An Evaluation of the Moderation of Images Depicting Women's Bodies' (2019) 42(2) *UNSW Law Journal*

discretion of internet companies to adjudicate -- a greater degree of real, substantial oversight is required. This can only be addressed through careful multistakeholder engagement and collaboration. In conclusion here, we commend the ongoing work of the The Office of the eSafety Commissioner, which has worked to develop strong relationships with internet companies and pursue novel co-regulatory approaches. We support the strengthening of the independence of the Office, and believe that the efforts to build ongoing multistakeholder dialogue to improve internet governance should be encouraged and supported.

II. The potential pitfalls and risks of content removal and how to mitigate them (Q 14)

There are a number of potential pitfalls and risks associated with content removal, including bias in processes for moderating content and possible violations of human rights. We focus on the risk of removing legitimate critical expression and valuable counter-speech, particularly from marginalised individuals and groups,[11] which directly and indirectly responds to hate speech and other pressing social issues. While technology firms are generally good at removing content that clearly transgresses their content policies, such as spam content or explicit sexual content, it is much more difficult to consistently identifying and moderating harmful or prohibited content more generally. A major problem, for instance, is that users can circulate hateful speech disguised through humour.[12] Given that there are differences in localised understandings of racism, content that may appear humorous in some cultural contexts may be highly offensive in others.

In designing regulatory regimes that address harmful content, it is particularly important to pay attention to the often disproportionate effects on already marginalised voices. Counter-speech is an important mechanism through which to combat and reduce the visibility of hate speech. Critical counter-speech also enable individuals to respond to and challenge expressions of hate and discrimination,[13] highlight existing social inequalities, and call out problematic attitudes toward marginalised and vulnerable populations who are the usual targets of hateful content.[14]

(in press) (arguing that applying standards of appropriateness to content is an immensely complex task).

[11] See, eg, Stefanie Duguay, Jean Burgess and Nicolas Suzor, 'Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine' (2018) *Convergence: The International Journal of Research into New Media Technologies* 1.

[12] Andre Oboler. 'Aboriginal Memes & Online Hate' (Melbourne: Online Hate Prevention Institute, 2012) <http://www.ohpi.org.au/reports/IR12-2-Aboriginal-Memes.pdf>.

[13] Jamie Bartlett and Alex Krasodomski-Jones, 'Counter-Speech Examining Content That Challenges Extremism Online' 2015, <http://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf>, 21.

[14] Anat Ben-David, and Ariadna Matamoros-Fernandez, 'Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain' (2016) 10 *International Journal of Communication* 1167.

Any attempt to moderate or remove hateful content, especially through blanket prohibitions and other regulatory measures, must pay attention to the risk of silencing valuable counter-speech and further marginalising vulnerable populations.[15] To mitigate this risk, there is a pressing need to make real commitments for extensive multi-stakeholder collaboration and the sharing of information and expertise to better understand hate speech and counter-speech. It is particularly important to engage with minority groups, those who are targets of hateful content, as well as anti-violence against women and anti-racism experts. We encourage technology firms, who currently undertake most of the work of regulating online content,[16] to better train their moderators to identify and address hateful content. As a first step, given the challenges of understanding localised hateful speech, we suggest that digital media platforms build on community consultation and collaboration to improve their knowledge and understandings of hateful content. Indeed, many of the themes discussed throughout the Consultation Paper highlight the need for community consultation and collaboration. Accurately understanding local context is critical to ensure that regulatory decisions about hate speech are consistently and accurately made. We recognise, however, that mistakes are inevitable, and as such, there must be transparency and accountability mechanisms in place for users to appeal decisions about content. We detail these mechanisms in the following Part.

III. What elements should be reported on and how can consistency of reporting be achieved? (Q 33)

As discussed in the Consultation Paper, technology firms including Google, YouTube, Facebook and Twitter, have published transparency reports and some information about the ways content is moderated behind closed doors. While these are steps in the right direction, a great deal more work needs to be done to make these systems more understandable, particularly for regulators, non-governmental organisations, academics, and journalists who seek to evaluate how well content moderation systems are working. In order to achieving consistent transparency reporting, technology firms must move beyond the mere provision of information. Aggregate statistics alone are insufficient to evaluate the effectiveness of moderation processes.[17]

As a statement of best practices, we highly recommend the *Santa Clara Principles*, a joint declaration developed by civil society organisations and academics that outlines minimum

[15] See, eg, Sarah Myers West, 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms' (2018) 20(11) *New Media & Society* 4366 <https://doi.org/10.1177/1461444818773059>.
[16] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, 1st edition, 2018).
[17] Nicolas P Suzor et al, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation' (2019) 13 *International Journal of Communication* 1526 <https://ijoc.org/index.php/ijoc/article/view/9736>.

levels of transparency and accountability in content moderation.[18] The Principles provide that, at a minimum, companies should publish the numbers of posts removed and accounts temporarily or permanently suspended as a result of violations of terms, guidelines and/or other policies. Second, firms should provide their users with notice about what content is taken down or account is suspended, the reasons for these decisions, and a description of the people and automated processes responsible for identifying content and making the decision. Third, the Principles emphasise the need for timely appeal decisions about content. Appeals should be undertaken by moderators who were not involved in initial decision-making processes.

We highlight the importance of binding reporting standards as a monitoring tool in any formal content regulation schemes. Any new legislative or co-regulatory regimes should include clear obligations to make available sufficiently granular data on a regular basis to enable public evaluation of the regime and the actions taken by relevant stakeholders. Ideally, these reporting obligations should be part of an accountability framework, where those who are responsible for enforcing content standards are subject to clear consequences for non-compliance with transparency standards.[19] Most importantly, we suggest that attempts to improve transparency around online content regulation should be reframed to focus on access to large-scale disaggregated data, in a way that can empower researchers and help improve public confidence in content moderation systems.

[18] ACLU Foundation of Northern California et al, *The Santa Clara Principles on Transparency and Accountability in Content Moderation* (7 May 2018) <https://santaclaraprinciples.org/>.
[19] See, eg, Julia Black, 'Constructing and Contesting Legitimacy and Accountability in Polycentric Regulatory Regimes' (2008) 2 *Regulation & Governance* 137, 150.