Submission to the Australian Government's Consultations on a new Online Safety Act draft exposure Bill

Submitted behalf of Australian Community Managers by:

Venessa Paech
ACM Co-Founder and Director
[redacted]

Dr Jennifer Beckett
The University of Melbourne
ACM founding member
[redacted]

Associate Professor Fiona Martin
The University of Sydney
ACM founding member
[redacted]

Date: 14/2/2021

Introduction

We welcome the opportunity to provide feedback on the Online Safety Act draft exposure Bill. Australian Community Managers is the national professional organisation for online community management practitioners. Our members plan, build and manage online communities across industries and contexts, on social media platforms and 'owned' or purpose-built digital environments. They facilitate, monitor and regulate the information posted in these environments, and interactions between community members.

Originally founded in 2009 as the Australian Community Manager Roundtables, ACM was formed to professionalise this field of critical digital regulatory work, and to build and enhance online community management practice by providing training, resources, mentoring and networking for members. We believe thriving online communities are important for the future of work and society, and that our members, who act as community designers and custodians, are essential to their growth, sustainability and shared value.

Professional community managers moderate content and ensure digital safety across a variety of different platforms, dealing with difficult issues such as hate speech, harassment and stalking, defamation and contempt of court. We are engagement, cultural development and governance specialists and are responsible for overseeing risk management around existing legislation, industry codes and best practice guidance. We ensure our organisations (ranging from corporate, to not-for-profit, to government) and our community members don't break the law, and that we insulate our social spaces from potential harms, via formal and informal governance mechanisms.

Australian Community Managers informs our members about significant legal developments, such as the recent Voller vs Nationwide ruling on the liability of media companies for third party comments. We also work with researchers who specialise in the areas of social media governance, online harm reduction (including digital safety and security), online dis- and misinformation, moderation and mediation techniques and the wellbeing of community management professionals. Associate Professor Martin and Dr Beckett have worked with ACM for the past five years, helping to enrich and inform practice in these areas. They are professional members of ACM and each have industry experience in the field. Assoc. Professor Martin is the Asia lead on a UNESCO funded International Center for Journalists project which is investigating online harassment of women journalists globally and means of combatting this problem. She is also, with Professors Terry Flew and Nic Suzor, and Associate Professor Tim Dwyer a co-investigator on the Australian Research Council Discovery Project Grant DP190100222 "Platform Governance: Rethinking Internet Regulation as Media Policy".

General position on The Act

Broadly, Australian Community Managers is supportive of the proposed Online Safety Act and its intention to create more consolidated, updated and wider oversight of citizen safety on digital platforms. We have long advocated for greater regulatory accountability from platforms, organisations that draw value from social media, such as brands, and the bad actors who perpetrate harm in digitally networked environments.

Online acts of intimidation, threats or menace, abuse and harassment are often far from casual, and have demonstrated relationships to serious threatening behaviour offline (Stevens, Nurse and Arief, 2020). We are happy to see their inclusion in the proposed Act, underscoring our own efforts to mitigate and manage these harms when they invariably emerge in our communities.

Our response to The Act is framed around five (5) core areas of concern:

1. The definitional scope of "online social interaction" and ambiguity regarding "business purposes" (Section 13 p. 19);
2. Details about complaints to the Adult Cyber-Abuse Scheme
3. The absence of reference to non-consensual, sexually violent and intimidatory sharing of photo-manipulated personal images such as deep fake porn images and mutilation images in the Image-based Abuse Scheme,
4. The exclusion of online hate speech and dis-information from the ambit of the Bill; and,
5. The exclusion of start-ups from safety and transparency standards

Additionally, we have provided recommendations that address these concerns, and which we believe will help create a more robust and practical legislative instrument.

Core concerns

1. *Concern: Definition of 'online social interaction' does not adequately cover the scope of business-related activity during which harms can occur*

We are concerned about limitations in the definition of online social interaction, starting with Section 13(1)(a)(i) of the draft exposure Bill, where a social media service is defined as a service where the "the sole or primary purpose of the service is to enable online social interaction between 2 or more end-users" (p.19 l. 11).

While we appreciate the need to draw boundaries around the scope of the Bill to allow its practical implementation, it is difficult in these environments to distinguish where

business ends, and sociality begins. Social media services are multi-sided markets where business and sociality combine: “federating and coordinating Internet actors in innovation and competition, and creating value by harnessing economies of scope in supply or/and in demand” (Flew, Martin and Suzor, 2019, p.36).

Many online interactions facilitated via social media and in online communities occur *as a result of business purposes*, including the trade or sale of items and the training of individuals or groups. For example, digital marketplaces such as Envato feature various and extended forms of online social interaction between members in a public forum, in the course of individuals advertising their services or products. Similarly, news and media companies across Australia use Facebook Groups to deepen their connection with audiences, posting stories to attract comment and traffic to the company website for a *primarily business purpose*. Further there are many examples of social groups which are hosted by individuals for a primarily business purpose, such as advertising to that social group, or recruitment to a subscription for training or specialist information. Communities often begin with functional and informational exchange and deepen into a stronger social context over time.

The digital communications platforms are primarily businesses in themselves, aggregating user data in order to attract and hosting targeted behavioural advertising which has been, on occasion both intimidatory and hateful, as was acknowledged in the recent global “Stop Hate for Profit” (2020) campaign against Facebook. Their primary purpose is not to support social interaction, but to gather, pattern, analyse and sell information gathered from that interaction.

We are concerned then that the draft Bill’s distinction between ‘social’ and ‘business’ purposes is not only unclear, but may exclude online settings where harms covered by the Bill may, and do, occur. This ambiguity is compounded by the following notes, which suggest that online social interaction “does not include (for example) business purposes.”

> 13.(1)(b) Online social interaction does not include (for example) online business interaction.
> 13 (2) Social purposes does not include (for example) business purposes.

Given the degree to which business and sociality blend in platform environments, and the way in which micro-targeted advertising on digital platforms can be mis-used, the explanatory clauses 3 (a) and (b) fail to clarify circumstances in which cyber-abuse would not occur.

2. *Concern: Details of complaints to the Adult Cyber-Abuse Scheme*

We welcome the creation of a dedicated Adult Cyber-Abuse Scheme to address the widespread problem of digital abuses and harms experienced by Australian adults as elucidated in the *Online Safety Reform Discussion Paper*.

We note that complaints procedure requires the complainant to first report the issue to the service provider and, if the relevant service provider "fails to address" the concern, the complainant may choose to escalate the complaint to the Commissioner. We are concerned that the phrase "fails to address" may then exclude complaints from those who receive a response, even if that response is inadequate or only addresses the complaint in part.

We would like to see the wording of the process changed, to that a complaint can be lodged with the Commissioner:

- if the response from the relevant service provider is inadequate or fails to address the complaint in full.

We would also like the complaints investigation procedure to acknowledge circumstances where the burden of reporting becomes a component of the abuse itself. The complaints reporting process, as with service provider reporting processes generally, can be and is sometimes weaponised by bad actors to increase the burden and trauma on a target individual or group (Gillespie, 2018). That is, actors can time attacks in bursts, so that complainants may be in the position of just having filed a complaint when the next attack occurs. In this respect the ongoing time and psychological burden of constant reporting can increase the trauma and economic cost of the harassment. We suggest the Commissioner allow complainants in the circumstance of ongoing harassment by an individual or group to add evidence to an existing complaint, rather than having to file new complaints for each batch of abusive messages.

3. *Concern: Inconsistency in the Image Based Abuse Scheme*

While we welcome the enhancements to the Image Based Abuse Scheme we note it only deals with intimate images, even though there is widespread misuse of personal images to intimidate and harass.

The UNESCO/ICFJ report which Assoc. Professor Martin is currently writing, documents numerous instances of women journalists receiving sexually violent non-consensually manipulated images of themselves, with their heads digitally mapped onto graphic video pornography, in so-called 'deep fake porn' videos, 2D pornographic images, or images of mutilated bodies/corpses. Such images are then often widely distributed with the intent

of shaming, menacing and/or silencing women in public positions. Martin has recorded women being subject to this type of abuse from the early 2000s.

This is graphic psycho-sexual image-based abuse but is not covered by the current Image Based Abuse legislation which only refers to the distribution of intimate sexual images. This gap in the law has been recently raised in the UK media (Royle, 2021). One of the reasons the ACM industry Code of Ethics for Community Managers (ACM, 2016) was initially created was that an Australian publisher, Zoo Weekly magazine, was using non-consensually digitally manipulated personal images of women to engage their followers on social media.

While this form of abuse will conceivably be covered by Adult Cyber Abuse provisions, it appears inconsistent to deal with it there, when both the intent of the perpetrators - to use sexual violence to menace, harass and silence the subjects - and the means of abuse - non-consensual use of personal images - is the same. The Act needs to acknowledge the online distribution of sexually violent non-consensually digital manipulated personal images as a crime of the order of so-called 'revenge porn'.

4. *Concern: Lack of consideration given to the intersection of online hate speech, dangerous organisations, disinformation and the harms they cause Australians*

We note that the draft Bill brings the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act (2019) under its purview. This is an excellent step in acknowledging the potential harms to Australians in viewing such material online.

Given this, we question as to why the Bill does not also address the linked issues of online hate speech, online recruitment to extremist groups and the networked spread of dis- and misinformation. All of these factors have been shown to increase the risks of harm to vulnerable individuals and groups (Chan, Ghose and Seamans, 2016; Williams, Burnap, Javed, Liu and Ozlap, 2020, Wright, Trott and Jones, 2020) and also endanger the security of the nation and trust in national institutions (Davis, M., 2019a and 2019b). Indeed, as the events at the US Capitol on January 6, 2021 have shown, the use of online spaces to foment hate, insurrection and peddle disinformation can lead directly to attacks on democratic institutions themselves.

In addition, it is now widely known that the Christchurch shooter's beliefs and actions were seeded, tested and reinforced on social media platforms prior to his streaming them online. Simply put, failing to tackle these issues consistently in online space is likely to lead to the development and spread of the types of Violent Abhorrent Material on the internet the Government was seeking to prevent via the 2019 Act.

Notes on the function of the eSafety Commissioner, at the intersection of the Online Safety Charter and the Online Safety Bill.

ACM strongly supports the continued development of the eSafety Commissioner's 'Safety by Design' initiative and the use of the *Online Safety Charter* to inform the development of the draft exposure Bill and to provide guidance to service providers and users after the introduction of the new Online Safety Act. However, the Charter should be amended as part of reforms to the existing legislation.

Guidance on strong, best practice governance, which includes acknowledgement of various regulatory and legal mechanisms available to service providers and users, is an integral part of ensuring community safety and setting effective platform norms and culture (Flew, 2015; Gillespie 2018; Roberts, 2019). Given that the Charter provides guidance to Australians, we believe it prudent to point to some potential inconsistencies between its current framing and that of the language in the Bill itself, specifically as it relates to the Charter Sections 2.1 - Scope and Application - and 2.2 - Empowering Users.

5. Concern: Carving start-ups out of a "common transparency standard" and to the need to "significantly invest in online safety tools" (Online Safety Charter, p. 5)

In the Scope and Application of the Charter (p.5), it suggests that start-up companies be exempted from consulting on or meeting transparency standards and incorporating safety by design mechanisms in the operation of their technology.

> "It would not be feasible, for example, for start-up firms to develop a common transparency standard for application across the industry, or to significantly invest in online safety tools such as content hashing."

We support the establishment of consistent and cohesive online safety expectations across all platforms, tools and digitally facilitated interactions and argue that this should be embedded in industry practice from the start of any relevant enterprise.

We believe it should not be optional for companies creating techno-social tools and environments to design for governance and safety. All platforms, tools and applications used to build and facilitate digital sociality should be required to prioritise the security impact of their design on the community, over and above the profit motive. They should also be required to embed industry transparency standards into their governance strategies. As long as we continue to treat content moderation and platform governance as optional extras in technological development, we will continue to see systemic incidences of the harms outlined in The Bill.

Good governance is the mechanism by which we “promote positive and respectful user behaviour” (Online Safety Charter, p. 4). Civil and constructive behaviours in digital social settings normalise rapidly and are difficult to unravel once set. The problem of retrofitting good governance functions into platforms and tools, is one that the Australian government is currently trying to tackle in the News Media Bargaining Code. There it is requesting that Facebook and other platforms provide news companies the tools to better moderate comments, to more effectively control legal offenses such as defamation. The fact that companies initially opted to ignore these functions is a decisive reason that we are faced with systemic issues of online abuse. This is what Gillespie (2018) has referred to as the “long hangover of Web 2.0” (p.202).

The Charter should also emphasise the need for human governance alongside technological solutions to online safety. “Content hashing” (Online Safety Charter, p. 5) and artificial intelligence based illegal content detection are only two algorithmic mechanisms of moderation and governance. As numerous researchers have pointed out, human moderation will continue to be an essential part of online governance structures (Gillespie, 2018; Flew, Martin and Suzor, 2019; Roberts 2019; Suzor, 2019). With that in mind, there is a proven link between the presence of community managers (for which moderation is one component of their work) and the reduction of harms such as abuse, harassment, threats of harm, hate speech, defamation and others.

Community managers develop community speech standards, cultivate civil and constructive social norms, moderate to remove harmful content and constrain bad actors, and educate community members to promote safe communicative environments. The presence of qualified community management is an inherent mitigant against many of the harms the draft Bill covers.

Community management additionally oversees risk management around threats of self-harm, ensuring these are managed proactively and escalated to the appropriate specialist support services. Lastly, community managers have expertise in building belonging, trust and other qualities of healthy and safe digital spaces. We believe community management should be named in the Scope and Application and the User Empowerment sections of the Charter as a means to promote awareness of the profession and its critical contribution to the mitigation of online harms.

Finally, while it is outside the scope of this consultation, ACM argues that best practice governance and moderation standards should be requirements for online platforms and start-ups seeking government investment funds, subsidies or grants. If a digital social product is receiving funding from the government, it is reasonable to require its adherence to safety by design principles, a minimum standard of community governance and tools to support moderation. Community management should be recommended as part of any basic online safety principles and governance strategy.

Recommendations

After consideration of the draft exposure Bill and the documentation accompanying it, we make the following recommendations based on the concerns outlined above:

1. Broaden the definition of 'online social interaction'

Overall the definition of social media services fails to capture the scope of business related environments in which cyber-abuse offenses and related digital harms may occur, and the overlap between business related activities and social media communication. Broadening the definition to include all sites of digital social interaction, regardless of their 'primary purpose' will ensure that Australians have recourse to safety mechanisms wherever they interact online, and reduce ambiguity in complainants pursuit of redress. References that exclude 'business purposes' from the scope of the Bill need reconsideration to better delineate their precise meaning.

2. Amend the working of complaint procedure for Adult Cyber-Abuse Scheme

We would like to see the wording of the process changed, to that a complaint can be lodged with the Commissioner:

- if the response from the relevant service provider is inadequate or fails to address the complaint in full.

Further, we note that in instances of domestic, racial, religious, sexual or sexuality-focused abuse part of the mechanism of abuse is forcing the ongoing labour of reporting. We recommend that in instances of ongoing abuse, the Commissioner allow the addition of evidence to existing complaints where the new material does not differ significantly from the original complaint in subject matter or intent.

3. Amend the Image Based Abuse provisions to include sexually violent non-consensual manipulation of personal images.

The new Bill should acknowledge the similarities between the existing offenses covered in this scheme and the non-consensual, sexually violent and intimidatory sharing of photo-manipulated personal images such as deep fake porn imagery and mutilation images. This will be critical as deep fakes develop in sophistication and the software to produce them becomes more readily available.

4. *Bring legislation regarding online hate speech and disinformation under the umbrella of this Bill to allow the eSafety Commissioner more power in requesting removals of harmful material*

   This recommendation seeks to bring consistency to the Bill regarding the removal of abhorrent violent content. We argue that in the face of the rise of political extremism online this Bill requires extension to include provisions for the timely removal of hate speech and disinformation, in order to reduce the potential for further harms of a more serious degree resulting from those offenses.

   Our final recommendations are broader, and refer to the eSafety Commissioner functions, its Online Safety Charter and its Safety by Design initiative.

5a. *Online Safety Charter should be revised to ensure that all digital companies, including starts-ups, which produce platforms or applications must incorporate safety by design principles and moderation and governance functionality.*

   Moderation tools help scale the management and mitigation of harms, reducing the sole burden on platform companies, and allowing communities and organisations to manage risk in contextually and culturally appropriate ways. By setting consistent expectations of safety by design, and requirements for governance and moderation tools we can reduce the potential that future online harms will become unmanageable. It is our firm belief that if a start-up or an existing platform is not willing to invest in effective governance and moderation from the outset, consistent with the principles of 'safety by design' then they should not be creating their product.

5b. *The Online Safety Charter and Safety by Design principles be revised to recommend that platforms and services which include online social interaction should be required to retain the services of a community manager*

   Community management is a critical intermediary force in the governance of online communities. The Online Safety Charter and the 'Safety by Design Principles' need revision to highlight the role of professional community managers. Further as part of its educational function the eSafety Commissioner's office should communicate to platforms, business, not-for-profit, government and NGO sectors the essential role community managers play in digital harm mitigation.

References

Australian Community Managers (2016) Community Manager Code of Ethics https://www.australiancommunitymanagers.com.au/codeofethics

Chan, J., Ghose, A. and Seamans, R. (2016) The internet and racial hate crime: Offline spillovers from online access; *MIS Quarterly* 40(2): 381-403

Davis, M. (2019a) Networked hatred: New technology and the rise of the right; *Griffith Review* 64: 83-94

Davis, M. (2019b) Transnationalising the anti-public sphere: Australian anti-publics and extremist online media in: *The Far Right in Contemporary Australia*, Palgrave Macmillan, pp. 127-149

Flew T. (2015) Social Media Governance; *Social Media + Society* 1(1):1-2

Flew, T, Martin, F, and Suzor, N. (2019) Internet Regulation as Media Policy: Rethinking the Question of Digital Communication Platform Governance. *Journal of Digital Media & Policy* 10(1): 33-50.

Gillespie, T. (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media,* Yale University Press

Roberts, S. (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*, Yale University Press

Royle, S. (2021) Deepfake porn images still give me nightmares. BBC News. January 6. https://www.bbc.com/news/technology-55546372

Stevens, F, Nurse, J R.C. and Arief, B (2020) Cyber Stalking, Cyber Harassmentand Adult Mental Health: A Systematic Review. *Cyberpsychology, Behavior, and Social Networking* (Online first http://doi.org/10.1089/cyber.2020.0253)

Suzor, N. (2019) *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge: Cambridge University Press.

Williams, M.L, Burnap,P., Javed, A., Liu, H. and Ozlap, O. (2020) Hate in the machine: Anti-black and anti-Muslim social media posts as predicators of offline racially and religiously motivated crime; *British Journal of Criminology* 60: 93-117

Wright, S.; Trott, V; and Jones, C. (2020) 'The pussy ain't worth it bro': Assessing the discourse and structure of MGTOW; *Information, Communication and Society 23(6)): 908-925*