Response to the Online Safety Act Exposure Draft

FEBRUARY 2021

FACEBOOK

Executive summary

Facebook supports the enhancement of online safety laws in Australia, as part of our ongoing commitment to ensuring the safety of Australians online. We support new regulatory frameworks for online content that ensure companies are making decisions about online speech in a way that minimises harm but also respects the fundamental right to free expression. Our previous submission to the Government outlined in detail our investments to protect the safety of Australians, particularly young Australians, on our services.[1]

There is a shared commitment amongst all stakeholders to online safety, and especially the safety of young people online. The shared commitment on online safety sits within a broader debate, currently underway in Australia and around the world, about what content should be allowed online, and how much discretion should be afforded to companies or governments to make those decisions.

We support the legislation as an opportunity to enhance the online safety of Australians and to establish a regime that holds companies to account for the commitments they make. There are many elements of the exposure draft legislation that we commend, in particular, the use of voluntary industry codes to set obligations for companies in addressing emerging policy concerns, rather than setting prescriptive requirements in law.

In recognition of concern that private companies are often making decisions about important content issues, we have not waited for legislation but regularly provide ongoing transparency through our Community Standards Enforcement Reports. These enable us to be held to account for our efforts to address harmful content on our service: for example, in the fourth quarter of 2020, we removed 6.3 million pieces of bullying and harassment content, up from 3.5 million pieces of content in the previous quarter. This was driven, in part, by increasing our automation abilities and improving our technology to detect and remove more English language comments, which helped our proactive rate increase from 26.4 per cent to 48.8 per cent.

To ensure greater accountability for our content governance, we have also taken proactive, voluntary steps to establish an Oversight Board to making binding rulings on difficult and significant decisions about content on Facebook and Instagram. Together, new legislation and industry governance initiatives like these will improve accountability for content on digital platforms.

[1] Attached.

Whether content removal decisions are taken by government or industry, however, content regulatory frameworks should set clear and reasonable definitions about otherwise legal content that warrants removal, and also require transparency and accountability.[2]

While we support the draft online safety legislation's intent, there are three areas where the scheme seems to go further than the stated policy intent of the legislation: expansion of cyberbullying takedown schemes to private messaging; the threshold proposed for the new adult cyberabuse scheme; and the need for greater transparency and accountability to ensure regulatory decision-making reflects the expectations of the community.

Firstly, the extension of cyberbullying takedown schemes to private communications (like messaging or email) seems to be a disproportionate response to bullying and harassment, given the existing protections and tools already available.

The eSafety Commissioner and law enforcement already have powers around the worst risks to online safety that can arises in private messages - like non-consensually shared intimate images and child sexual abuse material. And most private messaging services (including those provided by Facebook) provide tools and features to give users control over their safety in private messaging, like the ability to delete unwanted messages and block unwanted contacts.

Despite the fact that existing laws allow the most serious abuses of private messaging to be addressed, the draft legislation extends regulatory oversight to private conversations between Australians. Whilst no form of bullying and harassment should be tolerated, we do not believe this type of scheme is suitable for private messaging.

Human relationships can be very complex. Private messaging could involve interactions that are highly nuanced, context-dependent and could be misinterpreted as bullying, like a group of friends sharing an in-joke, or an argument between adults currently or formerly in a romantic relationship. It does not seem clear that government regulation of these types of conversations are warranted, given there are already measures to protect against when these conversations become abusive.

Moreover, the policy rationale of the Australian Government's cyberbullying scheme for social media does not apply in the same way to private messaging. Bullying over private messaging cannot go viral in the same way as a piece of bullying content on a

[2] M Bickert, *Charting a way forward: online content regulation*, white paper released February 2020, https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.

public social media platform; and regulators will rarely have the full context to determine whether a private conversation genuinely constitutes bullying. If people receive a hurtful message, they can easily delete it, block the sender and/or report it for action by the platform. These tools are designed to give people control over a very direct communications channel, given they will be best placed to have the context and to reduce the harm of a bullying message themselves.

Secondly, we believe the threshold set for an adult cyberbullying scheme could potentially lead to broader ramifications than anticipated and extend the eSafety Commissioner's regulatory powers to legitimate political speech and debate.

The draft legislation allows the eSafety Commissioner to order takedown by a digital platform simply if the content is offensive and was likely intended to cause serious harm. There are several reasons to be concerned about 'offensive' being a key threshold for the adult cyberbullying scheme. Firstly, it is a low threshold and has previously attracted controversy in Australia in other contexts, such as in relation to section 18C of the Racial Discrimination Act.

Contrary to the Government's stated intention in the consultation paper on the new online safety legislation, the use of the word 'offensive' sets a *lower* threshold for the adult cyberbullying scheme than for the existing scheme relating to children. In addition, the threshold of 'offensive' could create a challenging interaction between online safety legislation and existing laws, such as defamation, which have developed over time and which contain appropriate defences in order to balance between freedom of expression and addressing harms. In order to not unduly limit speech, comparable defences should be added to the online safety legislation to ensure an adequate balance, similar to other laws. The current threshold of 'offensive' fails to strike that balance. The risk is that an eSafety Commissioner could order the removal of 'offensive' content with public interest value (such as, posts from whistleblowers that contain allegations similar to stories considered by the Royal Commission into Institutional Responses to Child Abuse)

There is also the very real risk that this low threshold, as it has in other legislation, would capture political speech: the heat of political debate may result in legitimate political comments that could be considered offensive. Because the speech of political officials is within scope of the legislation, a regulator will have the discretion to potentially police what politicians say to each other.

Given the highly ambiguous thresholds for regulatory action under the proposed legislation such as 'serious harm' and 'offensive', there is significant likelihood that the adult cyberbullying scheme becomes de facto the regulation for all speech online.

To address this concern, a possibly more judicious threshold for the adult cyberbullying scheme would be replacing 'offensive' with 'grossly offensive' (in line with the New Zealand Harmful Digital Communications Act), which would more than adequately capture harmful bullying of adults without the same risks of overreach. Alternatively, the legislation could expressly require the eSafety Commissioner to consider the impact of exercising their discretion on the broader public interest and free speech.

Thirdly, the legislation grants a single regulator a considerable level of discretion and power over speech online. Clearer guidelines and greater checks and balances could assist in making sure this discretion is applied in ways that are consistent with the community's expectations. The nature of a complaints-based regulatory scheme means that any regulator only has sufficient information that will allow them to decide in favour of the person making the complaint: the eSafety Commissioner only considers the complainant's perspective of the possible harm resulting from a post and is not required to consider the relevant information from the other person engaged in the activity (when often this information will challenge or recontextualise the original complaint).

These comments are not intended to criticise the judgement or discretion of the regulator: online content decisions can be incredibly challenging, difficult to balance competing interests, and potentially significant to the individuals involved (or even society more broadly). Just as there should be transparency over the approach taken by digital platforms, there should also be transparency for decisions taken by regulators.

To that end, we encourage the Government to consider greater transparency and accountability measures for the relevant regulator, especially given the potentially significant impact of the systems enforced or the content decisions taken by the eSafety Commissioner. A specific measure could include the eSafety Commissioner reporting to Parliament with quantifiable metrics about reports and takedown notices. We also encourage the Government to consider stronger legal defences to protect legitimate political speech and requirements for the Commissioner to notify and reasonably consult with a digital platform prior to issuing a service provider notification.

We note that the draft legislation is near-identical to the Government's original consultation paper, and so, in an effort to be helpful, we have also attached Facebook's previous submission, which made a number of constructive suggestions in case they remain useful in the Government's consideration of the next draft of this legislation.

Facebook initiatives since the last submission

Facebook undertakes significant proactive work to protect the online safety of Australians who use our services. As we outlined in our previous submission to the online safety consultation process (attached), we have an industry-leading approach to safety that comprises policies, enforcement, tools and products, resources and partnerships.

There are two important updates on our proactive work that have occurred since our last submission.

Firstly, we have continued to improve in our efforts to combat harmful content on our platform, including the use of artificial intelligence to proactively detect and remove content. Our Community Standards Enforcement Report reports transparently on key metrics relating to harmful content, and is released quarterly.

According to the last Community Standards report (February 2021)[3], in the period October to December 2020, on Facebook, we took action against:

- 6.3 million pieces of content for bullying and harassment, up from 3.5 million in the previous quarter
- 26.9 million pieces of content for hate speech, up from 22.1 million in the previous quarter.

Both of these areas have traditionally been difficult to detect proactively via artificial intelligence, given they are so context-dependent. However, our proactive detection ability has continued to improve, with:

- 48.8 per cent of bullying and harassment content removed proactively via artificial intelligence. This is an increase from 26.4 per cent in the previous quarter and 16.1 per cent one year prior.
- 97.1 per cent of hate speech content removed proactively via artificial intelligence. This is an increase from 94.7 per cent in the previous quarter and 80.9 per cent one year prior.

Improvements to our artificial intelligence in areas where nuance and context are essential, such as hate speech or bullying and harassment, helped us better scale our efforts to keep people safe.

[3] G Rosen, 'Community Standards Enforcement Report - February 2021', *Facebook Newsroom*, 11 February 2021, https://about.fb.com/news/2021/02/community-standards-enforcement-report-q4-2020/

Secondly, there is a significant development since our last submission, namely the establishment of the Oversight Board. The members of the Oversight Board were appointed in May 2020[4] and they began taking cases in October 2020.[5]

The Oversight Board was borne out of the recognition that critical decisions about content should not be left to companies alone. Content decisions can have significant consequences for free expression and companies like Facebook - notwithstanding our significant investments in detection, enforcement and careful policy development - will not always get it right.

The Oversight Board comprises 20 experts in human rights and technology - including the Australian academic Professor Nic Suzor - and will increase over time to 40 members. The Board is entirely independent and hears appeals on Facebook's decisions relating to content on Facebook and Instagram (beginning with decisions where content was removed). We have agreed that the Board's decisions will be binding, and the Board is also able to make recommendations about Facebook's policies.[6]

The Oversight Board has begun issuing its decisions from January 2021.[7] Facebook has also announced that we have referred our decision to indefinitely suspend Former President Trump's Facebook and Instagram accounts to the Oversight Board, given it is a significant and difficult decision.[8]

We believe the Oversight Board is a significant innovation in content governance and a first-of-its-kind initiative. It will make Facebook more accountable for our content decisions and will help to improve our decision-making. When combined with other proactive industry initiatives (like Facebook's Community Standards Enforcement Report) and new legislation, it will provide greater confidence and accountability about the decisions underpinning content on our platform.

---

[4] N Clegg, 'Welcoming the Oversight Board', *Facebook Newsroom*, 6 May 2020, https://about.fb.com/news/2020/05/welcoming-the-oversight-board/

[5] B Harris, 'Oversight Board Selects First Cases to Review', *Facebook Newsroom*, 1 December 2020, https://about.fb.com/news/2020/12/oversight-board-selects-first-cases-to-review/

[6] B Harris, 'Establishing structure and governance for an independent oversight board', *Facebook Newsroom*, 17 September 2019, https://about.fb.com/news/2019/09/oversight-board-structure/

[7] M Bickert, 'Responding to the Oversight Board's First Decisions', *Facebook Newsroom*, 28 January 2021, https://about.fb.com/news/2021/01/responding-to-the-oversight-boards-first-decisions/

[8] N Clegg, 'Referring Former President Trump's suspension from Facebook to the Oversight Board', *Facebook Newsroom*, 21 January 2021, https://about.fb.com/news/2021/01/referring-trump-suspension-to-oversight-board/

Concerns with the legislation

There are three primary concerns Facebook has with the exposure draft legislation, relating to:

1. The inclusion of private messaging
2. The threshold for the adult cyberbullying scheme
3. Greater transparency and accountability for the regulator, given the increased powers of the eSafety Commissioner.

Private messaging

The draft legislation proposes extending the cyberbullying scheme to all private communication services (including private messaging and email). This grants the eSafety Commissioner the power to police not just social media posts but also private conversations between Australian users for potential bullying and harassment, and sets an expectation that digital platforms will be monitoring and policing private conversations.

There is no doubt that safety concerns can arise in relation to private messaging. However, it is overly simplistic to assume the existing cyberbullying takedown scheme should be applied wholesale to private messaging, for a number of reasons:

1. **Australian law already protects against serious harm in private messaging**. The eSafety Commissioner already has powers in relation to use of private messaging for the most serious harmful types of content (child exploitative content, and non-consensually shared intimate images).

   In relation to more complex and context-specific harms like bullying, people already have significant control over their experience when messaging. 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. Similar messaging controls exist on Instagram Direct.

2. **The nature of potential harm is different on private messaging to public forums like social media.** The type and severity of harm experienced via bullying or harassment on private messaging services is different from social media services, primarily because users have greater control over the interaction. This remains true for group conversations over private messaging

as well - the same tools are available on WhatsApp and Messenger for example for group conversations as for private direct messages.

3. **Takedown schemes are ill-suited for private messaging.** The existing cyberbullying scheme was developed to provide Australians with recourse when they experienced harassment online and they were unable to remove it themselves. On private messaging services however, Australians have a much greater level of control over their messages.

   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.

4. **There are technical limitations in the ability to monitor and police private messaging.** 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, end-to-end encrypted services would be otherwise unable to comply with a government order to remove a specific message.

While we accept the need to ensure the safety of users in private messaging, we do not believe a takedown scheme designed for social media is suitable nor should be applied wholesale to private communications.

Threshold for the adult cyber-bullying scheme

Facebook supports the creation of an adult cyber-bullying scheme in Australia. However, as the Government recognised in the consultation paper for this legislation, the scheme available to adults should set a higher threshold than the scheme for

children, in recognition "that adults can be expected to demonstrate a higher level of resilience and maturity than children".[9]

We highlighted in our previous submission that an adult-focussed regulatory scheme incurs greater risks than a scheme only available 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.

However, the exposure draft legislation does not seem to give effect to the Government's stated intention: it sets a *lower* threshold for adults than for children. It allows the eSafety Commissioner to seek the removal of content that is "menacing, harassing or offensive" and was likely intended to cause serious harm. The experience of other laws demonstrates that, 'offensive' is a low threshold -- for example, it has previously attracted controversy in Australia in other contexts, such as in relation to section 18C of the Racial Discrimination Act.

The design of this definition creates a challenging crossover between online safety legislation and existing laws, such as defamation, which have developed over time and which contain appropriate elements and defences in order to balance between freedom of expression and the protection of reputation. We do not believe it is appropriate for an adult cyberbullying scheme to cut across multiple pieces of well-established and tested legislation by creating a new, ambiguous standard that results in greater uncertainty for companies about content that may be within scope.

There is also no doubt that this low threshold would capture political speech: the heat of political debate may result in legitimate political comments that could be considered offensive. Furthermore, in determining what is offensive, there is no requirement to consider the public interest value of the material. Because the speech of political officials is within scope of the legislation, the eSafety Commissioner will have the power to police what politicians say to each other. It is not clear that this meets the policy intent of the draft legislation and perhaps greater clarification and safeguards are needed here.

[9] Department of Communications and the Arts, *Online safety legislative reform discussion paper*, December 2019, https://www.communications.gov.au/have-your-say/consultation-online-safety-reforms.

A potentially more judicious threshold for the adult cyberbullying scheme, could be replacing 'offensive' with 'grossly offensive' (in line with the New Zealand Harmful Digital Communications Act) would more than adequately capture harmful bullying of adults without the same risks of overreach. Alternatively, the legislation could expressly require the eSafety Commissioner to consider the impact of exercising their discretion on the broader public interest and free speech.

Greater regulatory transparency and accountability

Overall, the exposure draft legislation represents a significant increase in the breadth and scope of the eSafety Commissioner's powers. Many of the provisions of the legislation (for example, the basic online safety expectations or the definition of 'serious harm' in the adult cyberbullying scheme) are unclear, and rely on the discretion of the eSafety Commissioner.

To ensure that these powers are exercised in ways that are proportionate and balanced, the legislation requires greater regulatory transparency and accountability. We recommend the following checks and balances:

- Transparency and accountability of regulatory content decisions
- Simple appeal mechanisms for individuals, to correct regulatory errors
- Establishing clear legal defences
- Minimising the regulatory burden and ensuring best practice regulatory performance
- Statutory review of the scheme.

These are outlined in more detail below.

Transparency and accountability

Facebook agrees that there should be greater accountability and transparency about digital platforms' decision making in relation to content on their services. Particular instances of content can be difficult and significant, and we support frameworks (both legislative and industry-led, like the Oversight Board or our internal appeals processes) to provide greater confidence about content decisions. This is also why we voluntarily report quarterly, with consistent metrics, on our enforcement of Facebook's Community Standards.

We support the development of legislative frameworks to oversee content decisions by platforms, but those frameworks should be accompanied by comparable levels of transparency and accountability for the regulators making those content decisions in place of platforms.

For example, the eSafety Commissioner's Office already provides an annual report (like all government agencies), but -- to increase transparency -- this could contain information such as the number of complaints received, number of referrals sent to digital platforms, number of notices issued, or number of end-user notices issued.

Simple appeal mechanisms

There should be simple appeal mechanisms to allow individuals to raise appeals directly with the eSafety Commissioner's Office for adult cyberabuse. These should allow the respondent to express their point of view in response to a complaint when the Office makes a mistake or they disagree with the decision. The very nature of a complaints-based regulatory scheme means that the regulator will inevitably lean in favour of the person making the complaint: the eSafety Commissioner only considers the complainant's perspective of the possible harm resulting from a post and does not have the information nor imperative to consider the other side of the story. This may be particularly important when political speech is in question.

While there is broad and concrete agreement about the need to protect online safety and remove harmful content, there can be great contention about what constitutes harmful content when looking at the specifics of a particular piece of content. Human relationships can be very complex.

Even a considered and well-intentioned regulator will inevitably make mistakes; just as digital platforms' content enforcement is imperfect.

If an individual has their content or account taken down on the instruction of the eSafety Commissioner, they should have a simple way to raise this concern directly with the Office, without needing to resort to the Administrative Appeals Tribunal (AAT). AAT processes are long, complex, costly and, by their very nature, will not be an effective option for recourse for disadvantaged Australians.

Clear legal defences

Currently, the eSafety Commissioner has broad discretion over how to direct standards, issue take down notices and enforce the basic online safety expectations. To ensure the regulator's decision making approach is aligned with the expectations of the community, we recommend the Government consider whether it would be useful to encourage the Commissioner to have any regard to any countervailing factors, or recognition of any legitimate difficulties in complying with an eSafety Commissioner direction or regulation.

Some clear legal defences that the Government may wish to consider are provided below:

- **A "reasonable steps" defence.** There will be some instances where it is simply not possible to find the content or user that the eSafety Commissioner's Office is concerned about: we have experienced situations in which referrals to us are unclear, and we have not been provided with enough information to locate the content or user in question. We will always take additional reasonable steps to go beyond and attempt to find the material being referred by the eSafety Commissioner's Office but - under the current draft legislation - a company could be found non-compliant even if the information provided with the complaint is unclear.

  (We note that the adult cyberbullying scheme contains a "reasonable steps" test and requirement for the eSafety Commissioner to provide unique identifying information, but these requirements do not extend to the child cyberbullying scheme.)

- **A public interest defence.** As outlined earlier, given the low threshold set for the adult cyberbullying scheme, we anticipate that there would be 'offensive' material that could arguably result in serious harm to an individual that would still hold significant public interest. This is particularly the case in relation to content by whistleblowers, such as those who were part of the Royal Commission into Institutional Responses to Child Abuse. The eSafety Commissioner should be directed to consider public interest factors as they operate their wide-ranging discretion.

There could also be some consideration for instances where there is a conflict of laws (for example, where a regulated entity is unable to comply with an eSafety Commissioner's direction because of a conflict of laws where a regulated entity would be forced to choose between violating the online safety legislation to ensure compliance with international laws.

Minimising the regulatory burden

The eSafety Commissioner is granted a significant expansion in their powers over digital platforms as a regulator under this legislation. Some examples include:

- Under the basic online safety expectations, the eSafety Commissioner can issue a 'service provider notification' about any concerns with breaches of the expectations. The legislation allows for these notifications and any associated reports to be made publicly available, even before a company is notified. It

would be good regulatory practice to at least notify and reasonably consult with a digital platform, prior to publicly claiming they are not meeting basic expectations.

- While we welcome the use of voluntary industry codes to deal with emerging policy issues rather than legislation, the exposure draft allows the Commissioner to make a service provider determination (ie. set a new regulatory obligation on a single company or parts of industry), and does not require any notice or consultation prior to this determination being made.

- The legislation now provides for quasi-law enforcement style investigatory powers, which attract significant penalties for non-compliance.

Collectively, these changes could incur a marked increase in the regulatory burden faced by digital platforms. The regulatory burden of these requirements should be regularly reviewed.

In addition to some minor legislative changes, we suggest that it may be useful to include some additional governance structures in place to make sure the eSafety Commissioner's Office acts in accordance with best practice regulatory principles. As per the Government's regulatory performance framework, the Office should consult with regulated entities to gather feedback on its performance.[10]

Statutory review

We recommend that the legislation contain a statutory review of the operation of the legislation, to begin two years after the legislation's passage. This would be an entirely normal provision, consistent with other pieces of legislation of this level of significance.

[10] Department of the Prime Minister & Cabinet, *Regulatory Performance Framework,* https://pmc.gov.au/sites/default/files/publications/Regulator_Performance_Framework.pdf