Submission of Evidence to Online Safety Charter Consultation Paper

by Dr Kim Barker and Dr Olga Jurasz

March 2019

Dr Kim Barker
Stirling Law School
University of Stirling
Scotland

Dr Olga Jurasz
Open University Law School
Open University
England

We are responding to the call for evidence in our capacity as experts on social media abuse, online violence against women, online misogyny, and internet regulation. We have in the past made significant contributions to the UN calls for evidence on online violence against women, to the Bracadale Review on Hate Crime in Scotland, the Women and Equalities Committee inquiry into sexual harassment of women and girls in public spaces, and to Scottish Government's 'One Scotland: Hate Has no Home Here' Consultation on amending Scottish hate crime legislation.[1] In addition, we have made representations to the Scottish Government as to the need to amend legislation to cover a wider range of harassing and abusive behaviours online. We have recently published a world-leading volume '*Online Misogyny as a Hate Crime: a Challenge for Legal Regulation*' (Routledge 2019). We have been working on issues relating to harassment of women and girls in online spaces since 2013. We are possibly your only evidence respondents that have experience of the wider issues surrounding online harassment, and who take a holistic approach to the legal problems posed by such harassment, merging criminal law, gender, human rights, and internet law expertise. We would add that we are happy to provide further expertise or evidence if this is of use. We are only commenting on the questions posed from the perspective of our research, which generally focuses on internet regulation (broadly conceived) and the impact abusive content (especially its gendered manifestations) has on the participation and safety of women online. This expertise is placed within considerations of legal responses to addressing the challenges posed by illegal, harmful and abusive online content, with a specific focus on the gender dimension of these pressing issues.

[1] Kim Barker and Olga Jurasz, 'Submission of Evidence on Online Violence Against Women to the UN Special Rapporteur on Violence Against Women, its Causes and Consequences, Dr Dubravka Šimonović' (Open University, November 2017) http://oro.open.ac.uk/52611/; Kim Barker and Olga Jurasz, 'Submission of Evidence to Scottish Government Independent Review of Hate Crime Legislation (Bracadale Review)' (Open University, December 2017) http://oro.open.ac.uk/52612/; Kim Barker and Olga Jurasz, 'Written Submission of Evidence to the Women and Equalities Committee Inquiry into Sexual Harassment of Women and Girls in Public Spaces' (Open University, March 2018) http://oro.open.ac.uk/53804/.

1. *What are the examples of technology-facilitated solutions to enhance online safety, and how effective have these solutions been in addressing harms and mitigating risks?*

Technology facilitated solutions for the enhancement of online safety could include reporting websites offering a quick 'hide page' mechanism for users accessing assistance while at risk. Other examples include schemes such as 'Share Aware'[2] or the UK Safer Internet Centre's work across a range of demographics with initiatives including Digizen Game,[3] and Social Media Checklists[4] - including platform specific safety guides. Whilst these initiatives are not exclusively technology-facilitated, they blend technology with other mechanisms to offer guidance and 'real-life' examples of how to be safer online.

Other mechanisms that platforms have introduced in attempts to improve online safety include Facebook's Real Names Policy[5] although this has proved controversial,[6] and has even been deemed unlawful by some national courts.[7] Twitter has also introduced changes to its platform, and has introduced specific guidance within its Help Centre to address concerns surrounding abusive behaviour and stalking.[8] These initiatives have not been effective in addressing harms nor in mitigating risks – they are means of information rather than active measures but nevertheless give a 'nod' to awareness by the platforms themselves that there are issues that cannot be solved one-dimensionally, and not without oversight.[9]

2. *What tools are available and have been deployed to address safety issues for live-streamed content as it occurs?*

Platforms can – and should – deploy tools to identify unlawful content as it is being streamed. Microsoft, Facebook, Google and Twitter engineers have been working on tools to identify –

---

[2] NSPCC, 'Helping Children to Stay Safe Online' https://www.nspcc.org.uk/preventing-abuse/keeping-children-safe/online-safety/.

[3] Childnet International, 'Digizen Game' https://www.childnet.com/resources/digizen-game

[4] UK Safer Internet Centre, 'Social Media Guides' https://www.saferinternet.org.uk/advice-centre/social-media-guides.

[5] Facebook, 'What Names are Allowed on Facebook?' (Facebook Help Center) https://m.facebook.com/help/112146705538576.

[6] S Levin, 'As Facebook blocks the names of trans users and drag queens this burlesque performer is fighting back' The Guardian (28 June 2017) https://www.theguardian.com/world/2017/jun/29/facebook-real-name-trans-drag-queen-dottie-lux?CMP=Share_iOSApp_Other.

[7] J Kastrenakes, 'German court say's Facebook's real name policy is illegal' The Verge (12 February 2018) https://www.theverge.com/2018/2/12/17005746/facebook-real-name-policy-illegal-german-court-rules.

[8] Twitter, 'About Online Abuse' https://help.twitter.com/en/safety-and-security/cyber-bullying-and-online-abuse.

[9] A point Mark Zuckerberg has finally conceded in respect of platforms requiring input from regulators and lawmakers as well: M Zuckerberg, 'The Internet needs new rules. Let's start in these four areas.' The Washington Post (30 March 2019) https://abcnews.go.com/Business/facebooks-mark-zuckerberg-calls-government-regulations-op-ed/story?id=62068073.

for example – online child grooming.[10] Such tools indicate that technological solutions can be developed to identify online content and prevent it being streamed. Similarly, Governments have the means to empower law enforcement agencies to access any computer, and intercept information – something the Indian government has utilised.[11] This too is a particularly powerful tool in tackling streamed content which is illegal and / or harmful despite its controversy from free expression and privacy perspectives.

3. *What is the best way to establish a single 24/7 contact point for Australian authorities to ensure there is a timely response?*

What is meant by 'timely' here, and is it different to the expectation for 'expeditiously' referred to in the draft Charter (at pages 7&11)?

An online point of contact with an automated acknowledgement is perhaps the simplest mechanism but there is a risk that this could be overwhelmed with the volume of potential reporting. It may perhaps be more appropriate to require individual contact points for each platform / service, which would trigger a notification to a central repository so that there can be i) records, and ii) follow-ups after a response has been produced.

~~4. *Are there positive examples of flagging and content moderation? What makes these moderation systems work effectively and are they applicable to other services and applications?*~~

5. *Is there an acceptable error rate for inappropriately flagged or misidentified content?*

In a perfect world there would be no need for flagged content. Equally, there would be no error rate for inappropriately flagged or misidentified content. Errors ought to be minimal – certainly no more than a small percentage. It is difficult to conceive of this being achieved given the error rates encountered in other content flagging systems – one report indicates that there is a 38% error rate in copyright blocking injunctions.[12] Lessons must be learnt from the use of

[10] Home Office, 'New tool developed to tackle online child grooming' (13 November 2018) https://www.gov.uk/government/news/new-tool-developed-to-tackle-online-child-grooming.
[11] Press Trust of India, 'Government may amend IT rules to identify and curb origin of unlawful content' (25 December 2018) https://www.firstpost.com/tech/news-analysis/government-may-amend-it-rules-to-identify-and-curb-origin-of-unlawful-content-5786161.html; Tech2News Staff, 'Govt in talks to make amendments to s79 of IT Act to include breaking end to end encryption' (24 December 2018). https://www.firstpost.com/tech/news-analysis/govt-in-talks-to-make-amendments-to-sec-79-of-it-act-to-include-breaking-end-to-end-encryption-5781971.html.
[12] J Killock, 'UK Internet Regulation Part 1: Internet Censorship in the UK Today' Open Rights Group, (2019) https://www.openrightsgroup.org/blog/2019/formal-internet-censorship:-copyright-blocking-injunctions, 5.

copyright blocking injunctions which are error-strewn[13] because the risk to freedom of expression / speech rights should not be underestimated in any error rate – it is, in the words of the Open Rights Group, 'formal internet censorship.'[14]

Beyond this, lessons from New Zealand (and other states) provide recent examples of the difficulties posed by automated content scanning / removal. There is no accounting for nuance within automated systems.[15] Similarly, dealing with mass scanning of content is not something that can be error free. Where automated systems are relied upon, the error rate is likely – at least in the beginning – to be error-strewn.

6. What is an appropriate time frame for moderation and removal of content?

This should be dependent on the category of content. It should also be dependent on the category of the potential victim. For example, the most serious content should be removed within a very short time frame. The European Union is proposing that extremist or terror-related content be removed within one hour.[16] Other content which is illegal but perhaps of a lesser seriousness than terror-related content should be removed within a 24-hour period.

Content which is not illegal but harmful should be categorised differently and subjected to a different review process. Segregating content between that which is illegal, and that which is harmful but not illegal may be one method of mitigating the volume of content to be dealt with in short time frames. For example, content which encourages or demonstrates self-harm should be removed within 24 hours, whereas content which is abusive or otherwise harmful / unpleasant but not illegal should be removed within 3-5 days.

The European model in respect of time-frames for takedown and stay-down of illegal content online offers some benchmarks.[17] That said, content which is distasteful but not necessarily harmful should not be removed. A reasonable approach to content moderation has to be adopted otherwise sanitisation of online content becomes a risk, and that will lead to more and more content being taken to harder to access online locations. Sanitisation of online content also defeats the purpose of the internet.

---

[13] E Oswald, 'Out of control copyright bots are making a mockery of the DCMA' ExtremeTech (6 September 2012) https://www.extremetech.com/internet/135529-out-of-control-copyright-bots-are-making-a-mockery-of-the-dmca.

[14] J Killock, 'UK Internet Regulation Part 1: Internet Censorship in the UK Today' Open Rights Group, (2019) https://www.openrightsgroup.org/blog/2019/formal-internet-censorship:-copyright-blocking-injunctions, 5.

[15] L Woolery, 'Three lessons in Content Moderation from New Zealand and other High-Profile Tragedies' Center for Democracy & Technology (27 March 2019) https://cdt.org/blog/three-lessons-in-content-moderation-from-new-zealand-and-other-high-profile-tragedies/.

[16] Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online. 12.9.2018. COM(2018) 640 final 2018/0331(COD). Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online (C(2018) 1177 final).

[17] European Commission, 'State of the Union 2018: Commission takes action to get terrorist content off the web' (12 September 2018) http://europa.eu/rapid/press-release_MEMO-18-5711_en.htm.

7. How should content moderators be trained? What minimum standards should apply?

Training should be appropriate to the platform, and to the types of content that the moderators are likely to be exposed to. Training should prioritise and emphasise the support mechanisms available to moderators, and self-care should be a priority. In addition, given the volumes of abuse that is misogynistic, moderators should receive specific training relating to gender-based content. Gender perspectives are an essential element in tackling Online Violence Against Women (OVAW) and Online Violence Against Women in Politics (OVAWP) in particular – points highlighted by Julia Gilliard in her 'Sexism and Misogyny' Speech in 2012[18] but also illustrated through the incredibly high volume of abuse received by the UK's first black MP, Dianne Abbott in the 2017 General Election, who was targeted with over 8000 abusive tweets and messages sent directly to her Twitter account in the first six months of 2017.[19]

8. What sort of guidance should be available to moderators about dealing with vulnerable groups, such as children and Indigenous Australians?

Women should not be excluded from considerations of vulnerable groups, especially with the growth in technology-facilitated violence, online violence against women, and the use of technology in coercive control contexts.

~~*9. Are there positive examples of identification and content removal practices? What makes these practices effective and appropriate?*~~

~~*10. How should records of removed content be kept to ensure that evidence is available if needed by authorities?*~~

~~*11. Are there minimum requirements to uniquely identify content (for example, IP addresses of upload/posting source, geographic identifiers etc)? If so, please provide details.*~~

12. Can content be made invisible on a permanent basis? If so, how?

Making content invisible is not a solution to any problem concerning online safety. Content can be hidden by users (or by moderators), or removed by moderators. Within Europe, there are mechanisms by which Internet platforms are required to delist content, and make it much more difficult to access – colloquially referred to as the right to be forgotten.[20] That said, such a right is not about the making invisible of content – if content is to be made invisible on a permanent basis, this is really something falling squarely within the remit of takedown and staydown. Staydown procedures need to be carefully balanced with free speech rights and responsibilities.

[18] The Sydney Morning Herald, 'Transcript of Julia Gilliard's speech' (10 October 2012) https://www.smh.com.au/politics/federal/transcript-of-julia-gillards-speech-20121010-27c36.html.
[19] Amnesty International, #ToxicTwitter: Violence and Abuse Against Women Online (2018), 17.
[20] C-131/12 *Google Spain SL, Google Inc.* v *Agencia Española de Protección de Datos, Mario Costeja González* (2014).

Making content permanently invisible could be achieved after human verification procedures for reports of illegal or offensive content. The problems with requiring content to be made invisible and to not reappear on other platforms are currently being considered by Facebook after stinging criticism has been levied at platforms for not doing more to tackle the spread of illegal and seriously harmful content.[21] Where content is to be made invisible on a permanent basis, this should be a mechanism reserved for the most harmful of content, and should arise in situations where the harm of not doing so is severe. We would expect that such content would give rise to criminal proceedings being taken for the actions the content relates to.

13. Are there barriers to sharing of information about offensive content removed by an industry participant to prevent it being uploaded to another platform or distributed using another service?

Where barriers – relating to privacy rights or data protection for example – arise, it could be possible to overlook these if there is a national security or public protection interests. This could most notably occur in the context of terror-related content or coverage of terror content which if reposted or shared using other platforms would be particularly problematic. Beyond this, technical barriers can be circumvented – most feasibly manually depending on the volume of content in question.

14. What are the potential pitfalls and risks with content removal? How can these risks be mitigated?

Pitfalls and risks associated with content removal include missing harmful and illegal content and removing lawful content instead of unlawful content. Beyond this, risks also include removing content which does not violate standards of behaviour, or removing content and not replacing it should it be deemed not to be in violation of the law or behaviour standards.

Other risks include the expeditious or timely removal of content, and the pitfalls here include how content is classified and apportioned to timescales – for instance, is all illegal content to be removed before some harmful content or content which violates minimum standards of behaviour? Is there a graduated scale for different levels of seriousness, and how are these to be processed if so? Other pitfalls are likely to be similar as those experienced in respect of illegal downloads and measures to address infringing content. Where bots are relied upon, there are few opportunities to review due to the volume of requests for takedown.[22] Similarly, even where there is human oversight, mistakes will still be made. Beyond that, the volume that some platforms will have to deal with means that essentially the only feasible method is to rely

[21] J C Wong, 'Facebook finally responds to New Zealand on Christchurch attack' The Guardian (29 March 2019) https://www.theguardian.com/us-news/2019/mar/29/facebook-new-zealand-christchurch-attack-response?CMP=Share_iOSApp_Other.

[22] J M Urban, J Karaganis, and B Schofield,'Notice and Takedown in Everyday Practice', UC Berkeley Public Law Research Paper No. 2755628. http://dx.doi.org/10.2139/ssrn.2755628.

on bots, increasing the likelihood of errors.[23] One of the more successful systems is that of the YouTube Content ID System[24] – this may offer a basis for a model but is not without flaws and still generates errors.

If there is a default position of take-down, this could amount to a stay-down situation if the content (when reviewed by a human moderator) is deemed illegal. If there is a default position of take-down, this could be maintained as the first-stage until review and could be revoked if the content is deemed acceptable on review. This system would also allow for human moderators' workloads to be balanced as technological solutions should capture the majority of instances where the content is clearly illegal.

15. What should minimum standards of behaviour be? ~~Should they be higher for products and services directed at children, or that have a substantial number of child users?~~

An oft-cited principle for minimum standards of behaviour is that which does not cause harm to others, nor violate the terms of use policies. That said, it is difficult to proactively police this in an online platform. Notionally, what is illegal offline should be illegal online and standards of behaviour should be the same irrespective of the digital context wherever possible. The caveat to that is that there are some behaviours which are online specific, and therefore minimum standards of behaviour should be directed at the prevention of harm to users of the online platform by other users.

Precise minimum standards of behaviours are probably best left for each platform to decide upon given that they all offer unique experiences. For example, online games offer different experiences and whilst all have rules of play / terms of use, some games encourage immoral and /or violent, and/or harmful behaviours – such as for example, Grand Theft Auto. What is an acceptable minimum behaviour on one platform may be unacceptable on another. As such, one-size-fits-all is an undesirable, and perhaps unachievable aim.

16. How frequently should users be required to 'accept' or re-acknowledge terms of use, standards and policies?

Whenever there is a change to the terms of use, standards and policies users ought to be required to re-acknowledge them.

If significant emphasis is to be placed on the terms of use, there should be an obligation imposed on platforms to require acknowledgment of users on an annual basis. This could then

[23] J Pedersen, 'Automated Notices for Copyright Infringement: Pitfalls and Remedies.' Columbia Science & Technology Law Review, 28 November 2017. http://stlr.org/2017/11/28/automated-notices-for-copyright-infringement-pitfalls-and-remedies/?cn-reloaded=1.

[24] YouTube, 'How Content ID Works' https://support.google.com/youtube/answer/2797370?hl=en; J Bailey, 'YouTube Beta Testing Content ID For Everyone' Plagiarism Today (2 May 2018) https://www.plagiarismtoday.com/2018/05/02/youtube-beta-testing-content-id-for-everyone/.

tie in to the reporting obligations on platforms (see response to questions 33 & 34 below). It could for example, be useful to see how many users respond to changes in terms and conditions before e.g. access is restricted. Requiring more than simple acknowledgement of changes to policies could be one mode of increasing accountability and culpability of users.

Beyond this, terms and conditions should expressly include provisions that binds a user to other users and imposes obligations in terms of acceptable behaviours – something not seen in all terms and conditions.[25] These should be rigorously enforced by the platforms within the limitations provided in law that do not require active monitoring obligations by platforms.[26] Where there is a user in violation of the terms and conditions, enforcement action should be taken.

17. How should users be required to verify acceptance of terms of use, standards and policies?

If the emphasis is to fall on verifying and accepting terms of use / terms and conditions, something much more than scrolling to the end of the page, ticking a box and clicking 'continue' is required, especially if there are to be consequences arising from a breach of terms. For those of capacity (including age) this could involve a digital signature (akin to digital signatures for receipt of parcels, completed via mobile device) or for devices with biometrics built-in, fingerprint / face scan authorisation. This could follow in some form along the lines of the e-IDAS Framework[27] within the European Union for the verification of e-signatures – especially advanced electronic signatures – focussing on the unique identifier attached to the signer, followed by authentication.

18. Are there positive examples of improving user experience currently in use?

First, user experience ought to be understood as experience of *all* users, not just vulnerable groups such as children. Furthermore, whilst the current consultation heavily focuses on the safety of children online, greater consideration needs to be placed on the responsibilities of children and adolescents who actively participate on the Internet and through social media platforms in particular.

---

[25] K Barker, 'MMORPGing - The Legalities of Game Play', European Journal for Law and Technology, Vol. 3, No. 1, 2012, at [6].

[26] E.g. within the EU, platform providers benefit from protections under the E-Commerce Directive (Article 15) which does not impose any monitoring obligations on platforms. Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market.

[27] Regulation on electronic identification and trust services for electronic transactions in the internal market 910/2014/EU. (The 'e-IDAS' Regulation).

As the example of the tragic suicide of the British teenager, Molly Russell,[28] demonstrates, children and adolescents can be harmed by the content they view online and/or by the online abuse they experience. However, children and adolescents can act as – and often are – perpetrators of online abuse and the impact of their actions on other users can be severe – not least through cyberbullying but also through the perpetration of image-based sexual abuse. As such, a more complex understanding of children and adolescents is needed – one that views them as actors exercising their agency online, rather than solely as vulnerable actors and potential victims. An approach which fosters a greater understanding of responsibilities of participating online is greatly needed.

However, that certainly brings about legal complexities, not least due to different definitions of who is a child across various jurisdictions. Children are defined differently in various areas of the law – not least when it comes to the age of criminal responsibility or determination of legal capacity (for instance in the context of medical treatment[29]). Also, each of the platform's terms & conditions indicate the required age that a child needs to have reached in order to obtain in order access to the platform. Nonetheless, it is not uncommon for children and adolescents to falsely declare their age in order to participate on these platforms. Furthermore, the age limit prescribed by platform providers frequently stands in contrast with contractual understandings of the age of capacity.[30]

More broadly, it is crucial that improvements to user experiences are considered through experiences of all users and all potential vulnerabilities – including demographic characteristics. Within that, it is crucial that a gender perspective is meaningfully incorporated throughout and that steps are taken to improve the experiences (and safety) of users who experience gender-based and sexual abuse online, including gender-based online hate.[31] A gender perspective should also be a feature in designing any responses to children's and adolescents' online safety, reflecting the fact that children are not a homogenous group, and their user experiences will inevitably vary depending not only on their gender but also other characteristics. Furthermore, if harm perpetrated by children goes unchecked, or harm perpetrated against children is unregulated, these young Internet users will be exposed to behaviours and acts that seem 'acceptable' – this will breed future problems for the Internet and its users as they age.

[28] H Bodkin, 'Molly Russell: The "caring soul" who died after exploring her depression on Instagram' The Telegraph (27 January 2019) https://www.telegraph.co.uk/news/2019/01/27/molly-russell-caring-soul-died-exploring-depression-social-media/.

[29] E.g. 'Gillick competence.' *Gillick* v *West Norfolk & Wisbeck Area Health Authority* [1986] AC 112.

[30] See for example Facebook, which indicates that users are to be of age 13 before creating an account, and that to do so for an under-13 amounts to a violation of its terms. Facebook, 'How do I report a child under the age of 13?' Facebook Help Centre, https://m.facebook.com/help/157793540954833.

[31] K Barker & O Jurasz, *Online Misogyny as a Hate Crime: a Challenge for Legal Regulation* (Routledge 2019).

19. Are there positive examples of user support systems and processes currently in use? What are the factors and characteristics of these systems and processes that make them effective?

User support systems adopted by platform providers thus far are largely limited to mechanisms enabling muting abusive content / abusive users (Twitter / Facebook) or reporting. However, the latter has been largely unsuccessful in curtailing abusive behaviours online and offers limited redress to the victims. As for the former, muting merely masks the problem rather than constructively resolving it. This is particularly problematic in the context of online violence against women and text-based (sexual) abuse[32] as platform providers and law enforcement agencies generally ignore this as a problem requiring any tangible or lasting response.[33] Although gender stereotyping, violence against women, and misogyny are significant social problems which are pervasive both offline and online, the vast majority of regulatory efforts have been skewed towards addressing other problems, such as the posting of extremist content online.[34]

The effectiveness of the extremist / terror content online regulation discussions has largely arisen through the imposition of fines on the platforms for failing to act. In Germany, these fines imposed under the so-called 'NetzDG law' are even higher than those imposed beyond national state levels.[35] Whilst financial penalties are not necessarily the most useful means of ensuring action by platforms, it is possible that significant financial penalties can be used as a means of ensuring there is compliance.

20. What timeframe is reasonable to respond to complaints and reports?

The timeframes for responding to complaints and reports should be longer than the timeframes required for responding to takedown requests but should still be reasonable and timely so as not to unduly prejudice nor interfere with freedom of expression / free speech rights.

There should be longer timescales in operation for issues where a complaint is made. That said, an automated acknowledgement should be triggered when a complaint or report is made. This should outline the response times for the complainer, and inform the complainer as to the procedure that will be followed in producing a response. In situations where a complaint is

[32] K Barker & O Jurasz, *Online Misogyny as a Hate Crime: A Challenge for Legal Regulation?* (Routledge 2019) xiv.
[33] B Quinn, 'Met police chief backs calls to focus on violent crime not misogyny' The Guardian (2 November 2018) https://www.theguardian.com/uk-news/2018/nov/02/metropolitan-police-chief-cressida-dick-backs-call-focus-violent-crime-misogyny?CMP=Share_iOSApp_Other; E Stewart, 'Why is it so hard for women to get justice for online abuse?' ABC News (1 March 2016).
[34] Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online COM(2018) 640 final. See also: Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online C(2018) 1177 (final).
[35] Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (2017) [German Network Enforcement Law (2017)].

made as a result of takedown and stay-down, there should be an oversight body to which issues can be referred in the event that the complaint is not dealt with satisfactorily.

21. Should reporting and complaint response timeframes vary depending on the complainant (e.g. child or adult), the type of content or other factors?

Reporting and complaint response timeframes should vary depending on the complaint category, and the nature of the content. Categorising reporting based on the category of user could be problematic especially given that there is a – sometimes flawed – notion that children are always victims and never perpetrators of illegal or harmful online content (see answer to question 18 above).

If the reporting and complaint response timeframes were to vary depending on the complainant, this could give rise to situations where harmful content to children is addressed before – for example – extremist or terror related content. In an ideal scenario there would not have to be any need for a hierarchy. A more pragmatic approach suggests that a hierarchical response is exactly what is required, and responses should be prioritised according to the harm caused.

Where there are variations in response times, these must be based on the nature of the content and the categorisation of it rather than the person or persons concerned (see answer to question 6 above).

~~*22. What options are there for verifying age or ensuring that parental/guardian consent is provided? Is there an optimal method or methods?*~~

~~*23. Are there positive examples of parental settings currently in use?*~~

~~*24. Are there barriers to obtaining or using parental controls? How can these barriers be managed and overcome?*~~

~~*25. Are there positive examples of user content management options currently in use?*~~

~~*26. What user-controlled content management options should be available?*~~

~~*27. Are there positive examples of age appropriate products or services currently available?*~~

~~*28. To what extent do any technology firms restrict privacy and control settings as a default for younger users? If so, please provide detail.*~~

~~*29. Are there other positive examples of age guidance in the supply chain currently in use?*~~

~~*30. Do any technology firms have mandatory requirements for products and services to be designed and marketed as suitable for children?*~~

~~*31. Who should be responsible for ensuring built-in child safety?*~~

~~*32. Should relationships and engagement with independent experts be formalised, and what are the best mechanisms to achieve timely and productive input?*~~

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

Reporting should be required on the following (as a minimum):

- The number of posts (by jurisdiction e.g. Australian posts)
- The number of incidents reported (in total)
- The number of incidents of illegal content
- The number of incidents of harmful (but not illegal) content
- The total percentage / number of removals
- The percentage / number of removals of illegal content
- The percentage / number of removals of harmful (but not illegal content)
- The number of complaints received into removals
- The number of complaints which saw content reinstated
- The number of complaints which have a gender element to the threatening, abusive, or illegal content
- The number of employees reviewing illegal, and harmful content
- The number of employees with mental health and / or bias training
- A breakdown of action taken for the content removed e.g. removal of content, suspension of user, deletion of user account etc.
- Total number of 'flaggers'
- The contents of the terms of use / rules of conduct / acceptable behaviours terms and conditions
- The number of changes to the terms of use / rules of conduct / acceptable behaviours terms and conditions in the relevant reporting period

The UK Government draft transparency reporting template introduces 8 categories of content which could be the subject of a complaint.[36] This is a recommended starting point. As the interactive user-led nature of online content develops, these categories may need to be expanded. The list of categories should not therefore operate as a closed, exhaustive list.

[36] HM Government, 'Government response to the Internet Safety Strategy Green Paper' (May 2018) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/708873/Government_Response_to_the_Internet_Safety_Strategy_Green_Paper_-_Final.pdf, 67.

34. *How often should reporting take place? The UK requires country-specific reporting. To what extent should a similar arrangement be developed in Australia?*

The introduction of reporting in the UK requires social media companies to provide reports on an annual basis. This should be the minimum requirement. A more frequent requirement of reporting will be a significant burden. A less frequent reporting cycle than annually is likely to mean that should there be issues or a significant change in the approaches to content moderation, these could go unchecked for at least 12 months.

Any reporting requirements should operate in conjunction with an obligation imposed on technology firms to report on transparency and moderation guidelines.[37] This obligation should be enshrined in law and should be followed up with accountability to the appropriate Government body or Minister.

To ensure compliance with the annual reporting, and transparency obligations there must be sanctions imposed via the Online Safety Charter or through other legal mechanisms. Reporting requirements under a Charter which operates without sanction could lead to situations where large technology firms do not comply with the requirements. Compliance should be required to ensure a consistent approach across all firms and the sector more broadly. Failure to have consistency across the sector could lead to situations where vulnerable groups and/or victims are subjected to different standards and different responses across platforms. This is undesirable.

[37] HM Government, 'Government response to the Internet Safety Strategy Green Paper' (May 2018) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/708873/Government_Response_to_the_Internet_Safety_Strategy_Green_Paper_-_Final.pdf, 9.