facebook

Facebook's response to Australian Government consultation on a new Online Safety Act

19 February 2020

Executive summary

Facebook welcomes the opportunity to provide input to the Australian Government's consultation on a new Online Safety Act.

We recognise our responsibility to protect the safety of people who use Facebook's services - especially the safety of young people. It's essential to our business: Australians and other people around the world will only continue to use our platform if they feel welcome and safe.

Industry, government and the community all have a role to play in working towards online safety. To uphold our responsibility, we invest significantly in developing: policies, tools and reporting infrastructure for our platforms; technology that detects and removes harmful content proactively; and programs to support young people to have a safe and positive experience online. Our investments in technology have substantially increased our capability to protect people online: in 8 of the 10 policy areas that we cover in our Community Standards Enforcement Report, we proactively detected over 90% of the content we took action on before someone reported it.[1]

In addition to our global investments in online safety, we have steadily increased our efforts to protect online safety and wellbeing of Australians, especially young people, people living in regional Australia and Indigenous girls. Through the Digital Ambassadors Program, PROJECT ROCKIT has delivered online safety training to over 4,500 young Australians from 138 schools in all states and territories.[2] To ensure that young people in regional Australia are receiving the same opportunity as those living in metropolitan areas, we have delivered online safety training to over 3,210 young people across 9 schools in regional areas, as part of our Community Boost program. And we partnered with the Alannah and Madeline Foundation and the Stars Foundation to host online safety workshops for Indigenous girls as part of the Safe Sistas initiative in ten locations including Tennant Creek, Yirrakala and Mildura.[3]

We have also worked with headspace to help them provide targeted support to communities that have recently experienced suicides of young people.[4] To encourage better wellbeing and self-esteem around body image, we released the Own Your Feed and The Whole Me campaigns with

[1] See Facebook, *Community Standards Enforcement Report*, https://transparency.facebook.com/community-standards-enforcement; see also, https://about.fb.com/wp-content/uploads/2019/11/FB_CSER_Highlights_1113.pdf
[2] Project Rockit, *Digital Ambassadors*, https://www.projectrockit.com.au/digitalambassadors/
[3] Alannah & Madeline Foundation, *Helping Sistas be safer*, https://www.amf.org.au/news-events/latest-news/helping-sistas-be-safer/
[4] Headspace, *headspace and Facebook combat youth suicide*, https://digitalworkandstudy.org.au/blog/headspace-and-facebook-combat-youth-suicide/

the Butterfly Foundation.[5] And to support parents in engaging with young people about using Instagram safely, we released the Instagram Parents Guide together with ReachOut.[6]

To support the work of the Australian Government in promoting online safety, we have worked with the Office of the Australian eSafety Commissioner, since its establishment in 2015, to promptly respond to all complaints. The overwhelming majority of reports are resolved in under 30 minutes and - across all reports we have received from the eSafety Commissioner since 2015, including complicated cases that required further review - the median turnaround time is around 16 hours. We were also one of the first companies to publicly support the Commissioner's Safety by Design Principles and in July 2019, we hosted a Safety By Design Youth Jam to bring the Principles to life by engaging directly with young Australians and New Zealanders.[7]

We recognise the important role that legislative frameworks play, to hold companies to account for the steps they take to protect online safety and combat harmful content online. As our CEO Mark Zuckerberg has outlined, we are committed to working with governments to develop effective regulation for online content. Our CEO recently said: *"People need to feel that global technology platforms answer to someone, so regulation should hold companies accountable when they make mistakes."*[8] And we have released a white paper called *Chartering a Way Forward: Online Content Regulation* to propose models for best practice regulation of online content.[9]

In line with our strong commitment to safety in Australia, we broadly support the Government's enhancement of Australia's online safety regulatory framework. We believe this presents an opportunity to enhance the online safety of Australians and to establish a regime that holds companies to account for the commitments they make. In some areas, the paper could go further, such as strengthening the governance of the eSafety Commissioner and extending the Commissioner's remit to offline bullying and harassment.

However, we are concerned that the well-intentioned proposals in the paper inadvertently expand the government's powers so far that a regulator could intrude on private conversations between adults, have broad discretion over what Australians can say online, or get dragged into arbitrating online debates.

[5] *The Butterfly Foundation*, Instagram launches Own Your Feed campaign, https://thebutterflyfoundation.org.au/about-us/media-centre/media-releases/instagram-launches-own-your-feed-campaign-in-response-to-body-image-survey-fingings/; see also https://thebutterflyfoundation.org.au/support-us/the-whole-me/

[6] *ReachOut*, Instagram Parents Guide, https://parents.au.reachout.com/landing/parents-insta-guide

[7] Antigone Davis, *How can Facebook design products with interests of young people at their core?* https://www.facebook.com/notes/facebook-australia-new-zealand-policy/how-can-facebook-design-products-with-interests-of-young-people-at-their-core/2374348519559074/

[8] Mark Zuckerberg, *Big tech needs more regulation*, https://www.ft.com/content/602ec7ec-4f18-11ea-95a0-43d18ec715f5

[9] Monika Bickert, *Charting a way forward on online content regulation*, https://about.fb.com/news/2020/02/online-content-regulation/

There are three key areas where we have concerns.

Firstly, we are concerned about the proposal to extend online safety regulation to private messaging - especially between adults. The challenge is that a government takedown scheme can be a blunt instrument that is ill-suited for considering the nuances of social interactions between Australian adults. The consultation paper proposes granting a regulator the power to police not just social media posts but also private conversations between Australian adults for potential bullying and harassment. Given Australians can already use the tools available to them to block and delete harmful content in messaging apps, it is not clear that government intervention in private conversations is the best approach to protect online safety.

Secondly, while we support a set of baseline online safety principles, the proposed Basic Online Safety Expectations are confusing in scope. They go beyond what could be reasonably considered as “basic expectations” and instead represent a collation of all best practice safety tools and practices in the industry. Our submission proposes some alternative principles that we believe could help meet the expectations of governments and the community, via a more certain baseline for companies.

Finally, we believe the proposed new scheme for “harmful content” is overly broad and could be highly contested. The interaction with other legislation needs to be considered: for example, both the harmful content scheme and a cyberbullying scheme for adults would inevitably include hate speech or other types of content that needs to be carefully considered within the context of the Racial Discrimination Act, especially Section 18C.

Given the complexity of regulating harmful content, we encourage the Government to undertake a separate consultation with respect to harmful content, including hate speech, similar to the Online Harms White Paper process in the United Kingdom. This should involve consideration of the best metrics to measure success, such as prevalence. At Facebook, we consider prevalence to be a critical metric because it helps us measure how content violations impact people. We care most about how often content that violates our standards is actually seen relative to the total amount of times any content is seen on Facebook.

Summary of Facebook responses to the consultation paper questions

Questions	Facebook's Response
Objects of the new Act *1. Are the proposed high level objects appropriate? Are there any additions or alternatives that are warranted?* *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?*	We **support in principle** the proposal that Australia's online safety legislation includes high level objects, but we have concern about the lack of clarity around the concept of "online harms" and suggest that it would be preferable to initiate a new, separate review process to consider a regulatory scheme for harmful content more broadly.
Basic Online Safety Expectations *3. Is there merit in the BOSE concept?* *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?* *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)?* *6. Should there be sanctions for companies that fail to meet the BOSE, beyond the proposed reporting and publication arrangements?*	We **support in-principle** the concept of safety-based regulatory principles**, subject to** the following clarifications: • The role of the principles requires clarification: whether they are (as the name suggests) a basic set of minimum expectations, or whether they are (as the content suggests) an aspirational bar, set by collating all best practice safety tools and practices in the industry. a collation of all safety tools and practices in the industry. • The principles require much greater specificity. • The principles should be principle-based and technology-neutral, rather than seeking to prescribe specific business processes. • The principles should apply to all regulated entities under the scheme, not just social media companies. In an effort to provide constructive assistance, we have suggested a number of principles that we believe could form the basis of a clearer and effective principles-based regime.
General comments	We **do not support** the expansion of cyberbullying schemes to private messaging, as it is (1) a blunt instrument for government to intervene into private conversations; (2) not clearly beneficial given tools that digital platforms already make available to allow people to remove harmful content and block harmful content in messaging; and (3) does not reflect technical complexities of private messaging platforms.

	We **note** the proposal to focus on turnaround times as a marker of success for the scheme and to shorten it to 24 hours across all regulatory schemes, as it may result in unintended consequences, such as incentivising investment to respond more to content removal instead of investment in technology to reduce the prevalence of harmful online content.
Expansion of cyberbullying scheme for children *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?* *8. Is the proposed take-down period of 24 hours reasonable, or should this require take-down in a shorter period of time?* *9. What are the likely compliance burdens of the proposed changes to the cyberbullying scheme on small and large businesses?* *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?*	We **support** the expansion of the scheme, subject to reservations outlined above about the inclusion of private messaging and shorter takedown times. We do not support the proposed other tools proposed for the eSafety Commissioner. Instead, we believe that greater consideration should be given to the Commissioner's ability to target bullying and harassment more broadly in our society.
New cyberbullying 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?* *12. Is the proposed take-down period of 24 hours reasonable, or should this require take-down in a shorter period of time?* *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?* *14. Should the penalties differ under a cyber abuse scheme for adults and the cyberbullying scheme for children?* *15. What additional tools or processes, in addition to removal notices, could be made available to the eSafety Commissioner to address cyber abuse*	We **support** the creation of the scheme, subject to reservations outlined above about the inclusion of private messaging and shorter takedown times. We agree that bullying and harassment experienced by adults requires a higher threshold than for children, and we have made some suggestions about the definition that are intended to limit unintended consequences for freedom of expression. The scheme should be carefully designed to avoid inserting the government into political debates and discussion. We recommend the new cyberbullying scheme for adults should include an appeal mechanism for affected individuals that is faster and easier than existing broad appeal mechanisms for government decisions.

occurring across the full range of services used by Australians?	
Image-based abuse scheme *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?* *17. Does the image-based abuse scheme require any other modifications or updates to remain fit for purpose?* *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?*	We **support** a complaints-based scheme to ensure the swift removal of non-consensually shared intimate images and schemes that incentivise companies to invest in technology and systems to proactively prevent the sharing of this type of content on their services.
New Online Content Scheme *19. Is the proposed application of the take-down powers in the revised online content scheme appropriate?* *20. Are there other methods to manage access to harmful online content that should be considered in the new Online Safety Act?* *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?* *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?* *23. Which elements of the existing co-regulatory requirements should be retained under the new Act?*	We **support** integrating Schedules 5 and 7 of the Broadcasting Services Act into the new Online Safety Act. However, before determining a position on the design of the scheme, we believe much more discussion and work is required to clarify the types of content that should be captured - particularly content that is not illegal but is considered to be "harmful". The scope of content should be defined in legislation rather than set by legislative instrument. We believe it would be preferable to initiate a new, separate review process to consider a regulatory scheme for harmful content more broadly - rather than try to retrofit a safety legislation scheme onto all types of content.
Accreditation scheme *24. To what extent would an expanded accreditation scheme for opt-in tools and services assist parents and carers in mitigating the risk of access by minors to potentially harmful content?* *25. What categories of tools and services should be included in an accreditation program, aside from content filters?* *26. What are the likely costs of developing and maintaining an accreditation scheme for opt-in*	We **support in principle** the proposed accreditation scheme, subject to greater clarification about how this would operate.

tools and services to assist parents and carers in managing access to online content by minors? *27. When evaluating opt-in tools and services for accreditation, what criteria should be considered?*	
Content blocking for ISPS *28. Is the proposed scope of content blocking for online crisis events appropriate?* *29. Are there adequate appeals mechanisms available?* *30. What other elements of a protocol may need to be considered?*	We have **no comment** on this approach, as it does not relate to digital platforms
Ancillary service provider scheme *31. Is there merit in the concept of an ancillary service provider notice scheme?* *32. Are there any other types of services that should be included in the definition of ancillary service provider?* *33. Should the definition of search engine provider be broadened to include search functions housed in other services, such as social media services, video hosting services or other services with internal search functionality?* *34. Is the requirement that 3rd parties be systemically and repeatedly facilitating the posting of cyberbullying or cyber abuse material, image-based abuse or hosting illegal or harmful content appropriate before the eSafety Commissioner can issue a notice to an ancillary service provider? Should a different threshold be contemplated?* *35. Is there merit to making compliance with the ancillary service provider notices mandatory?*	Before determining a position on the design of the scheme, we believe much more discussion and work is required. It is not clear what harms the proposed ancillary service provider scheme is intended to address. Social media services are already captured under the takedown scheme (and potentially the proposed new Online Content Scheme), so it does not seem necessary to extend an ancillary service provider scheme to social media services.
Governance of the eSafety Commissioner *36. Are the eSafety Commissioner's functions still fit for purpose? Is anything missing?* *37. To what extent should the existing functions of the eSafety Commissioner be streamlined? Are there particular functions that need to be maintained, or new functions that should be specified?* *38. To what extent should the functions of the eSafety Commissioner be prioritised?*	We **support** the establishment of the eSafety Commissioner as an independent office, with the same oversight as comparable regulators like the Privacy Commissioner. We **recommend** the expansion of the remit of the eSafety Commissioner to include combatting offline bullying and harassment, to enable the Commissioner to undertake proactive steps to change the culture of bullying and harassment among children (including undertaking education programs, and consultation with students, parents and teachers).

Table of Contents

Facebook's commitment to safety

As context to the Government's consideration of enhancing Australia's online safety legislation, we wanted to share an overview of Facebook's work to promote the safety and wellbeing of all Australians, particularly young Australians.

Facebook is strongly committed to enhancing the online safety of people who use Facebook services. That's why we have invested in an industry-leading program of online safety that comprises five components:

1. Policies
2. Enforcement
3. Tools and products
4. Resources
5. Partnerships.

Firstly, we develop policies to keep people safe. Because we take seriously our role in keeping abuse off our service, we've developed a set of Community Standards[10] that outline what is and is not allowed on Facebook. Safety is a core value of our Community Standards.[11]

Our policies are based on feedback from our community and the advice of experts in fields such as technology, public safety and human rights, including experts based in Australia. To ensure that everyone's voice is valued, we take great care to craft policies that are inclusive of different views and beliefs, in particular those of people and communities that might otherwise be overlooked or marginalised. The Community Standards are regularly updated to keep pace with changes happening online and offline around the world.

Every two weeks, members of our Product Policy team, who sit in 11 offices around the world, run a meeting called the Product Policy Forum to discuss potential changes to our Community Standards, ads policies and major News Feed ranking changes. A variety of subject matter experts participate in this meeting, including members of our safety and cybersecurity policy teams, counterterrorism specialists, Community Operations employees, product managers, public policy leads and representatives from our legal, communications and diversity teams. In keeping with our commitment to greater transparency, the minutes of these meetings are made publicly available.[12]

Our Community Standards prohibit various categories of harmful content, from violent content, objectionable content and content that contravenes people's safety (including suicide and self-

[10] Facebook, *Community Standards,* https://www.facebook.com/communitystandards/

[11] Monika Bickert, *Updating the values that inform our community standards*, https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/

[12] See, for example, https://about.fb.com/news/2018/11/content-standards-forum-minutes/

injury, child exploitation, adult sexual exploitation, bullying and harassment, human exploitation, and privacy violations).

Secondly, to enforce our policies, we use a combination of automation and human review. We use technology and automation in a number of ways to enforce our policies, to help our teams of human reviewers perform faster and smarter, and to proactively detect and remove content that violates our Community Standards.

We publicly report our progress in enforcing our policies through the Community Standards Enforcement Report (CSER), which indicates the volume of content actioned across various categories of our Community Standards.

In our latest CSER, we disclosed that between July and September 2019:

- We removed 11.6 million pieces of *child nudity or child sexual exploitation* content, 99.5 per cent of which we identified proactively.
- We removed 5.2 million pieces of *terrorist propaganda* content, 98.5 per cent of which we identified proactively.
- We removed 25.2 million pieces of *violent and graphic* content, 98.6 per cent of which we identified proactively.
- We removed 3.2 million pieces of *bullying and harassment* content, 16.1 per cent of which we identified proactively. Because bullying and harassment are highly personal by nature, in many instances, we need a person to report this behaviour to us before we can identify or remove it. This means that using technology to proactively detect bullying and harassment can be more challenging than other violation types. We continue to invest in our proactive detection technology to ensure we're tackling the problem and protecting our community.

To assist others within industry and government to leverage our investments in automatic detection of harmful content, we made a major announcement that we were sharing some of our AI technologies that detect harmful content (called PDQ and TMK+PDQF) on an open source basis, to help build the capacity of our industry partners, smaller developers and not-for-profits.

In addition to insights on content that we have actioned, one of the most significant metrics we provide in the CSER is prevalence.[13] Focusing on prevalence allows consideration to be given to how often content that violates our standards is actually seen relative to the total amount of times any content is seen on Facebook.

The way content causes harm on the internet is by being seen. Given the nature of the internet, the amount of times content is seen is not evenly distributed. A small amount of content could go viral and get a lot of distribution in a very short span of time, whereas other content could be on

[13] Facebook, *Measuring Prevalence of Violating Content on Facebook*, https://about.fb.com/news/2019/05/measuring-prevalence/

the internet for a long time and not be seen by anyone. Any measure we use to understand our enforcement of harmful content should take that into consideration.

For this reason, we consider prevalence to be a critical metric because it helps us measure how violations impact people on Facebook. We care most about how often content that violates our standards is actually seen relative to the total amount of times any content is seen on Facebook.

Ideally, we would remove all violating content before anyone ever sees it, if it was possible to perfectly moderate content. In some cases, however, the content never being detected or reported in the first place is a bigger reason harmful content is seen. We need a measure that captures all of these reasons people may be exposed to harmful content. We believe prevalence is that measure. We now report prevalence across 7 of the 10 policy areas (up from 5 policy areas in the previous report) that we disclose in our CSER.

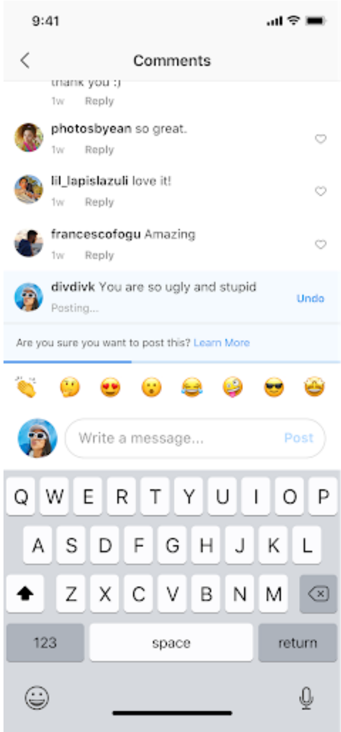
Thirdly, we design tools within our products to empower people to address potentially negative impacts from the use of our services.

In addition to the long-standing tools of Block, Report, Hide, Unfollow[14], some of the latest features we have recently announced include: tests on Facebook and Instagram in Australia and other countries to minimise social comparison by hiding the total number of likes on posts[15]; a Restrict tool in Instagram that protects accounts from unwanted interactions without making the people involved aware[16]; and a new feature on Instagram that encourages people to pause and reflect on a potentially hurtful comment or caption before posting it. A screenshot of the latter is provided below.

[14] An overview of these and other tools is available in the Facebook Safety Center: https://www.facebook.com/safety/tools

[15] Antigone Davis, *How can Facebook design products with interests of young people at their core?* https://www.facebook.com/notes/facebook-australia-new-zealand-policy/how-can-facebook-design-products-with-interests-of-young-people-at-their-core/2374348519559074/

[16] Instagram, *Introducing the Restrict feature to stand up against bullying*, https://about.instagram.com/blog/announcements/stand-up-against-bullying-with-restrict



thank you :)
1w Reply

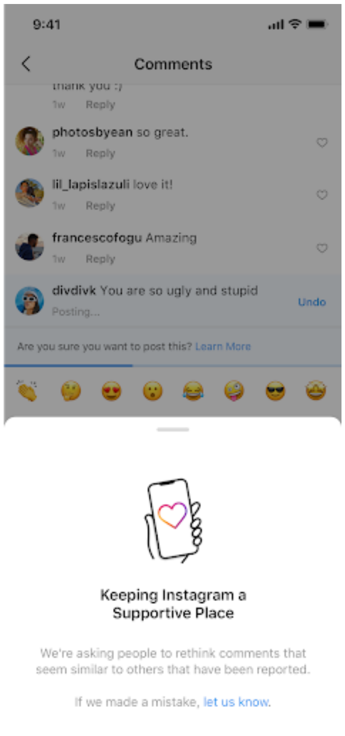
photosbyean so great.
1w Reply

lil_lapislazuli love it!
1w Reply

francescofogu Amazing
1w Reply

divdivk You are so ugly and stupid
Posting...

Are you sure you want to post this? Learn More



thank you :)
1w Reply

photosbyean so great.
1w Reply

lil_lapislazuli love it!
1w Reply

francescofogu Amazing
1w Reply

divdivk You are so ugly and stupid
Posting...

Are you sure you want to post this? Learn More

Keeping Instagram a Supportive Place

We're asking people to rethink comments that seem similar to others that have been reported.

If we made a mistake, let us know.

Fourthly, we make a variety of resources, that we have worked on often with third party experts, available to assist teachers, parents and other people to be safe online. This includes the Instagram safety and wellbeing hub[17] and the Facebook Safety Center[18]. Within the Facebook Safety Center, there are various additional resources including:

- Our Bullying Prevention Hub developed in partnership with the Yale Centre for Emotional Intelligence;
- A Digital Literacy Library, which is a collection of lessons to help young people think critically and share thoughtfully online, developed together with the Berkman Klein Center for Internet & Society at Harvard University;
- The Parents Portal that contains information and tips for parents to help foster conversations among parents and their children about staying safe online; and,
- Not Without My Consent, which provides information about our pilot to combat non-consensually shared intimate images.

Finally, we have launched a number of partnerships - including in Australia - as part of our comprehensive safety programs.

[17] *Instagram safety and wellbeing hub*, https://about.instagram.com/community
[18] *Facebook Safety Center*, https://www.facebook.com/safety/educators

1. We have invested $1 million in a Digital Ambassadors program delivered by PROJECT ROCKIT. Digital Ambassadors is a youth-led, peer-based anti-bullying initiative. A Digital Ambassador aims to utilise credible strategies to safely connect and tackle online hate.[19]
2. We also work with the Alannah and Madeline Foundation and the Stars Foundation on the Safe Sistas program, which support the online safety of young Indigenous women to respond to the issue of non-consensually shared intimate images.[20]
3. As part of our Community Boost with Facebook for Regional Australia initiative, we are investing in the digital skills of regional communities. Technology can be a great equaliser for a country defined by distance such as Australia and through our Community Boost initiative, we are working to ensure that people in regional Australia have access to the same information about how to have a safe and positive experience online, as people who live in metropolitan areas. Since 2018, we have trained over 3,210 young people across 9 schools in regional Australia with tips and insights about how to have a safe and supportive experience online.
4. In September 2018 and in December 2019 respectively, we released the Own Your Feed and The Whole Me campaigns with the Butterfly Foundation, to help people make sure their time on social media is positive, inspiring, safe and empowering.[21] The campaigns aim to support better mental health and self-esteem in relation to how people feel about how they look.
5. To ensure that young Australians get support messages when they need it, we have worked with headspace since 2019 to provide them with support to promote messages and advice to people in their News Feeds, who are living in towns, regions or communities that have recently experience the suicide of young people.
6. To support parents to understand the tools that are available on Instagram, we worked with ReachOut to develop an Instagram Parents Guide that contains suggested conversation starters to better understand how their teens are using Instagram and how to ensure they are using it safely and positively.[22] We released the Guide in September 2019.

And we partnered with a number of organisations - including the eSafety Commissioner's Office - in July 2019 to deliver a Safety by Design Jam. The Safety by Design Jam was a workshop designed specifically for young people, which sought to gather insights and feedback from the people best placed to talk about youth safety online - young people themselves. It was the first of five age-appropriate Design Jams around the world, aiming to bring together policymakers, academics,

[19] Project Rockit, *Digital Ambassadors*, https://www.projectrockit.com.au/digitalambassadors/

[20] Alannah & Madeline Foundation, *Helping Sistas be safer*, https://www.amf.org.au/news-events/latest-news/helping-sistas-be-safer/

[21] *The Butterfly Foundation*, Instagram launches Own Your Feed campaign, https://thebutterflyfoundation.org.au/about-us/media-centre/media-releases/instagram-launches-own-your-feed-campaign-in-response-to-body-image-survey-fingings/; see also https://thebutterflyfoundation.org.au/support-us/the-whole-me/

[22] *ReachOut*, Instagram Parents Guide, https://parents.au.reachout.com/landing/parents-insta-guide

safety and privacy experts, and, of course, young people, to share new ideas and perspectives about how we can build age-appropriate experiences on our products that meet the needs and expectations of our young community.

General comments

Before responding to the specific questions asked in the consultation paper, we wanted to share some general comments on themes that emerge across the proposals, to assist the Government to shape the overall legislative amendments. These general comments relate to setting the regulatory frameworks for harmful content, the inclusion of private messaging within regulatory scope, and the focus on time-to-action rather than prevalence.

Regulating harmful content

Several aspects of the proposed new online safety regulation propose to regulate online harms or harmful content. Facebook has been at the global forefront of calling for regulation of harmful content. Twelve months ago, our CEO Mark Zuckerberg called for liberal democracies to develop new regulation in relation to online content (along with privacy, data portability and elections)[23], a call that he reiterated in the Financial Times in the last fortnight.[24]

At the same time, we released a white paper called *Charting a Way Forward - Online Content Regulation*, which raises a series of questions to assist in designing effective content regulation. Many of those questions relate to the questions posed in the Australian Government's consultation on a new Online Safety Act.

We have called for content regulation because, in many aspects of online content[25], to date, Facebook has had to effectively self-regulate and make decisions on where to draw the line between freedom of expression and harmful content. We strongly believe that designing the rules of the internet should not be left to private companies alone, and we recognise that new regulation can make a real positive difference. It is preferable for this regulation to be developed in

[23] Mark Zuckerberg, 'The Internet Needs New Rules', *Washington Post*, https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html

[24] Mark Zuckerberg, *Big tech needs more regulation*, https://www.ft.com/content/602ec7ec-4f18-11ea-95a0-43d18ec715f5

[25] Notwithstanding the small number of content regimes in some areas around the world, like the *Enhancing Online Safety Act*.

liberal democracies that uphold values of free expression and human rights, rather than countries that adopt a more authoritarian view of speech.

In the past, laws regulating expression have generally been implemented by law enforcement officials and the courts, but internet content moderation is fundamentally different. The white paper asks whether governments should create rules to address this complexity, recognising user preferences, differences in services, the need to enforce at scale, and flexibility across language, trends and context.

We believe there is now an opportunity for governments like the Australian Government to develop effective, thoughtful, world-leading content regulation that reflects democratic values and protects its citizens.

In order to assist, we have outlined a number of principles that we believe reflect an effective content regulatory scheme:

- It should incentivise best practice content moderation, rather than encourage a compliance mindset.
- It should be principles-based and able to accommodate changes in technology.
- Regulation should balance the need to effectively reduce harmful speech, while preserving free expression.
- Policymakers should develop frameworks that do not rely excessively on proxy metrics which may not properly account for the impact of harmful content.
- Regulatory frameworks should account for the difference between illegal content and harmful content.
- Regulation should recognise that enforcement is not perfect.
- Regulation should not expect companies (who are intermediaries) to assume the responsibilities of those who produce the content in the first place.

Given the breadth and nuance of the content that is potentially covered by regulation targeting harmful content, we encourage the Australian Government to undertake a deeper consultation process, similar to the UK Government with its *Online Harms White Paper* consultation.

Inclusion of private messaging within the online safety legislative scheme

The consultation paper's proposed inclusion of all private messaging within scope of safety laws raises concern about government regulation of individual conversations. The paper proposes extending the cyberbullying scheme to all private messaging, granting the eSafety Commissioner to power to police not just social media posts but also private conversations between Australian adults for potential bullying and harassment. The paper also explicitly identifies private messaging services that are end-to-end encrypted.

The eSafety Commissioner already has powers in relation to use of private messaging for the most harmful types of content (child exploitative content, and non-consensually shared intimate images), but the paper proposes extending these powers in relation to potential bullying and harassment.

We already take significant measures to prevent the sharing of the most harmful types of content in private messaging. For example, Facebook was one of the first to use artificial intelligence and machine learning to identify newly created child exploitation imagery and potentially inappropriate interactions between adults and minors. We use PhotoDNA to proactively scan for child exploitation imagery, including on unencrypted surfaces in WhatsApp such as user and group profile photos and user reports. If there is a match to known child sexual exploitation content, the image is blocked from being uploaded, reported to the National Center for Missing and Exploited Children, and the accounts in question are banned. This practice occurs across Facebook, Instagram, Messenger and WhatsApp.

While we are committed to providing a safe experience on messaging services, we have serious concerns about the inclusion of all private messaging in regulatory schemes for the following reasons:

- The existing cyberbullying scheme was developed to provide Australians with recourse when they experienced harassment online and they were unable to remove it themselves. There are many tools in messaging services that Australians can use to manage the messages they receive. They can delete any message they receive and block any person from contacting them. Within Messenger, people also have the option to Ignore any conversation, which moves those messages into a separate inbox, so they don't have to see it every time they open Messenger. People already have significant control over their experience when messaging.

- Government takedown schemes are a blunt instrument when applied to private messaging, and they will struggle to capture the context and complexity of human relationships. Because private conversations do not have a public or shaming component, decisions about whether content constitutes bullying or harassment will require finer judgement and a full understanding of the context of the relationship and offline context in which the conversation occurs. While takedown schemes for social media can help stem the further sharing and continued harm of a piece of bullying or harassment content, they are less suitable for bullying or harassment that occurs privately than laws or schemes that are directly targeted at stopping the perpetrator from continuing the behaviour.

- The extension of the cyberbullying scheme to end-to-end encrypted messaging services would not account for the technical challenges and features of encryption. The core principle behind end-to-end encryption is that only the sender and recipient of a message have the keys to "unlock" and read what is sent. No one can intercept and read these messages. To protect people who use WhatsApp, for example, we have protections in place to help keep people safe from unwanted contact and offer them the ability to block and report inappropriate behaviour. Those found violating our terms of service are removed from the platform. However, beyond these individual controls, WhatsApp would be otherwise unable to comply with a government order to remove a specific message.

The type and severity of harm experienced via bullying or harassment on private messaging services is different to social media services, primarily because users have greater control over the interaction. Also, the paper suggests the capacity for victims to effectively address cyberbullying conduct occurring within message groups – where there is typically a large number of recipients – is more limited. The protections of blocking and reporting in WhatsApp are the same in both one-to-one and group message settings.

We recognise people could be added to groups they did not wish to be part of so we recently updated our product to improve group settings and give users control over who can add them to a group - 'everyone', 'contacts only or 'selected contacts'. This is a significant product change that has been requested by users, policy makers, and privacy advocates to help prevent phone numbers from being exposed to unwanted groups and also to give people greater control over the content they see.

Time-to-action v prevalence

We share the view of the Australian Government and the community that harmful content should be removed from digital services as quickly as possible. The overwhelming majority of reports we receive from the eSafety Commissioner are resolved in under 30 minutes and - across all reports we have received from the eSafety Commissioner since 2015, including complicated cases that required further review - the median turnaround time is ~16 hours.

That said, establishing a requirement (backed by legal sanction) that all harmful content should be removed within 24 hours may have perverse unintended consequences. As we outline in our recent whitepaper *Chartering a Way Forward: Online Content Regulation*, *"companies focused on average speed of assessment would end up prioritising review of posts unlikely to violate or unlikely to reach many viewers, simply because those posts are closer to the 24-hour deadline, even while other posts are going viral and reaching millions."* And specifically, given the concern about private messaging, we note that imposing a 24-hour timeframe means *"*[c]*ompanies would have a strong incentive to turn a blind eye to content that is older than 24 hours* (*and unlikely to be seen by a government*)*, even though that content could be causing harm. Companies would be disincentivised from developing technology that can identify violating private content on the site, and from conducting prevalence studies of the existing content on their site."*

An alternative metric to measure the performance of a digital platform to action harmful content could be *prevalence*. Generally speaking, some kinds of content are harmful only to the extent they are ever actually seen by people. A regulator would likely care more about stopping one incendiary hateful post that would be seen by millions of people than twenty hateful posts that are seen only by the person who posted them. For this reason, regulators (and platforms) will likely be best served by a focus on reducing the prevalence of views of harmful content. Regulators might consider, however, the incentives they might be inadvertently creating for companies to more narrowly define harmful content or to take a minimalist approach in prevalence studies.

Specific questions raised in the consultation paper

Objects of the new Act

We **support in-principle** the proposal to develop an objects section for the new Act, accompanied by a statement of regulatory policy.

However, we have concerns about the lack of clarity around the concept of "online harms" throughout the paper, and that is also proposed to be included in the objects of the new Act.

There is little clarity provided in the consultation paper about the definition of online harms: when describing the intended scope of “online harms”, the consultation paper (on page 9) refers to “cyberbullying, abusive commentary or ‘trolling’, the non-consensual sharing of intimate images (image-based abuse), grooming for the purpose of child sexual abuse, cyberflashing, doxing and cyberstalking”, which are all examples that are covered under online safety and existing legislation, but in a different section states that “Online safety measures extend to mitigating user exposure to illegal or harmful content, such as extremely violent content, terrorist propaganda or child sexual abuse and exploitation material.” All of these examples are covered by existing legislation. Consequently, it is not clear what the additional content is that is intended to be captured by the general term of “online harms” or “harmful content”.

“Harmful content” is a very broad concept, and it can be highly contested (for example, some claim that climate change scepticism should be classified as harmful content). Different solutions will be appropriate for different types of content.

We believe it would be preferable to initiate a new, separate review process to consider a regulatory scheme for harmful content more broadly - rather than try to retrofit a safety legislation scheme onto all types of content.

Structure of the Act (including Basic Online Safety Expectations)

We **support** the concept of principles-based requirements outlining the expectations that the Australian Government has for online safety, however, we believe more work is required on the Basic Online Safety Expectations to provide clarity and certainty to industry about the standards they are expected to meet. The Basic Online Safety Expectations should constitute principles-based and technology-neutral regulation that sets sufficiently precise obligations for companies, while retaining flexibility to adapt to changes in technology.

The paper recommends adapting the existing Online Safety Charter and the eSafety Commissioner’s Safety by Design Principles to constitute the set of Basic Online Safety Expectations. Facebook is one of the first companies to have publicly supported the eSafety Commissioner’s Safety by Design Principles and already fulfils many of the expectations outlined in the Online Safety Charter. The underlying concept of safety-based principles (as reflected in the Basic Online Safety Expectations) is useful, and has some parallels with other regulation - like the Australian Privacy Principles in the Privacy Act.

Notwithstanding the potential benefit of safety principles, the purpose and practical effect of the Expectations is confusing. We raise four concerns and suggest a series of principles that may be more effective.

Firstly, it is not clear whether the Expectations are supposed to set minimum requirements (as the name would suggest) or highlight aspirational goals for online safety efforts. While the Online Safety Charter is described as a “benchmark for best practice”[26], the same requirements when incorporated into the Expectations are considered to be a basic, minimum obligation for all social media companies. Many of the proposals contained in the Expectations would set an aspirational bar, by collating all best practice safety tools and practices in the industry, regardless of whether they are all suitable for specific products or services. We do not agree that it is appropriate for legislation to articulate opinions on best practice: this role is better suited for non-regulatory initiatives, like the Charter and Safety by Design Principles that already exist.

It is also premature for the Expectations to incorporate the provisions of the voluntary transparency protocol being developed by the Organisation for Economic Cooperation and Development (OECD). The OECD work has not yet completed and it is too early to assess whether the contents of that protocol should be incorporated in legislation. Even once the OECD work has been completed, the voluntary transparency protocol has been developed with the purpose of being voluntary rather than contained in regulation.

Secondly, various aspects of the Expectations seek to prescribe features as detailed as technologies used to detect content, the structure of teams to develop safety policies, user reporting processes, internal risk management processes, promotion of online safety resources, and support inboxes for responding to user reports. Regulation may be more effective if it is technology-neutral, but prescribes the standards or policy outcomes that companies are obliged to work towards, allowing companies to respond in ways that best match their particular services.

Thirdly, any expansion of the eSafety Commissioner’s powers should apply across all aspects of the Online Safety Act, allowing for the Commissioner to apply the same scrutiny to all entities that should take responsibility for online safety, not just social media companies. At present, the powers required for the eSafety Commissioner to assess the efforts of digital platforms and investigate non-compliance with the Act are only attached to the Basic Online Safety Expectations (which cover only social media companies).

Fourthly, the proposed new provisions around transparency reporting do not provide sufficient clarity or certainty for companies. It seems to propose that the regulator seek any type of either public or direct reporting about any issue with respect to any company on which the Commissioner considers to be large or on which “online harms have been occurring”. This is very broad and vague, and does not provide certainty to industry. While we are supportive of transparency, it is important to set the right metrics for reporting.

[26] Page 22 of the consultation paper

We encourage the Government to establish functions and oversight for the eSafety Commissioner consistent with other regulators such as the Australian Privacy Commissioner. Requests for information (whether directly to the Commissioner or in publicly available reports) should follow robust, accountable processes.

To give effect to the intention of best practice, principles-based expectations, we propose a set of principles (below) that could represent more realistic minimum expectations for companies.

- Companies should seek to prevent known illegal material from being made available to users or accessible on their platforms and services, take appropriate action under their terms of service, and report to appropriate authorities.

- Companies should seek to identify new illegal content on their platforms and services, take appropriate action under their terms of service, and report to appropriate authorities. This principle should focus on the systems in place, rather than assuming the existence of any such content is a reflection of non-compliance on behalf of the company.

- Companies should regularly publish or share meaningful data on their efforts to combat illegal content. Although we do not believe it will be possible for all regulated entities to report data on all types of harmful content (and there is not consensus on the best metrics for reporting harmful content), the industry is moving in this direction and the principle should be drafted in a way to enable the Minister to compel this reporting in future.

- Companies should enable users to easily report content, and provide avenues for users to appeal content decisions.

- Companies should allow for external oversight or involvement for some aspects of their content moderation systems (such as their content policies or enforcement decisions). This principle requires further consultation across digital platforms, as different levels of external oversight may differ between services.

- Companies should have publicly-available policies to govern the content they will remove from their services that is not otherwise illegal. These policies could be articulated in terms of service, community standards or minimum standards.

- Companies should enhanced protections for minors. These requirements should take into account the best interests of the child, including balancing the child's right to protection, privacy and online participation.

Expanded cyberbullying scheme for children

We **support** the expansion of the cyberbullying scheme for children, subject to earlier comments relating to private messaging and takedown times. The existing scheme provides an avenue for Australians to raise concerns with problematic content that targets children.

Throughout the operation of this scheme, we have worked to ensure we are responsive and prompt in responding to any issues raised with us by the eSafety Commissioner's Office and other stakeholders in the child safety, mental health, and education sectors.

We support the extension of the same obligations to other service providers (including gaming) where there may be online safety considerations for children.

We do not support the proposed other tools for the eSafety Commissioner beyond content takedowns, including notices to apply account restrictions, to enforce terms of service or request other enforcement actions. The primary benefit from a cyberbullying scheme is to alert a social media platform to the existence of a piece of bullying and harassment-related content. In many instances, once becoming aware of the content, Facebook will take steps beyond simple removal of the content, in line with our terms of service. Social media services have terms of service because it is in their interests to have them and to enforce them and, for this reason, it is not necessary to allocate the proposed additional powers to the eSafety Commissioner.

New cyberbullying scheme for adults

We **support in-principle** the expansion of the cyberbullying scheme to adults, subject to earlier comments relating to private messaging and takedown times, and careful consideration given to the scope of this new scheme.

Bullying and harassment is already prohibited on Facebook and Instagram for both children and adults, as it violates our Community Standards. We're deeply committed to ensuring Facebook is a safe place for everyone. That's why we offer very easy ways for people to report content or accounts that make them feel uncomfortable. Each of those reports is reviewed by a team who review content 24/7, so we can quickly take action if that content violates our standards.

Beyond that, we've invested in a series of other tools and resources to help protect people from bullying and harassment on our platforms, including the ability to block, unfollow or unfriend people, Over 5 years ago, in coordination with the Yale Center for Emotional Intelligence we also built social resolution tools that provide people with the opportunity to let someone know when they have posted something that upsets them. The tools include lightweight language prompts and in the vast majority of cases when used the two people resolve the issues independently. More recently, we have developed offensive comment filters, and the ability to ignore unwanted messages in Messenger and restrict unwanted comments on posts on Instagram without the other person knowing, thereby avoiding escalating any conflict.

A regulated scheme provides an avenue for Australian adults to seek content to be taken down, in line with our Community Standards.

There are, however, difficulties that arise in relation to a regulated scheme for adults that do not arise in relation to children. The potential vulnerability of children means it is appropriate to err on the side of taking down content that may constitute bullying and harassment, but adults have much more complex and nuanced relationships. Context becomes much more important and judgements about meaning and impact are more subjective, which means that reasonable minds can differ about what should constitute bullying or harassment. There is also a risk that, in removing content that one person finds to be bullying and harassment, another person may consider that their legitimate self expression has been curtailed or censored.

We believe a new cyberbullying takedown scheme for Australian adults should have the following features:

- It should not be extended to private messaging, as outlined above.

- As suggested in the consultation paper, the threshold for determining whether content is bullying or harassment should be substantially harder to meet for adults than for children. The threshold proposed in the paper (content is menacing, harassing or offensive and causes serious distress or harm) still captures broad swathes of material that reflects normal relationships between adults.

Further limitations should be added, including replacing the term “offensive” with “grossly offensive” (in line with the New Zealand Harmful Digital Communications Act). We also recommend that public figures (such as politicians) should be excluded from the legislation, recognising that they can attract strident but legitimate criticism and creating an avenue to silence that criticism could have grave implications for free expression.

The Government has indicated its intention that this scheme should not be specifically targeted at hate speech, but may remove some hate speech incidentally. This is an imperfect solution that does not provide clarity about the types of content that are within scope (for example, differentiating between hate speech directed at an individual, and hate speech directed at a group of people).

- Just as social media platforms have appeals proposals for users who believe content has been wrongly taken down, a cyberbullying scheme for adults should have an appeals mechanism for impacted individuals to dispute decisions made by the eSafety Commissioner. Any cyberbullying scheme runs a high risk of incorrect decisions made, because of the complexity of adult relationships. Existing appeal avenues for government decision making (like the Administrative Appeals Tribunal) are not fast enough to give individuals recourse if they suffer from an incorrect decision.

- In line with comments provided about the cyberbullying scheme for children, we do not support the proposed other tools for the eSafety Commissioner beyond content takedowns, including notices to apply account restrictions, to enforce terms of service or request other enforcement actions.

Image-based abuse scheme

We **support** a complaints-based scheme to ensure the swift removal of non-consensually shared intimate images.

We do not allow the non-consensual sharing of intimate images at Facebook, nor do we allow threats to share those images without consent. The sharing of, or threat to share, intimate images online can have serious emotional, psychological and physical consequences for those depicted in the image. As a result, we respond seriously to violation of our policies in this area. Not only do we remove intimate images shared without permission from the people pictured and threats to share intimate images, in most cases, we will disable the account for sharing intimate images without permission.

Facebook has also been investing in technology and systems to proactively prevent the sharing of this type of content on our services.[27] We use photo matching technologies to help stop future attempts to share this content on Facebook, Messenger and Instagram. By using machine learning and artificial intelligence, we can now proactively detect near nude images or videos that are shared without permission on Facebook and Instagram. This means we can find this content before anyone reports it, which is important for two reasons: often victims are afraid of retribution, so they are reluctant to report the content themselves or are unaware the content has been shared.

Facebook has also taken a number of steps of our own to provide a more supportive experience for someone who is subject to non-consensually shared intimate images.

[27] Antigone Davis, *Detecting Non-Consensual Intimate Images and Supporting Victims,* https://about.fb.com/news/2019/03/detecting-non-consensual-intimate-images/

We built a proactive reporting tool in partnership with international safety organisations, survivors, and victim advocates to provide an emergency option for people to provide a photo proactively to Facebook, so it never gets shared on our platforms in the first place. And we have been re-evaluating our reporting tools and processes to ensure they are straightforward, clear and empathetic when a victim reports image-based abuse to us.

We look forward to continuing to work with the eSafety Commissioner, civil society groups, not-for-profits, academics and experts to continually improve our collective ability to respond to non-consensual intimate images.

New Online Content Scheme

We **support** the proposal to shift to Schedules 5 and 7 of the Broadcasting Services Act into the new Online Safety Act, for consistency and alignment of technologies.

We also agree that it is essential that there be a conversation within Australia about the regulation of harmful content on the internet. However, we are concerned with the ambit of the proposals in relation to the eSafety Commissioner's ability to address 'seriously harmful content'.

First, the proposed scope of seriously harmful content is unclear. To operate with any certainty, platforms and service providers will need a clear-cut definition of content to fall under the scheme. Consideration should be given to how this scheme would interact with other legislation or legal frameworks - such as existing anti-discrimination laws, abhorrent violent material, defamatory content and also, proposed religious freedom laws.

Given the broad nature of the scheme, allowing the Minister to capture additional types of content without Parliamentary process raises concerns. The implications for free expression could potentially be so significant that it should be considered by the Parliament.

Second, in relation to content hosted outside of Australia, the requirement that platforms be required to take down (rather than geo-restrict) content that is illegal but which does not violate companies' policies, goes beyond what is reasonably required to protect Australians. This proposal would apply Australian law extraterritorially in relation to content that may not be illegal in other jurisdictions, and that there is no global consensus on what constitutes seriously harmful content. The Australian Government would be unlikely to welcome similar extra-territorial claims to be made by other governments in determining what Australians can view online. The same outcome for Australians can be achieved by enabling geo-restrictions rather than takedowns.

We believe it would be preferable to initiate a new, separate review process to consider a regulatory scheme for harmful content more broadly - rather than try to retrofit a safety legislation scheme onto all types of content. “Harmful content” is an overly-broad concept, that is highly contested (for example, some claim that climate change scepticism should be classified as harmful content), and different solutions will be appropriate for different types of content.

Accreditation scheme

We **support in principle** the proposed accreditation scheme to enable informed purchase of family-friendly tools, subject to greater clarification about how this would operate. Accreditation schemes will only be effective if they are meaningful and provide relevant information concisely, at the point that consumers need it.

The consultation paper’s discussion about a proposed accreditation scheme references the age verification systems that companies use to ensure that children are not using their services in violation of their terms of service. Companies’ systems for verifying the age of their users vary significantly between services, which would make it impractical to include this consideration within the scope of an accreditation scheme.

Content blocking for ISPs

As these proposals are put forward in relation to internet service providers, rather than digital platforms, we have **no comment** on the proposed content blocking regime for internet service providers.

Ancillary service provider scheme

We would be grateful for further clarity around what the ancillary service provider scheme is intended to encompass.

In relation to digital distribution platforms (such as app stores), we agree that there should be means for the eSafety Commissioner to identify and alert providers to apps or games which are systemically facilitating abuse. We would appreciate further clarity on the specific thresholds for the ancillary service provider scheme and would be grateful for further engagement with the eSafety Commissioner on the proposed ancillary service provider scheme.

Governance of the eSafety Commissioner

We **support** the restructuring of the eSafety Commissioner (independent of the Australian Communications and Media Authority), provided that the functions and oversight are consistent with other comparable regulators such as the Australian Privacy Commissioner.

We also strongly recommend extending the remit of the Commissioner in relation to offline bullying and harassment. Bullying and harassment are complex phenomena, and their online manifestation is often only part of a broader bullying and harassment that also occurs offline.

Focussing only on the online component will limit the effectiveness of the Commissioner. For this reason, we strongly support extending the eSafety Commissioner's remit to include bullying and harassment more generally, to enable the Commissioner to undertake proactive steps to change the culture of bullying and harassment among children (including undertaking education programs, and consultation with students, parents and teachers).

There needs to be additional work to establish an evidence base about the levels of bullying and harassment, and non-consensually shared intimate images, in Australia. Success should not be measured by the number of content referrals, but a reduction in the overall level of bullying and harassment (offline and online) and in the prevalence of harmful content online.