GLOBAL PARTNERS DIGITAL

Online Safety Legislative Reform

Global Partners Digital response

February 2020

About Global Partners Digital

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

Table of contents

Introduction

We welcome the consultation on a new Online Safety Act to consolidate Australia's current legislative framework and update it in light of changes to the online environment. GPD recognises the legitimate desire of the Australian government to tackle unlawful and harmful content online, and the majority of the proposals put forward in the discussion paper are reasonable and sensible.

Based on our analysis, however, we believe that certain aspects of the proposals, if taken forward in their current form, pose risks to individuals' right to freedom of expression or privacy online and could be inconsistent with Australia's international human rights obligations. It is particularly important that these obligations are upheld given that Australia has limited constitutional protections for freedom of expression and privacy.

While we respond to the relevant questions posed in the discussion paper and make a series of recommendations on how the proposals should be refined, these refinements should be accompanied with further commitments by the government to ensure that the proposals do not put the right to freedom of expression or privacy at risk. Through a full human rights analysis of the proposals, we make further specific recommendations on how the proposals should be revised in order to mitigate those risks as far as possible, including through the incorporation of further safeguards.

Framework for analysis of the proposals and consultation

Our analysis of the proposals in the discussion paper and the consultation questions asked is based on international human rights law, specifically the International Covenant of Civil and Political Rights (ICCPR). The most relevant human right impacted by the proposals is the right to freedom of expression. This is recognised by the government in the discussion paper itself, where it states that "the Act is seeking to: (…) balance the competing objectives of user safety and freedom of expression". (p. 19)

Article 19 of the ICCPR guarantees the right to freedom of expression, including the right to receive and impart information and ideas of all kinds regardless of frontiers.[1] Restrictions to the freedom of expression or privacy guaranteed under international human rights law are only lawful when they can be justified. In order to be justified, any restriction must meet a three-part test, namely that: (1) restrictions are provided by law; (2) restrictions pursue one of the purposes set out in Article 19(3) of the ICCPR - to protect the rights or reputations of others, to protect national security or public order, or public health or morals; and (3) restrictions must be necessary and proportionate, which requires that the restriction be the least restrictive means required to achieve the purported aim.[2]

It is important to remember that Australia's obligation to ensure that this right is not unjustifiably restricted exists both in relation to restrictions which stem from the actions of the state itself as well as those caused by third parties, such as private companies. As such, it makes no difference from the perspective of the individual affected whether any restrictions are imposed and enforced directly by the state (e.g. through creating criminal offences which are enforced by the police and the courts) or through third parties, particularly when the third party is acting in order to comply with legal obligations.

[1] See, in particular, Article 19 of the International Covenant on Civil and Political Rights. The right to freedom of expression is also protected in other treaties, such as Article 13 the Convention on the Rights of the Child.

[2] UN Human Rights Committee, General Comment No. 34, Article 19: Freedoms of opinion and expression, UN Doc. CCPR/C/GC/34, 12 September 2011.

With respect to the actions of private companies specifically, the United Nations Guiding Principles on Business and Human Rights (UNGPs) makes clear that a state's international human rights obligations include establishing a legal and policy framework which enables and supports businesses to respect human rights. Principle 3 notes that this general obligation includes ensuring "that (...) laws and policies governing the creation and ongoing operation of business enterprises, such as corporate law, do not constrain but enable business respect for human rights".

Given the impact that online platforms have upon the enjoyment and exercise of the rights to freedom of expression and privacy, the government has a clear obligation to ensure that these rights are respected by these platforms. This includes ensuring that legislation and other measures do not constrain online platforms' ability to respect the right to freedom of expression or privacy themselves, nor should they directly or indirectly constitute a restriction on the enjoyment and exercise of those rights by those that use those platforms.

Our analysis of the regulatory measures proposed in the discussion paper and our subsequent recommendations are based on these frameworks. Given the limited existing interpretation and case law of these frameworks as they apply to measures comparable to those proposed in the discussion paper, we also make reference, as appropriate, to relevant commentary from the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (the UN Special Rapporteur).

Though not a framework for the purpose of our analysis, we note that Australia has, through its membership of the Freedom Online Coalition, signed up to a number of commitments which are relevant to the subject. These includes commitments made in the "Recommendations for Freedom Online, Adopted in Tallinn, Estonia on April 28, 2014 by Ministers of the Freedom Online Coalition":

> "We, the members of the Freedom Online Coalition
>
> 4. Dedicate ourselves, in conducting our own activities, to respect our human rights obligations, as well as the principles of the rule of law, legitimate purpose, non-arbitrariness, effective oversight, and transparency, and call upon others to do the same,
>
> (...)
>
> 6. Call upon governments worldwide to promote transparency and independent, effective domestic oversight related to electronic surveillance, use of content take-down notices, limitations or restrictions on online content or user access and other similar measures, while committing ourselves to do the same".[3]

More recent commitments were made in the Freedom Online Coalition's "Joint Statement on Internet Censorship":

> "In 2017, the world witnessed state-sponsored Internet censorship in various forms: states have manipulated and suppressed online expression protected by international law, have subjected users to arbitrary or unlawful surveillance, have used liability laws to force ICT companies to

[3] Recommendations for Freedom Online, Adopted in Tallinn, Estonia on April 28, 2014 by Ministers of the Freedom Online Coalition, available at: https://www.freedomonlinecoalition.com/wp-content/uploads/2014/04/FOC-recommendations-consensus.pdf.

> self-censor expression protected by international law, have disrupted networks to deny users access to information, and have employed elaborate technical measures to maintain their online censorship capabilities. Further unlawful efforts included state censorship in private messaging apps and systematic bans of news websites and social media. Likewise certain states have introduced or implemented laws which permit executive authorities to limit content, on the Internet broadly and without appropriate procedural safeguards. Individuals who may face multiple and intersecting forms of discrimination, including women and girls, often faced disproportionate levels of censorship and punishment.
>
> (...)
>
> The FOC firmly believes in the value of free and informed political debate, offline and online, and its positive effects on long term political stability. The Coalition calls on governments, the private sector, international organizations, civil society, and Internet stakeholders to work together toward a shared approach - firmly grounded in respect for international human rights law - that aims to evaluate, respond to, and if necessary, remedy state-sponsored efforts to restrict, moderate, or manipulate online content, and that calls for greater transparency of private Internet companies' mediation, automation, and remedial policies".[4]

Finally, we note that while Australia has a long tradition of commitment to human rights and supporting human rights internationally, international processes have recently cited concerns about the risk to freedom of expression and privacy posed by Australian legislation. Australia received a recommendation in its most recent Universal Periodic Review (UPR) that the government "[t]ake concrete measures in order to ensure that any interference with the right to privacy comply with the principles of legality, proportionality and necessity, regardless of the nationality or location of the individuals affected".[5] Moreover, the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression and the UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism noted their conerns in a recent comment on the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Law 2019.[6] As Australia will face scrutiny over its compliance with international human rights law during its upcoming UPR in 2020, developing the proposed legislation in manner that protects human rights online would provide an opportunity for Australia to demonstrate a commitment to human rights on the global stage.

[4] The Freedom Online Coalition, Joint Statement on Internet Censorship, available at: https://freedomonlinecoalition.com/wp-content/uploads/2018/05/FOC-Joint-Statement-on-Internet-Censorship-0518.pdf.

[5] Report of the Working Group on the Universal Periodic Review, UN Doc. A/HRC/31/14, 13 January 2016, Para 136.227.

[6] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, UN Doc. OL AUS 5/2019, 4 April 2019

1. The basic online safety expectations

Penalties for non-compliance

While we have few concerns in relation to the basic online safety expectations (BOSE), we are concerned that reporting companies under BOSE may be incentivised to proactively remove permissible forms of expression to avoid potential penalties for non-compliance with the proposed reporting requirements. The discussion paper provides that penalties "will include the capacity for the eSafety Commissioner to publish a statement that a reporting social media service is not complying with the basic online safety expectations". This presents a potential risk, albeit a limited one, to freedom of expression as this could encourage companies to proactively remove permissible content in an effort to remain in compliance and avoid the public fallout associated with a public statement by the eSafety Commissioner. This risk could be mitigated if companies did not feel that they were at risk of censure if they made a mistake moderating content, but were instead able to receive support from the eSafety Commissioner as an initial step.

Recommendation 1: The BOSE should make clear that any social media service or potential entity within scope has the ability to request further guidance from the eSafety Commissioner where they reasonably believe that upholding the BOSE might undermine their ability to safeguard the right to freedom of expression.

The use of automated processes and artificial intelligence

Our response to question 4 highlights our concern with the Online Safety Charter's Service Provider Responsibilities and, in particular, section 1.5 which provides that services should "[p]ut processes in place to detect, surface, flag and remove illegal and harmful conduct, contact and content with the aim or prevent harms before they occur". The scale of content that is shared online today would require the use of automated processes to comply with a general monitoring obligation as it is impossible for human moderators to review all content. This appears to be recognised by section 1.5 itself which goes on to require that services "[w]here feasible and appropriate to the service, utilise technology to 'fingerprint' content that has been identified as illegal or harmful and deploy systems to prevent the attempted upload, re-upload or sharing of this material". While part of this responsibility appears to relate to content already identified as illegal, we are concerned about the use of automated processes outside this limited context and the risks to freedom of expression.

Automated processes have had some success in relation to content moderation with types of images, including the ability to identify copies of images that have already identified by humans as constituting child sexual abuse and exploitation. However, automated processing has been less effective when identifying speech or less specific forms of illegal or harmful content.[7] As noted by the UN Special Rapporteur:

> "AI-driven content moderation has several limitations, including the challenge of assessing context and taking into account widespread variation of language cues, meaning and linguistic and cultural particularities. Because AI applications are often grounded in datasets that incorporate discriminatory assumptions, and under

[7] See, for example, Center for Democracy & Technology, "Mixed Messages? The Limits of Automated Social Media Content Analysis", 28 November 2017, available at: https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf

> circumstances in which the cost of over- moderation is low, there is a high risk that such systems will default to the removal of online content or suspension of accounts that are not problematic and that content will be removed in accordance with biased or discriminatory concepts. As a result, vulnerable groups are the most likely to be disadvantaged by AI content moderation systems."[8]

To give two examples, while a human might have the ability to discern the difference between a joke made to a friend and a legitimate threat of violence, even the most advanced automated processing tools are unable to determine context and therefore differentiate between the two. Similarly, a journalist could upload a video of a war crime as a means of drawing attention to a horrific crime, but an automated process might flag and remove this as terrorist content.

Due to this inability to recognise context, and the evidence of inaccurate decision making by automated tools when it comes to many forms of content moderation, the use of automated content moderation tools would risk the inadvertent removal of content which is lawful and harmless.

Recommendation 2: The BOSE should not require companies within scope to filter content at the point of upload nor require the use of artificial intelligence or other forms of automated decision-making. If automated decision-making is undertaken by companies within scope, this should be accompanied by requirements to ensure the use of open source tools, transparency around standards, and appropriate appeals mechanisms.

2. The four schemes (cyberbullying, cyber abuse for adults, image-based abuse, and illegal and harmful content), which would all require removal of certain types of content, on order of the eSafety Commissioner, within 24 hours

Scope of harms

We have no concerns in relation to the scope of the first three schemes (cyberbullying, cyber abuse for adults, image-based abuse). With regards to the fourth (seriously harmful content), while we are pleased that the proposed definition would be based on content that is currently illegal under the Commonwealth Criminal Code, we are concerned about the potential expansion of this definition by the Minister. The discussion paper indicates that the Minister would be provided with the power to make a legislative instrument that captures additional types of content based on the advice of the eSafety Commissioner. However, any change to the proposed definition should be made via primary legislation to ensure that there is comprehensive democratic oversight of any types of content.

The current proposed definition (child sexual abuse material, abhorrent violent material, and content that promotes, incites or instructs in serious crime), is limited to clearly defined and illegal content. It is therefore likely to meet the legality and legitimate aim requirements for a permissible restriction on freedom of expression. This has the potential to change if the Minister includes additional forms of content, such as virtual reality or animated content, that are not clearly defined or prohibited by law.

[8] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc. A/73/348, 29 August 2018. Para 15.

Recommendation 3: Any change to the proposed definition of seriously harmful content made by the Minister should be made via primary legislation.

Recommendation 4: There should be a requirement that any new form of seriously harmful content be already clearly defined and prohibited under Australian law.

Appeals mechanism

We are concerned over the lack of adequate appeals mechanisms for take-down notices issued by the eSafety Commissioner to private companies and end-users. The discussion paper does not propose making private entities decide whether a particular piece of content is unlawful or not. It would instead be the eSafety Commissioner, a public body, that would make this decision. This is a welcome approach that avoids the privatisation of law enforcement. However, while decision making by a public body can provide a far greater level of transparency and accountability, judicial redress should be provided for across all schemes. Part 10 of the Enhancing Online Safety Act 2015 provides some ability for private companies or end-users to challenge decisions of the eSafety Commissioner in the Administrative Appeals Tribunal, but it is unclear that this will extend to the new legislation. We are also concerned that this ability to challenge decisions, even if incorporated, would be insufficient. It is important that there be an appropriate appeals mechanism for take-downs as international human rights law, specifically Article 2(3) of the ICCPR, requires that any person whose rights or freedoms are violated shall have an effective remedy. This is especially needed for the proposed legislation as civil proceedings and other forms of redress are often cumbersome and expensive. Meaningful opportunities to challenge decisions should be readily available and accessible to the public.

Recommendation 5: The proposed take-down schemes should enable all end-users and private companies the opportunity to challenge decisions made by the eSafety Commissioner. The eSafety Commissioner itself should have the resources to provide an effective remedy.

3. The codes of practice for other forms of illegal and harmful content.

We are concerned about one aspect of the proposed principles-based industry codes to address other forms of illegal and harmful content, namely the requirement that all sectors of industry provide their users with access to the best available technology solutions to prevent children's access to harmful content. Some of the new technologies mentioned in the discussion paper, including forms of facial recognition technology, pose a potential risk to the right to privacy. Article 17(1) of the ICCPR provides that "no one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence".[9] As provided for in the most recent Human Rights Council Resolution on "The Right to Privacy in the Digital Age":

> *"Recognizing* that, despite its positive effects, the use of artificial intelligence that requires the processing of large amounts of data, often relating to personal data, including on an individual's behaviour, social relationships, private preferences and identity, can pose serious risks to the right to privacy, in particular when employed for identification,

[9] ICCPR, Article 17.

> tracking, profiling, facial recognition, behavioural prediction or the scoring of individuals,"[10]

Accordingly, it is unlikely that the utilisation of these technologies would be considered the least restrictive instrument to achieve the purported goal. Other forms of technology that pose less serious risks to the right to privacy could instead be utilised to prevent children's access to harmful content.

Recommendation 6: Industry codes should not require service providers to use the "best available technology", such as facial recognition technologies, if there are any concerns that usage could lead to an impermissible restriction on human rights. We also recommend that the codes require companies to consider the potential human rights impact of new technologies used.

4. Blocking measures for terrorist and extreme violent material online

Development of the protocol

While we welcome the improvements made to the scheme for blocking terrorist and extreme violent material online, we are concerned that the proposed power of the eSafety Commissioner to direct ISPs to block domains containing terrorist or extreme violent material might still lead to impermissible restrictions on freedom of expression. The eSafety Commissioner is required to develop a protocol for the use of the new power. This protocol will provide guidance on the circumstances in which it is anticipated that this power may be used. It will also set out the processes to be used to determine whether the terrorist or extreme violent material is sufficiently serious to warrant blocking action, the means of determining which ISPS would be subject to the blocking orders, and for how long.

Despite assertions in the discussion paper that this new power would only be utilised for limited periods of time and in the event of an online crisis event, there is no guarantee that the final protocol will adhere to the principles of proportionality and necessity. To satisfy the third limb of the three-part test for a permissible restriction on freedom of expression, restrictions must be necessary and the least restrictive means required to achieve the purported aim. Here, it is possible that the protocol will establish a low threshold for the circumstances that the power may be utilised, as there is no clear definition for what constitutes an "online crisis event". Furthermore, the protocol could contain processes which allow the government to "play it safe" and either include ISPs that aren't necessary or maintain the block for longer than necessary.

Recommendation 7: The protocol developed by the eSafety Commissioner should be designed in a way that adheres to the principles of necessity and proportionality, providing for a framework that only utilises the least restrictive means of tackling an online crisis event. It should provide a limited set of circumstances in which the power may be used, establish a high threshold for determining which ISPs are subject to blocking orders, and limit the time that ISPs are required to implement the blocks to the shortest time frame possible.

[10] Resolution adopted by the Human Rights Council "The right to privacy in the digital age", UN Doc. A/HRC/42/15, 7 October 2019.

Appeals mechanism

The discussion paper makes it clear that the eSafety Commissioner is required to notify owners of affected domains that their services have been blocked, and provide for appropriate appeal and review mechanisms. We welcome this approach as international human rights law, specifically Article 2(3) of the ICCPR, requires that any person whose rights or freedoms are violated shall have an effective remedy. Yet, we are concerned that the "appropriate appeal and review mechanism" might be insufficient as it may not promptly rectify the violation or provide meaningful reparation. The time sensitive nature of these potential violations would make a remedy meaningless if the applicable blocking period had already passed.

Recommendation 8: The "appropriate appeal and review mechanism" established should ensure that potential human rights violations are promptly managed, and that violations be utilised to inform more human rights respecting policies moving forward.

Further considerations

We are concerned that this proposed power does not consider how the blocking mechanism might infringe upon the ability of journalists to report on terrorist-related events or similar online crisis related activity. Restrictions on the right to freedom of expression require an assessment of the proportionality of the relevant measures, with the aim of ensuring that restrictions "target a specific objective and do not unduly intrude upon the rights of targeted persons".[11] The absence of exceptions or considerations of journalists risks disproportionately impairing the public's right to access vital reporting on terrorist related events. Section 474.37 of the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019 provides a defence for journalists who live stream abhorrent violent material in their professional capacity as journalists. While this particular interpretation of journalism has been criticised as overly narrow by several UN Special Rapporteurs,[12] it does acknowledge the need for exceptions in relevant legislation.

Recommendation 9: The proposed blocking measures for terrorist and extreme violent material online should provide exceptions for journalists as to not impair the public's right to access vital information. It should also include exceptions for researchers who might require access to information that becomes blocked.

The consultation questions posed in the discussion paper

Objects of the new Act

1. **Are the proposed high level objects appropriate? Are there any additions or alternatives that are warranted?**

   As the purpose of the objects section is to set out the underlying purposes for a piece of legislation which can be used to aid interpretation of detailed provisions, including by courts, it is important that the objects reflect a commitment to human rights.

[11] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc. A/HRC/29/32, 22 May 2015, para 35.
[12] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, UN Doc. OL AUS 5/2019, 4 April 2019

Recommendation 10: We recommend adding a high level object that reflects a commitment to human rights. For example, "ensuring the protection of human rights online, including the right to freedom of expression and privacy".

2. **Is the proposed statement of regulatory policy sufficiently broad to address online harms in Australia? Are there aspects of the proposed principles that should be modified or omitted, or are there other principles that should be considered?**

   These proposed principles should be modified to more clearly recognise the impact these proposals might have on human rights. Australia has limited constitutional or legislative protection of freedom of expression so it is critical that this new online safety legislation contains effective protections and explicitly recognises the importance of human rights online. Additional references to human rights would be a welcome modification to these proposed principles.

Recommendation 11: We recommend that the proposed principle "balance the competing objective of user safety and freedom of expression" be modified to "uphold the right to freedom of expression". This alternative text emphasises the importance of freedom of expression as a standalone objective, rather than linking it to user safety.

Recommendation 12: We recommend that the proposed principle "encourage the development and use of new technologies and safe products and services" be modified to "encourage the development and the use of human rights-respecting technologies and safe products and services".

Basic online safety expectations

3. **Is there merit in the BOSE concept?**

   There is merit in the BOSE concept, particularly since it recognises the importance of existing initiatives and collaboration with government and civil society. It is designed to improve transparency and intended to align with voluntary efforts already underway in Australia. By building on existing efforts, BOSE is likely to be less onerous and more likely to achieve results when compared to an entirely novel process. It will hopefully enable tech companies to learn from previous failures and build on current successes.

4. **Are there matters (other than those canvassed in the Charter) that should be considered for the BOSE? Are there any matters in the Charter that should not be part of the BOSE?**

   While most matters included in the Charter should be considered for the BOSE, we have some concern over section 1.5 of the Charter. This section indicates that service providers should "[p]ut processes in place to detect, surface, flag and remove illegal and harmful conduct, contact and content with the aim of preventing harms before they occur". This could encourage proactive monitoring of content and the unintentional removal of permissible content. Given the scale of content which is generated and shared online, companies will turn to automated processes, including AI, to meet their obligations under the BOSE. As discussed above, automated processes may detect and remove content that is not actually illegal or harmful in a particular context. Existing research demonstrates that AI is limited in its ability to effectively analyse categories of speech as it is not able to

recognise context or nuance, which has been shown to lead to over-removal of content.[13] Automated processes have also been shown to risk further marginalisation and censorship of groups that already face discrimination.[14]

5. **What factors should be considered by the eSafety Commissioner in determining particular entities that are required to adhere to transparency reporting requirements (e.g. size, number of Australian users, history of upheld complaints)?**

   A risk-based approach should be taken into consideration by the eSafety Commissioner in determining particular entities that are required to adhere to transparency reporting requirements. The size, number of Australian users, and history of upheld complaints should all be considered. However, it is important for the eSafety Commissioner to keep up to date with new entities that become popular with the Australian public and vulnerable groups in particular. It is also imperative that the means for determining the particular entities that are required to adhere to transparency reporting requirements do not incentivise potential entities to impermissibly restrict legitimate expression online.

6. **Should there be sanctions for companies that fail to meet the BOSE, beyond the proposed reporting and publication arrangements?**

   The discussion paper notes that the government is not proposing to impose sanctions for non-compliance with the BOSE at this stage. The government should continue this policy as heavy or disproportionate sanctions will skew incentives and only increase risks to freedom of expression. The introduction of sanctions for non-compliance might encourage platforms to "play it safe" and simply remove permissible content rather than risk a potential sanction. This could create a chilling effect on the right to freedom of expression.

Cyberbullying scheme

7. **Is the proposed expansion of the cyberbullying scheme for children to designated internet services and hosting services, in addition to relevant electronic service and social media services, appropriate?**

   The proposed expansion is appropriate as it recognises that cyberbullying and harms suffered by children online do not only take place on large social media services. Broadening the range of service providers covered may however place an unreasonable burden on certain companies to comply with particular elements of the new scheme, which is expanded upon above and in response to the following questions.

8. **Is the proposed take-down period of 24 hours reasonable, or should this require take-down in a shorter period of time?**

   The proposed take-down period of 24 hours might be reasonable to larger companies that have the means and capacity of quickly removing content. Companies that have not been included under the existing scheme, including gaming services or 'confessional' platforms, may need more time to respond to such orders if considered relevant entities. A more

[13] See, for example, Center for Democracy & Technology, "Mixed Messages? The Limits of Automated Social Media Content Analysis", 28 November 2017, available at: https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf
[14] *Ibid*., pp 8-9.

flexible time frame could be beneficial for certain entities that may need more time to comply.

9. **What are the likely compliance burdens of the proposed changes to the cyberbullying scheme on small and large businesses?**

   Large businesses should be able to comply with the proposed changes, particularly those in Tier 1 or Tier 2 of the existing cyberbullying scheme. On the other hand, new services and smaller companies may initially face challenges in responding to take-down orders by the eSafety Commissioner. These companies may need to create devoted internal structures or invest significant financial resources. The success of the existing scheme, which involved the development of an effective and collaborative partnership between online service providers and social media providers, could be replicated to offset these substantial burdens with smaller entities.

10. **What other tools could the eSafety Commissioner utilise to effectively address cyberbullying in the circumstances where social media service and end-user notices are not well suited to the particular service upon which the cyberbullying has occurred?**

    Some of the additional tools proposed in the discussion paper could effectively address cyberbullying in circumstances where social media service and end-user notices are not adequate, such as requiring platforms to enforce their terms of service in relation to a user. Other means such as requiring account restrictions might be a disproportionate response and impermissibly restrict valid forms of online expression. It is important for the eSafety Commissioner to utilise the least restrictive means of achieving a particular objective and provide a means for appealing any restriction.

Establishing a new cyber abuse scheme for adults

11. **Is the proposed application of the cyberbullying and cyber abuse schemes to designated internet services and hosting services, relevant electronic service and social media services, appropriate?**

    Yes, the proposed application of the cyberbullying and cyber abuse schemes to designated internet services and hosting services, relevant service and social media services is appropriate. The previous construction of the scheme was not comprehensive enough to tackle the issue. Both children and adults face abuse and harassment across a range of platforms and services. The reporting and take-down scheme should therefore apply to adults as well. It also focuses on minimising harm to the victim as opposed to strictly restricting or removing content.

12. **Is the proposed take-down period of 24 hours reasonable, or should this require take-down in a shorter period of time?**

    As noted in our response to question 8, the proposed take-down period of 24 hours should be reasonable to larger companies that have the means and capacity of quickly removing content. A more flexible time frame could be beneficial for certain entities that may need more time to comply.

13. **Do the proposed elements of a definition of adult cyber abuse appropriately balance the protection from harms with the expectation that adults should be able to express views freely, including robust differences of opinion?**

    The proposed elements of a definition of adult cyber abuse do appropriately balance the protection from harm with the expectation that adults should be able to express views freely, including robust differences of opinion.

14. **Should the penalties differ under a cyber abuse scheme for adults and the cyberbullying scheme for children?**

    The penalties should differ under a cyber abuse scheme for adults and the cyberbullying scheme for children. The cyberbullying scheme for children pertains to "harmful" but not necessarily illegal content, whereas the cyber abuse scheme for adults involves content that mirrors the construction of offence provisions under Australian law. Therefore, it is appropriate to have penalties for adult end users that are in violation of the cyber abuse scheme. Applying this to the cyberbullying scheme for children risks creating two separate regimes for what is considered legal expression on and offline. In addition, the violators of the cyberbullying scheme for children will likely be children themselves. It would be inappropriate to penalise them as you would an adult.

15. **What additional tools or processes, in addition to removal notices, could be made available to the eSafety Commissioner to address cyber abuse occurring across the full range of services used by Australians?**

    N/A

Non-consensual sharing of intimate images (image-based abuse)

16. **Is the proposed take-down period for the image-based abuse scheme of 24 hours reasonable, or should this require take-down in a shorter period of time?**

    Yes, the proposed take-down period for the image-based abuse scheme of 24 hours is reasonable. As noted in our response to questions 8 and 12, the proposed take-down period of 24 hours should be reasonable to larger companies that have the means and capacity of quickly removing content. A more flexible time frame could be beneficial for certain entities that may need more time to comply.

17. **Does the image-based abuse scheme require any other modifications or updates to remain fit for purpose?**

    N/A

18. **What additional tools or processes, in addition to removal notices, could be made available to the eSafety Commissioner to address image-based abuse being perpetrated across the range of services used by Australians?**

    N/A

Addressing illegal and harmful online content

19. Is the proposed application of the take-down powers in the revised online content scheme appropriate?

Yes, this is appropriate as it streamlines the process for determining harmful content, which would be done by the eSafety Commissioner and not require referral to the Classification Board. However, we are concerned about the risk of extraterritorial removal of content as the proposed scheme would enable the eSafety Commissioner to issue take-down notices to content hosted outside of Australia. Extraterritorial take-downs could risk creating inconsistent regimes which apply in the same jurisdiction, particularly when a company might not be able to geo block certain content.

20. Are there other methods to manage access to harmful online content that should be considered in the new Online Safety Act?

N/A

21. Are there services that should be covered by the new online content scheme other than social media services, relevant electronic services and designated internet services?

N/A

22. Is the proposed take-down period of 24 hours for the online content scheme reasonable or should this require take-down in a shorter period of time?

The proposed take-down period of 24 hours for the online content scheme is reasonable for seriously harmful content. Seriously harmful content should be prioritised for take-down by relevant services, but some entities may still have trouble complying with these orders. For example, international entities may require more time to comply with such requests and may triage their efforts based on the specific content in question. We understand that overseas-hosted material forms the vast majority of the prohibited online content actioned by the eSafety Commissioner so the ability of these international entities needs to be considered.

Opt-in tools and services to restrict access to inappropriate content

27. When evaluating opt-in tools and services for accreditation, what criteria should be considered?

We are concerned that if industry code is updated to require service providers to use the best available technology to prevent children's access to harmful content that some tools may pose risks to human rights. For example, the discussion paper makes reference to facial recognition technology as a potential tool. As indicated above, this technology may lead to a potential infringement upon the the right to privacy in certain cases. Human rights considerations should be incorporated for all opt-in tools and services utilised or required under the codes.

Blocking measures for terrorist and extreme violent material online

28. Is the proposed scope of content blocking for online crisis events appropriate?

The proposed scope of content blocking for online crisis events is an improvement from its current form. Unlike the existing power, which is broad in nature and has led to concerns about its potential impacts on freedom of expression, the proposed scope establishes a specific and targeted power. It is proposed that the eSafety Commissioner would only be able to block websites for a certain time period and not on an ongoing basis. This more narrowly tailored response is preferred as it is less likely to infringe upon permissible forms of expression and is a proportionate response.

We welcome this approach with regard to blocking measures for terrorist and extreme violent material, but are concerned that the final protocol developed by the eSafety Commissioner would not establish appropriate arrangements and processes for implementing these powers. The means of determining which ISPs would be subject to blocking orders and the length of time that the ISPs will be required to implement the blocks may result in a broad or unlimited power as opposed to a specific and targeted one. Moreover, guidance on the circumstances in which it is anticipated that this power may be used by the eSafety Commissioner should be limited in scope as to only apply to necessary circumstances.

29. Are there adequate appeals mechanisms available?

The requirement of the eSafety Commissioner to notify owners of affected domains that their services had been blocked and provide for an appropriate appeal and review mechanism could be adequate. International human rights law requires the right to effective remedy when rights have been infringed upon, which seems to be satisfied through the proposed notifications and appeals scheme. Yet, we are concerned that the "appropriate appeal and review mechanism" might be insufficient as it may not promptly rectify the violation or provide meaningful reparation. The time sensitive nature of these potential violations would make a remedy meaningless if the applicable blocking period had already passed.

30. What other elements of a protocol may need to be considered?

It is important that the protocol consider how the blocking mechanism might infringe upon the ability of journalists to report on terrorists or other online crisis related activity. Furthermore, there should be exemptions provided for academics who engage in research on the topic and would require access to certain material. If not, this might be an impermissible restriction on freedom of expression that would be at odds with the discussion paper's emphasis on balancing online safety during crisis events with "broader principles of freedom of expression".

Role of eSafety Commissioner

36. Are the eSafety Commissioner's functions still fit for purpose? Is anything missing?

It is important that the eSafety Commissioner's office has a dedicated staff with sufficient knowledge and expertise to effectively meet the proposed functions. The eSafety Commissioner will need to make informed decisions that could potentially encroach or infringe upon freedom of expression, particularly when issuing take-down orders for content, and especially those which require an understanding of the context of the

content. We believe it is important that these decisions be made according to clear criteria that requires a consideration of freedom of expression.

Recommendation 13: The eSafety Commissioner should seek external advice to ensure that they have the necessary expertise to carry out the respective functions of the office. We further recommend that decisions on take-down orders require an explicit consideration as to whether the decision constitutes an impermissible restriction on freedom of expression. The eSafety Commissioner should consult with relevant experts and develop further guidance or a policy statement as to how they will be making determinations, particularly in relation to types of content which require an understanding of the context of the content.

39. What are the likely impacts, including resource implications, on other agencies and businesses of a new Online Safety Act?

A new Online Safety Act would undoubtedly require additional resources for the eSafety Commissioner and the Australian court system if implemented according to the proposals outlined in the discussion paper. These entities will require significant financial resources, personnel and expertise to meet their responsibilities under the proposal. It is also important to consider how the higher burdens on private companies imposed by the proposals will affect the market and freedom of expression more broadly. We are concerned that higher regulatory burdens will reduce competition in the market, and power may be concentrated on a few large platforms or entities. This would lead to less places for individuals to express themselves online, which might ultimately affect freedom of expression in the aggregate.