Consultation on a draft Online Safety (Basic Online Safety Expectations) Determination 2021

GLOBAL PARTNERS DIGITAL

DIGITAL RIGHTS WATCH

Global Partners Digital & Digital Rights Watch - Joint Submission
November 2021

About Global Partners Digital

Global Partners Digital is a social purpose company dedicated to fostering a digital environment underpinned by human rights.[1]

About Digital Rights Watch

Digital Rights Watch is a charity organisation founded in 2016 whose mission is to ensure that people in Australia are equipped, empowered and enabled to uphold their digital rights.[2]

Introduction

Global Partners Digital and Digital Rights Watch welcome the opportunity to provide feedback on a draft Online Safety (Basic Online Safety Expectations) Determination 2021. We recognise the legitimate desire of the Australian government to place greater responsibility on service providers to ensure they provide safer services to Australian end-users. Based on our analysis, however, we believe that particular aspects of the draft Expectations, if taken forward in their current form, may pose risks to individuals' rights to freedom of expression, security, and privacy online and could be inconsistent with Australia's international human rights obligations.

In this joint response, we relay our concerns and make a series of recommendations on how the proposals could be revised to mitigate these risks. We believe these considerations and recommendations, if incorporated into the final instrument, will help safeguard freedom of expression, security, and privacy online.

6 Expectations - Provider Will Take Reasonable Steps to Ensure Safe Use

- *Additional expectation* - (2) The provider of the service will take reasonable steps to proactively minimise the extent to which material or activity on the service is or may be unlawful or harmful.

We are particularly concerned about the inclusion of this additional expectation and the corresponding *reasonable step (a) "developing and implementing processes to detect, moderate,*

[1] Learn more about our work at: https://www.gp-digital.org
[2] Learn more about our work at: https://digitalrightswatch.org.au

report and remove (as applicable) material or activity on the service that is or may be unlawful or harmful" as they may pose considerable risks to freedom of expression online.

Given the scale of content which is generated and shared online, providers will increasingly turn to automated processes, including AI, to meet this core expectation without the inclusion of any additional expectations. Larger companies tend to develop their own bespoke tools, whereas smaller companies may have to purchase or license generic tools for adaptation to their platform. However, the risk of encouraging or mandating the use of AI is that automated processes will detect and remove content that is not actually unlawful or harmful in a particular context.

Automated processes have had some success in relation to content moderation with types of images, including the ability to scan for copies of images that have already been identified by humans as constituting child sexual abuse and exploitation. But automated processing has been less effective at interpreting speech or less specific forms of content, as highlighted in our previous submissions on the Online Safety Act.[3] Cyber-bullying, cyber abuse, and material which promotes abhorrent violent conduct, as well as other types of harms addressed under the framework, may include a mixture of audio, visual and text content. Automated processes for the detection of such material thus rely on a combination of natural language processing, image recognition and contextual knowledge-mapping for detection, technologies which, at present, are somewhat limited. For example, a recent survey of machine learning techniques for cyber bullying detection on Twitter demonstrated a huge variation in accuracy of different models, which ranged from 30 to 80 percent.[4]

It is broadly recognised that these automated technologies struggle with novel content and domains and with inferring users' intentions through context. There is, therefore, a substantial risk that relying upon automated processes to proactively minimise harmful content will result in the removal of content which is entirely permissible due to error. We recommend that this additional expectation be removed from the proposal and that no additional expectations or reasonable steps be included which incentivise the use of automated processes to proactively monitor and remove content. If automated decision-making is undertaken to meet this additional expectation, platforms must be required to use open source tools, commit to transparency around standards, and implement appropriate appeals mechanisms.

Recommendation 1: The proposal should not compel or incentivise the use of automated processes to proactively monitor and remove harmful content, which has been proven to result in the removal of lawful and legitimate content online. Therefore, we recommend that the proposal exclude the additional expectation which would require services to proactively minimise the extent to which material or activity on the service is or may be unlawful or harmful, as well as the corresponding reasonable step to develop and implement processes to detect, moderate, report and remove (as applicable) material or activity on the service that is or may be unlawful or harmful.

Recommendation 2: If automated processes, such as those used for content flagging, are undertaken by entities to comply with the core expectation, the eSafety Commissioner should

[3] Digital Rights Watch, Submission on the proposed Online Safety Bill 2020 (February 2021), available at: https://digitalrightswatch.org.au/wp-content/uploads/2021/02/Submission_-Online-Safety-Bill-February-2021.pdf; and Global Partners Digital, Submission on Draft Online Safety Bill (February 2021), available at: https://www.gp-digital.org/wp-content/uploads/2021/02/GPD-Draft-Submission-for-new-Online-Safety-Act-2021.pdf

[4] Amgad Muneer and Suliman Mohamad Fati, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter", 12(11) Future Internet (October 2020), available at: https://www.mdpi.com/1999-5903/12/11/187

be required to offer rules and guidance regarding rigorous testing of automated tools prior to roll-out through expert consultation and trials. Such trials must be accompanied by human oversight and adequate appeals mechanisms, and be regularly assessed for their impacts on users' human rights.

We are also concerned that the inclusion of this additional expectation and reasonable step would result in discriminatory implementation, posing risks to individuals' right to non-discrimination. This is because proactive monitoring can be influenced by algorithmic bias, which is inevitable in virtually any automated process due to the availability of particular types of data for training the algorithm, the types of value judgements used to tag that data for training, or the biases and blind spots of those developing and testing the tool. In our previous submissions, we pointed out that automated content moderation tools have been shown to disproportionately penalise content from - and afford unequal protection to - marginalised communities and minority groups. For example, automated content moderation is more likely to flag nudity of Black, Indigenous, or people of colour, as well as overweight people.[5] When tested, automated language processors trained on widely used datasets of hate speech have been shown to be up to two times more likely to label tweets by Black people as offensive compared to other users.[6] The use of such tools thus poses a clear risk to individuals' right to non-discrimination.

In addition to removing this additional expectation and reasonable step, we believe the proposal could be further improved by requiring robust impact assessments when AI tools are voluntarily employed - specifically with regard to bias - to assess whether the use of automated processes results in, or could result in, differential treatment of any group based on a prohibited ground of discrimination under domestic law or under Article 26 of the International Covenant on Civil and Political Rights.

Recommendation 3: We recommend the proposal include explicit recognition of Australia's obligation to uphold the right to freedom of expression and non-discrimination under international human rights law.

Recommendation 4: We recommend the proposal require robust impact assessments of AI tools - specifically with regard to bias - to assess whether providers use of automated processes does not result in differential treatment of any group based on a prohibited ground of discrimination under domestic law or under Article 26 of the International Covenant on Civil and Political Rights. The eSafety Commissioner should provide guidance to providers regarding the requirements of impact assessments.

[5] See, Nosheen Iqbal, "Instagram censorship of black model's photo reignites claims of race bias" The Guardian (2021), available at: https://www.theguardian.com/technology/2020/aug/09/instagrams-censorship-of-black-models-photo-shoot-reignites-claims-of-race-bias-nyome-nicholas-williams; and Kevin Rennie, "Nude Photos of Australian Aboriginal Women Trigger Facebook Account Suspensions", Global Voices (2016), available at: https://advox.globalvoices.org

[6] Maarten Sap & Al, "The Risk of Racial Bias in Hate Speech Detection" Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019), available at: https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf

8 Additional expectation—provider will take reasonable steps regarding encrypted services

- *Additional Expectation* - If the service uses encryption, the provider of the service will take reasonable steps to develop and implement processes to detect and address material or activity on the service that is or may be unlawful or harmful.

We are particularly concerned with the inclusion of this additional expectation and the detrimental impact it would have on individual and collective digital security. This additional expectation currently frames encryption as an inhibitor to safety, which runs counter to the consensus in the cybersecurity industry who view encryption as vital to *facilitate* safety. Encryption is essential for all businesses, individuals, and digital security at a national level. Encryption facilitates the security of our online activities; protecting data from potential cybercriminals, enabling secure online transactions, and maintaining the privacy and security of our online communications, including those of children. For example, encryption plays a crucial role in preventing malicious actors from accessing networked devices, including tapping into users' webcams or baby monitors. This additional obligation would potentially undermine the security of Australians' encrypted services, jeopardizing the safety of the millions of people who rely on them each day. Prompting services to weaken, undermine, or otherwise bypass encryption threatens the digital security of Australia at a national level by introducing security weaknesses into everyday services used by Australians. Further, encryption is essential for the protection of vulnerable groups, including LGBTQ+ persons[7] and survivors of domestic violence, who rely on encryption to protect the sharing of information about safe relocation, the integrity of digital evidence, and to guard against unauthorised access to survivors' details or communications.[8] The inclusion of this additional expectation risks undermining the first core expectation of the BOSE: that the digital platform will take reasonable steps to ensure that end-users are able to use the service in a safe manner.

We are also concerned that the inclusion of this additional expectation would potentially have a detrimental impact on individuals' right to privacy and freedom of expression. Privacy is a gateway to the enjoyment of other rights, particularly the right to freedom of expression. Encryption provides individuals with a zone of privacy online to hold opinions and exercise freedom of expression. Mandating that the providers of encrypted services take reasonable steps to detect and address material would almost certainly amount to an unjustifiable restriction on individuals' right to communicate privately. This is because such services use, almost universally, end-to-end encryption, limiting (although not eliminating) the ability of providers to filter or monitor content which is generated or shared using them. Compliance with this additional expectation would be unfeasible unless those services weakened or ceased to use end-to-end encryption, which would amount to an unjustifiable restriction on the right to privacy and a potential chilling effect on freedom of expression.

The utility of this additional expectation is also questionable at best. Research by Tech Against Terrorism indicates that, contrary to the rationale underpinning policymakers' calls for weakening encryption or inserting backdoors, there is no guarantee that systematic monitoring of encrypted content will be efficient in countering and disrupting criminal activities, as

[7] LGBT Tech & ISOC, Encryption - Essential for the LGBTQ+ Community, available at: https://www.lgbttech.org/post/2019/11/22/lgbt-tech-release-encryption-one-sheet

[8] ISOC, Understanding Encryption Fact Sheet: The Connections to Survivor Safety, (2020) available at: https://www.internetsociety.org/resources/doc/2020/understanding-encryption-the-connections-to-survivor-safety/

undermining end-to-end encryption on mainstream platforms contributes to the migration of threat actors to alternative or potentially non-cooperative platforms.[9]

We recommend that this additional expectation, as currently worded, be removed from the proposal because of the risks that it may pose to individuals rights, collective digital security, and the failure of the government to establish the necessity and proportionality of this obligation. Alternatively, it should be limited to, and reflective of, the reasonable steps included in the accompanying FAQ document - which explicitly notes that reasonable steps include a range of actions, such as detecting misuse through behavioral, account or online signals including routing information and metadata and closing accounts.

We further suggest that the proposal explicitly note that companies are not required to cease, restrict or in any way weaken their use of encryption or other privacy-enhancing technologies to satisfy core expectations. For example, Section 317ZG of the Telecommunications and Other Legislation Amendment (Assistance and Access) Act 2018 provides that designated communications providers must not be requested or required to implement or build a systematic weakness, or a systemic vulnerability, into a form of electronic protection, which includes one or more actions that would render systematic methods of authentication or encryption less effective.[10] While we do not necessarily support the precise scope and limited nature of this limitation, we believe that this precedent would support the inclusion of additional safeguards within the proposal.

Instead, we recommend that the proposal include expectations or reasonable steps which do not seek to undermine encryption or pose risks to individuals' right to communicate privately. For example, having providers of encrypted services focus on strengthening user-reporting systems, encouraging victims to report abuse, improving response times to such reports, and ideally providing a contact point for victims throughout the process to build trust in the reporting process. The platform should also communicate clearly to victims what they can expect of the platform in terms of redress.

Recommendation 5: The additional expectation which would require providers to take reasonable steps regarding encrypted services should, as currently worded, be removed from the proposal because of the risks that it may pose to individuals rights, collective digital security, and the failure of the government to establish the necessity and proportionality of this obligation.

Recommendation 6: If this additional expectation is to be included, we recommend that it should be limited to, and reflective of, the reasonable steps included in the accompanying FAQ document - which explicitly notes that reasonable steps include a range of actions, such as detecting misuse through behavioral, account or online signals including routing information and metadata and closing accounts. We further suggest that the proposal explicitly note that providers are not required to cease, restrict or in any way weaken their use of encryption or other privacy-enhancing technologies to satisfy core expectations.

[9] Tech Against Terrorism, "Terrorist Use of E2EE: State of Play, Misconceptions, and Mitigation Strategies" Report Summary, available at: https://www.techagainstterrorism.org/wp-content/uploads/2021/09/TAT-Terrorist-use-of-E2EE-and-mitigation-strategies-report-Executive-summary.pdf

[10] Telecommunications and Other Legislation Amendment (Assistance and Access) Act 2018, available at: https://www.legislation.gov.au/Details/C2018A00148

Recommendation 7: We recommend that the proposal only include expectations or reasonable steps which do not seek to undermine encryption or pose risks to individuals' right to communicate privately. For example, having providers focus on strengthening user-reporting systems, encouraging victims to report abuse, improving response times to such reports, and ideally providing a contact point for victims throughout the process to build trust in the reporting process.

9 Additional expectation—provider will take reasonable steps regarding anonymous accounts

- *Additional expectation* - (1) If the service permits the use of anonymous accounts, the provider of the service will take reasonable steps to prevent those accounts being used to deal with material, or for activity, that is or may be unlawful or harmful.

- *Reasonable steps that could be taken* - (2) Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following:
  (a) having processes that prevent the same person from repeatedly using anonymous accounts to post material, or to engage in activity that is unlawful or harmful; (b) having processes that require verification of identity or ownership of accounts.

We are concerned with the inclusion of this additional expectation and the two reasonable steps as these pose risks to individuals' right to privacy and freedom of expression online. Anonymity, like encryption, enables individuals to exercise their rights to freedom of opinions and expression in the digital age. Members of vulnerable communities, dissidents, human rights activists, journalists, whistleblowers, and victim-survivors of domestic abuse rely on anonymity online to maintain their safety.[11]

Restrictions on anonymity, as envisioned here, must be provided for by law, must be in pursuance of a legitimate aim, and must conform to the strict tests of necessity and proportionality.[12] We do not believe that this additional expectation and the corresponding reasonable steps would be considered necessary or proportionate. We note that the Australian Privacy Principles under the Privacy Act provides that individuals must have the option of dealing anonymously or by pseudonym with an APP entity (APP 2). While there are exceptions to this, we are concerned that the inclusion of this expectation would amount to a broad carve out from APP 2, and undermine the essence with which it was written. It is also unclear whether the expectation or reasonable steps would be technically feasible or effective. And even if technologically feasible, these restrictions would still be considered disproportionate and unjustified due to their widespread negative impacts on individuals' right to privacy.

These elements of the proposal would undermine anonymity through the use of technical approaches to verify user accounts. At present this often involves two-factor authentication with a phone or email address, but such processes are easy to manipulate for users who want to remain anonymous. Other methods involve the use of ID verification, which requires individuals to provide identifying information or documents, including through cross-referencing systems or using third-party identification services. These systems pose risks to individual privacy, including through data leaks or the mismanagement of personal information. Moreover, any requirements for ID may result in access issues for those who are less likely to have official forms of ID. These

[11] Digital Rights Watch, "Anonymity Online is Important", (2021) available at: https://digitalrightswatch.org.au/2021/04/30/explainer-anonymity-online-is-important/
[12] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report on encryption, anonymity, and the human rights framework, (2015) A/HRC/29/32.

types of requirements may also pose heightened risks to marginalised groups, specifically LGBTQ youth or victim-survivors of domestic violence, who may have their documents held by parents or abusers, allowing them to monitor or restrict online activity.

The necessity of this additional expectation and the reasonable steps are unsubstantiated. We are unaware of relevant data on the prevalence of harms perpetrated by anonymous accounts vs. non-anonymous accounts, or the degree to which anonymity encourages or incites greater incidence of NCSII, cyber-abuse, cyber-bullying, or violent abhorrent content in Australia. On the contrary to these prevalent assumptions, research has indicated that anonymity is often a tool that empowers and protects individuals.[13] It is therefore unclear what removing anonymity would concretely achieve without any evidence or analysis to support this. This is previously supported by the eSafety Commissioner, who has rejected the practicality of such efforts, as well as noting that "it would create a range of other issues and that removing the ability for anonymity or to use a pseudonym is unlikely to deter cyberbullying and the like".[14]

We therefore recommend that the additional expectation and reasonable steps be removed from the proposal. If the government does keep the additional expectation, we recommend that they provide for alternative reasonable steps to satisfy the additional expectation which does not undermine anonymity or privacy online. For example, requiring platforms to increase user controls for user interaction with anonymous accounts. This could involve limited verification of accounts to simply determine whether they are real or not, or making verification optional for users, and then allowing them to choose the types of content or accounts they interact with. This approach would incentivise verification, but not mandate it, and thus reflect a more proportionate approach.

Recommendation 8: We recommend that the additional expectation relating to anonymous accounts and corresponding reasonable steps be removed from the proposal.

Recommendation 9: If the government does keep the additional expectation, we recommend that they provide for alternative reasonable steps to satisfy the additional expectation which does not undermine anonymity or privacy online. For example, requiring platforms to instead increase user controls for interaction with anonymous accounts, or to incentivise as opposed to mandating user verification.

11 Core expectation—provider will take reasonable steps to minimise provision of certain material

- The provider of the service will take reasonable steps to minimise the extent to which the following material is provided on the service:
  (a) cyber-bullying material targeted at an Australian child;
  (b) cyber-abuse material targeted at an Australian adult;
  (c) a non-consensual intimate image of a person;
  (d) class 1 material;
  (e) material that promotes abhorrent violent conduct;
  (f) material that incites abhorrent violent conduct;

[13] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report on encryption, anonymity, and the human rights framework, (2015) A/HRC/29/32.
[14] Asha Barbaschow, "eSafety thinks identity verification for social media would be impractical", ZDNET: https://www.zdnet.com/article/esafety-thinks-identity-verification-for-social-media-would-be-impractical/

(g) material that instructs in abhorrent violent conduct;
(h) material that depicts abhorrent violent conduct.

Prompting services to proactively remove material that **depicts** abhorrent violent material is likely to result in proactive removal of content documenting human rights abuses, or incidents of state violence (such as disproportionate use of force by police). This content, while upsetting, is essential in the pursuit of justice and for accountability. There is a difference between material that **promotes, incites or instructs** abhorrent violent conduct and material that depicts it.

Simply preventing people from seeing violent material does not solve the underlying issues causing violence in the first place, and it does not create justice or avenues of redress. It is essential that this expectation does not facilitate the hiding of state use of violence or abuses of human rights.

As a core expectation, we understand that it cannot be amended in the determination, however, we would recommend that some additional safeguards or alternatives are included to ensure that vital political content is less likely to be proactively removed, at the detriment of justice, accountability, and democratic processes. We therefore recommend that an additional expectation be included to expressly note that service providers are not expected to remove content of political importance, but that they may consider alternative reasonable steps, such as placing violent or offensive content behind a sensitive warning, so that users can actively choose to reveal the content behind it.

Recommendation 10: We recommend an additional expectation to expressly note that service providers should not remove content which has important political value, and to provide for alternative reasonable steps to minimise the content, for example, by placing it behind a sensitive warning. The eSafety Commissioner should provide guidance on determining political importance, to reduce the onus of ambiguity on providers.