Online Hate Prevention Institute

Online Hate Prevention Institute submission to the Statutory Review of the Online Safety Act 2021

July 2024

Introduction

This submission to the statutory review into the operation of the Online Safety Act 2021 by Ms Delia Rickard PSM is from the Online Hate Prevention Institute, a harm prevention charity established in January 2012 and focused on preventing and minimising the harm to people that results from online hate and extremism. We cover all forms of online hate and extremism against individuals and groups in Australia and we contribute significantly to international online safety efforts and public policy discussion in this area. Our expertise is often sought by foreign governments and intergovernmental organisations as well as UN bodies and our reports on the topic are widely cited. This submission has been prepared by our CEO, Dr Andre Oboler.

Dr Andre Oboler is an internationally recognised expert who has been working on tackling online hate since 2004, initially in relation to search engines and later social media. From 2008 he served for nine years as a co-chair of the Global Forum for Combating Antisemitism, convened by the Israeli government, where he has responsibility for co-leading the working group on online antisemitism. Since 2015 he has served as an expert member of the Australian Government's delegation to the International Holocaust Remembrance Alliance, where he served on the Committee on Antisemitism and Holocaust Denial. He served on the advisory group for the Swedish Government's Malmo Forum in 2020. He has assisted the governments of Greece and The Netherlands with conferences looking at online hate, and served as an invited panellist for the UN Forum on Minority Issues when it focused on online hate in 2020.[1] Dr Oboler also serves as Chair of the Global Public Policy Committee of the IEEE,[2] the international professional body representing over 460,000 professionals in the technology and engineering fields across 190 countries.[3] He is also general chair of the IEEE Conference on Digital Platforms and Societal Harms.[4] In addition to a strong focus on antisemitism, Dr Oboler through the Online Hate Prevention Institute carried out the first research on racist Aboriginal Memes in 2012,[5] the world's first report on online Islamophobia in 2013,[6] and is a co-author of the book *Cyber-Racism and Community Resilience* (2017),[7] the result of a major ARC funded research project into online hate with Australia's leading experts in the field. He is the author of book chapters and peer reviewed research papers on the topic of online hate, and of particular note for this review is his paper on *Legal Doctrines Applied to Online Hate Speech.*[8] Dr Oboler holds a PhD in Computer Science from Lancaster University (UK), and an LLM(Juris Doctor) and Honours degree in Computer Science from Monash University.

We open by noting that the Online Hate Prevention Institute pre-dates the establishment of the office of the eSafety Commissioner and we were one of the first to call for the establishment of such an office during an interview on online racism with SBS in August

---

[1] https://ohpi.org.au/address-to-the-united-nations/
[2] https://globalpolicy.ieee.org/about/
[3] https://www.ieee.org/about/at-a-glance.html
[4] https://tech-forum.computer.org/societal-harms-2024/
[5] https://nla.gov.au/nla.obj-888711478/view
[6] https://nla.gov.au/nla.obj-1971792213/view
[7] https://link.springer.com/book/10.1007/978-3-319-64388-5
[8] Andre Oboler, "Legal Doctrines Applied to Online Hate Speech", *Computers & Law*, Number 87, pp 9—15, July. 2014. https://www.austlii.edu.au/au/journals/ANZCompuLawJl/2014/4.pdf

2012.[9] We were also part of the initial civil society group consulted by the then Liberal Opposition when developing their policy to establish an online Safety Office, and from the start we have advocated for an online safety policy that covers cyber-racism and other forms of hate targeting entire groups within society.[10] Even in the limited context of children's online safety, a policy tackling group based hate (targeting the group in general rather than specific individuals) was needed, and unfortunately absent.

Recommendation 1: The remit of eSafety should explicitly include the safety of individuals and groups within society that are impacted by online hate.

We have also advocated from the start for online safety to protect more than just children. We have made submissions to many reviews and inquiries into online safety by the relevant department and the parliament over the years and repeatedly advocated for such an expansion.[11] We have welcomed the various expansion of the scope of the eSafety Commissioner as they have occurred. In our most recent submissions,[12] we advocated for the notice power to secure removal of content to be expanded to make it a general power to issue a notice on referral from a competent authority for any content that breaches state, territory, or Commonwealth laws administered by that authority. This would streamline the removal process, empowering more government agencies to take action in relation to online breaches of law, and reducing the burden on platforms of dealing with multiple authorities. In this submission we also take a queue from Europe's *Digital Safety Act* noting the ability of civil society organisations (CSOs) to contribute to online safety, and the value in government supporting and empowering this role.

Recommendation 2: The power to issue notices and takedown orders should be extended beyond the current schemes into a general power that allow eSafety to act, on referral from a relevant government authority, in response to any content likely to be unlawful under Commonwealth, state, or territory laws. The existing scheme provides sufficient protections to allow such notices to be challenged through the courts where a company feels content is likely to be legal.

While welcoming the expansion of eSafety, online hate against groups (beyond cyberbullying of individuals) continues to be excluded, despite its growth into a major pandemic which has been addressed by governments around the world, as well as by a major new push from the

---

[9] SBS World News Australia, first broadcast 17 AUG 2012, 6:36 PM. Available at: https://www.youtube.com/watch?v=cYTMkWu30ow

[10] 2012 Submission to the Coalitions Review of Online Safety for Children, Liberal Party of Australia and the National Party of Australia (available on request). Extract: "While cyberbullying is recognized as a significant online threat to children, it has a fairly narrow definition and other forms of hate are excluded, for example online racism targeting communities at large rather than specific individuals… Online racism falls into a gap between those policies that address racism (but not online) and those policies that address online bully, but not hate targeting victim groups. Other forms of group hate fall into a similar gap, include: homophobia, hate targeting those with disabilities and hate targeting groups based on their religion or nationality."

[11] See past submissions at: https://ohpi.org.au/public-policy-law-reform/

[12] Andre Oboler, Mark Civitella, David Wishart and Simon Katterl (2021). *Online Hate Prevention Institute Submission to the Inquiry into the Online Safety Bill 2021*, Senate Standing Committee on Environment and Communications, Australian Parliament.

United Nations as recently as June 2024.[13] We have also found that resourcing has not kept pace with the increased scope of eSafety, nor has the culture shifted, meaning areas outside of Children's online safety often appear to be an added burden interfering with the core work (around children's safety), rather than an equally important part of the role. This applies even to work countering terrorism. The level of engagement with civil society organisations focused on children's safety, and to a degree women's safety, is far stronger than the engagement with an organisation like ourselves that focuses on online safety related to hate and extremism. There needs to be a wider consultation with CSOs for eSafety to properly fulfil its broader objectives and to improve collaboration and support for civil society organisations operating in the broader online safety space.

Recommendation 3: eSafety needs to have a wider and more regular engagement with civil society organisations working across different issues related to online safety.

While we have met with a range of staff at eSafety and shared information with them for working in different areas, given our role as a civil society organisation dedicated to online safety, a closer working relationship and stronger links including funding and support for the work we do might be expected. Such views have been expressed by a past Minister responsible for eSafety, only for it to be pointed out that the legislation itself prevented this as the power to make grants was limited to the initial children's safety remit. While the legal limitations have, we believe, been removed, limitations remain within the culture. Our critical work on online safety, such as on anti-Asian hate during Covid, and racism during the referendum, have instead been supported by grants from industry and have been dependent on good will from industry.

Recommendation 4: eSafety grants need to cover a wider scope of online safety, not just the work related to the dedicated schemes eSafety manages.

While Australia's Online Safety Act was world leading at the time it was created, and many changes to the Act have significantly enhanced it, other countries (and the European Union) have not only caught up, but in many ways now surpassed us. This review is timely and further reform is needed, but it should build upon the existing legislation, and focus on closing the long term gaps. Some restructuring (or rather, addition of structure) may be needed along with additional staffing and funding to facilitate a cultural shift to have a part of the Office with a dedicated focus outside of children's online safety (or other specific target areas like image based abuse) and the capacity to engage more deeply with a broader cross section of civil society and with other government agencies.

Recommendation 5: eSafety should be restructured to add one or more Deputy Commissioners, and related support staff, who can maintain a focus on different areas of online safety with at least one focus on general online safety beyond the current specific schemes.

---

[13] Edith N. Lederer, "UN launches global principles to combat online hate and demands big tech take action now", AP, 25 June 2024. https://apnews.com/article/un-combat-online-hate-lies-information-integrity-9a3b13fd27fb46574d45750dadcdcd3e

Australia's regulatory approach to online services, systems and processes

1. Are the current objects of the Act to improve and promote online safety for Australians sufficient or should they be expanded?

The issue paper opens in Part 2 by saying "The objects of the Act are to improve and promote the safety of Australians online". While it is the primary objective, as the paper also noted in Part 1, "Australia is one of many countries regulating online safety in a global regulatory environment that is not confined to national borders". As an advocate for the international rules based order, an additional objective of the act should be to facilitate Australia's contribution to reducing global harms. Practically, this might include:

- Taking action in Australia where content published from Australia is designed to cause serious harm overseas (such that if the harm were to be focused on people in Australia it would be a breach).
    - We have on various occasions been made aware of Australia social media users and website hosts who, operating from Australia target people overseas in a manner that, had those people been in Australia, would have seen them protected. At the same time regulators / law enforcement overseas may wash their hands of the matter on learning the perpetrator is outside their jurisdiction.
- Taking action globally where Australia has the strongest association with harmful content. Abhorrent Violent Content produced in Australia should be globally banned by Australia, not just blocked within Australia. This improves global safety and Australia has the responsibility to do this when the content originates in Australia.
- Representing Australia internationally in groups and forums related to online safety might be explicitly incorporated. This might require some consideration of the relationship between eSafety and DFAT.

Recommendation 6: An additional objective of the Online Safety Act should be to fulfil Australia's international human rights obligations, particularly in addressing Australian based or generated content that is causing or contributing to harms overseas.

2. Does the Act capture and define the right sections of the online industry?

Online services might be better defined by their functionality rather than their advertised purpose. The expectations on a read only service (like an informational website) will be lower than those on a system that allows users to upload and share content (without pre-moderation), which may be lower than those on a system that allows networking and the type of social interactions and sharing that enabled content to "go viral". Specific expectations may apply to systems that allow live streaming or search. Whether the service is presented as a gaming platform, a social media platform, an entertainment platform, a dating application, a file sharing system, is less relevant than the functionality the system provides and how that functionality (or combination of functionalities) might cause harm.

This also means systems with a variety of functions might be subject to different regulations for each function. Imagine a social media platform that integrates its own search engine, or a metaverse application with a search function and a messaging function, possibly with advertising and suggested content. There is a convergence of technologies and an interplay between different functionality. Regulating features more gradually would provide a better approach, and may provide a lighter-touch approach for platforms with a more limited set of features. It would also allow platforms to plan new features with an eye to the safety requirements needed for that type of content / interaction. We note that generative AI agents may in some sense already be replacing functions such as search without being a search engine under the act.

In 2013 the Online Antisemitism Working Group of the Global Forum for Combating Antisemitism (which I had the honour to co-chair) produced the TEMPIS taxonomy for classifying online content.[14] This is only a subset of online services, but the working group, far ahead of its time in discussing online regulation, suggested different expectations be set for different types of online content rather than for different platforms. The table below outlines the factors the Working Group suggested be considered when setting an expectation. The same pattern of factors could occur for multiple styles of content on multiple platforms, setting a common expectation for them. New platforms would know the expectations for different types of content before they launch features using such content (e.g. live streaming) and could design to meet them the expected standard, or delay release until the capacity for compliance was available.

Timing	Empowerment	Moderation	Publicness	Identity	Social Impact
Real Time	None	Pre-moderated	Personal	Verified real ID	No sharing
Stored	Power	Post Moderated	Private Group	Unverified Consistent ID	Sharing with associates
-	Responsibility	Exception Moderated	Associates	Anonymous possible	Public Sharing
-	-	Complaint Moderated	Public	Anonymous	-

Recommendation 7: The Online Safety Act should make reference to eSafety or the Minister maintaining a list of functionalities of online services and expectations in relation to them. The Act could then apply to any online services having any of those functionalities, rather than categories such as “search” and “social media” which may become less applicable as technologies change.

[14] Andre Oboler and David Matas (2013). *Online Antisemitism: A systematic review of the problem, the response and the need for change.* Online Antisemitism Working Group - Global Forum for Combating Antisemitism.
https://www.jewishvirtuallibrary.org/jsource/anti-semitism/onlineantisem2013.pdf

3. Does the Act regulate things (such as tools or services) that do not need to be regulated, or fail to regulate things that should be regulated?

The act is very broad in its scope. It is hard to see how something might fall outside of the categories of social media service, relevant electronic service, or designated internet service given that relevant electronic service may include any “electronic service specified in the legislative rules”.[15]

It is appropriate that distinction be maintained between connectivity providers, those controlling the physical infrastructure or giving end users access to the internet through it, and those making available online content. The connectivity providers should have the same immunity in relation to content as telephone companies (which some in fact are) for what travels through their wires or wireless networks. These services, provided they adhere to net neutrality,[16] are fundamentally different to website hosts, search engines, social media, games, or other content related services.

There is currently a debate over Canada’s proposed act as some provisions of it seek to impose obligations on connectivity providers which from a technical perspective do not belong there.

Recommendation 8: The Act should maintain a distinction between connectivity providers that respect net neutrality, which should have immunity in relation to content they carry, and services which relate to content and which should have obligations to mitigate or prevent harms from that content.

4. Should the Act have strengthened and enforceable Basic Online Safety Expectations?

The issues paper notes that, “The Commissioner may also prepare and publish service provider notifications about a service’s failure to comply with a reporting notice. These measures boost transparency for users and create reputational risks for service providers, encouraging improvements to policies, processes, and human and technological interventions to keep Australian end-users safe on their platforms.”[17] With some service providers promoting themselves as “free speech platforms” and as rebels against government tyranny, the risk of reputational harm to a platform for non-compliance may be low. In fact non-compliance may even be a positive for certain target audiences. The use of civil penalties when a service provider fails to comply with a reporting notice or determination issued by the Commissioner are the more important element and need to be readily used, we would recommend the option of warning only be used when there are mitigating circumstances.

---

[15] *Online Safety Act 2021* (Cth) S 13A(g).
[16] https://www.fcc.gov/net-neutrality
[17] Page 15.

The lack of penalties for a service provider failing to comply with the expectations outlined in the Basic Online Safety Expectations Determination are a clear weakness. Willful non-compliance has demonstrated the need for a stronger approach. At the same time, some of the expectations, and those consulted on in November 2023, are too nebulous to properly measure. Basic expectations may either be a tick box exercise, e.g. do you have a process for dealing with online hate, or they may be measurable with acceptable limits being provided (as occurs for quality of service in some sectors), or level of pollution (in other sectors). A systems approach must have elements that are measurable and which indicate when the system is performing at or above acceptable standards and when it is failing to do so. The measurements must be meaningful, appropriate, verifiable, and not prohibitively expensive to obtain.

To give an example, “service providers review and respond to reports and complaints within a reasonable period of time, and provide feedback to users on the actions taken”[18] is a nice idea, but not measurable. What is a reasonable time? Does it vary based on the nature of the contents (a video may take longer to review than a comment)? Further, if there is no quality of service standard (i.e. how accurate the reviews must be) measuring how quickly it occurs can be harmful as it can incentivise the creation of a poor quality system that quickly gives the wrong answer most of the time. If a quality of service is applied it requires measuring to verify if the quality standard is being met.

Recommendation 9: At least some Basic Online Safety Expectations should be enforceable and some of these should be expressed in a form that can be measured, with acceptable limits indicated. The measurements must be meaningful, appropriate, verifiable, and not prohibitively expensive to obtain.

5. Should the Act provide greater flexibility around industry codes, including who can draft codes and the harms that can be addressed? How can the code drafting process be improved?

Greater engagement with civil society, and other government agencies, may improve the process. This could help define societal expectations. Once sent, a “comply or explain” model as used for directors obligations may be useful.

Recommendation 10: To protect the public interest, civil society organisations, including community representative bodies, charities with an online safety or human rights objective, professional bodies with relevant expertise, and academics should have a means to provide feedback on industry codes and their appropriateness, or the need for changes to them.

---

[18] “Amending the Online Safety (Basic Online Safety Expectations) Determination 2022 (BOSE Determination)—Summary of the BOSE Determination and proposed amendments”, *Department of Infrastructure, Transport, Regional Development, Communications and the Arts*, November 2023 https://www.infrastructure.gov.au/sites/default/files/documents/amending-the-online-safety-basic-online-safety-expectations-determination-2022-bose-determination-summary-of-the-bose-determination-and-proposed-amendments-november2023.docx

6. To what extent should online safety be managed through a service provider's terms of use?

There is a danger that turning voluntary commitments expressed in the terms of service into enforceable rights could lead to protections in the terms of service being removed so that only the minimum required by law is included. This would be a reasonable risk mitigation process for companies to engage in and would have the opposite effect of what the proposal intends..

We have found that as a civil society organisation we have far more traction than governments with many social media companies. We can secure rapid action on harmful content because it is the right thing, and doing it as our request is essentially the platform taking positive action themselves. By contrast, if the same request came from a government, the platform feel they need to ensure there isn't any overreach and that the rights of users to protection from the government are protected. This is largely based on the culture in the United States where most platforms are based. That culture holds that governments should be the absolute last ones to decide what people can or can't say, particularly online. As most online content is a form of speech, this can lead to strong resistance and careful checks on any compliance with government requests.

We recommend expectations be established by government outside of the terms of service and be consistent for different providers with respect to equivalent services. We also recommend that the terms of service be (as stated by a platform to their users) required to meet or exceed the minimum legal standard. This allows for a layered response where expectations are in the law, expected norms are in the terms of service, and guiding principles can be provided in a service's statement of mission or purpose.

To make an analogy, online safety standards should be compared to other safety standards e.g. the electrical safety standards AS/NZS 3760:2022 ("Wiring Rules") which are not optional, and not to a voluntary scheme such as "Australian Made". A commitment to online safety should not be a "good idea" with potential market value, it should be a legal requirement. This is a fundamentally different mindset. Terms of services may provide an opportunity for companies to go above and beyond, which may provide them with a commercial advantage, but this should be in addition to and not in place of expected safety standards.

Recommendation 11: Enforceable expectations should be set by government and exist outside of the terms of service of any one company, and should be consistent for different providers with respect to equivalent services.

7. Should regulatory obligations depend on a service provider's risk or reach?

Higher risks justify higher regulations.

If we define risk as impact x likelihood we can consider reach a major factor in likelihood. In this case a platform's reach can impact its risk, but other factors are also important, including the capacity for harmful content to be posted in the first place, its ability to go viral and be seen, and the speed and effectiveness of the platform's response. Reach should therefore

not be a factor on its own, and other factors may mean a platform with high reach in fact has lower risk than one with a much smaller reach but other significant risk factors.

Two examples:

The ABC website, https://abc.net.au, is a very popular website in Australia. It received 10.4 million unique visitors a month in 2019-2020 from within Australia and based on 2022 figures its international audience is almost as large.[19] The likelihood of harm occurring if harmful content was posted on the ABC site could be considered high because of its reach, but once we consider that only ABC staff can post content to the site (no user posts, comments, pictures etc), and that even the content from staff goes through an editorial process, the likelihood of harm would have to be considered very low.

The 8chan website in 2019 attracted around 15 million unique visitors a month and was particularly popular in the US.[20] Only part of its audience visited /pol/, a forum we are interested in here. The reach of /pol/ to Australian audiences would be only a part of the (global) visitors to /pol/. Put another way, it would be only a tiny fraction of the Australian public compared to the number of Australians visiting the ABC. The reach is therefore lower, but there are other factors that make the likelihood of harm far higher: unlike the ABC, 8Chan /pol/ allowed users to post content and upload images and links, users were anonymous by default with identifiers changing every 24 hours, content vanished after 24 hours, and the culture of dedicated to far-right ideology.[21] Such a site posed a very high risk which is why we initiated calls for the forum tobe closed. These of course are extremes at opposite ends of the spectrum. They highlight that reach alone may not be a good metric, but it does contribute to risk.

Regulatory obligations should depend on a service provider's risk. Reach should be considered a factor that raises the likelihood of a risk leading to harm, though other factors (such as a lack of user content) may almost entirely nullify it, and other factors such as effective moderation, may substantially lower the risk. Regulatory obligations should be seen as risk mitigations. They should be obligatory on large platforms unless the platform demonstrates the risk either does not exist or has been mitigated in some other acceptable manner. For example, regulation for live streaming services might not be needed for a platform that only allows live streaming by trusted content providers who are acting in a professional manner (rather than offering the service to the public at large).

Recommendation 12: Regulatory obligations should depend on risk, taking into account reach, functionality, and mitigations.

---

19 https://www.transparency.gov.au/publications/communications-and-the-arts/australian-broadcasting-corporation/australian-broadcasting-corporation-annual-report-2021-22/audience-data-and-analysis/audience-reach

20 https://www.npr.org/2019/08/05/748166877/the-website-where-violent-white-supremacists-state-their-case

21 Andre Oboler, William Allington and Patrick Scolyer-Gray (2019). *Hate and violent extremism from an online sub-culture : the Yom Kippur terrorist attack in Halle, Germany.* Online Hate Prevention Institute. https://nla.gov.au/nla.obj-2286730824/view

Protecting those who have experienced or encountered online harms

We note that outside of the specific schemes managed by eSafety, there is no funding through eSafety to help protect those who have experienced or encountered other online harms. This is despite the existence of civil society organisations that tackle those harms. Australians who are impacted by online harms other than those eSafety deals with directly are therefore doubly disadvantaged, first by eSafety not addressing those harms, and then again through a lack of funding to support the civil society organisations, such as the Online Hate Prevention Institute, working in this space.

We note that this is somewhat glossed over in the issue paper where it says "eSafety provides grants to organisations working to improve online experiences for Australians". It goes on to more specifically explain that this means children's online safety. As explained, the funding almost all went to education providers endorsed under the Trusted eSafety Provider program, which in 2022-23 educated 1.4 million people including 1.1 million school students, 140,000 parents and 31,000 educators.[22] The rest of the funding is focused on preventing violence against women and children under a separate scheme.

Recommendation 13: Grants from eSafety need to be more widely available to address a greater variety of online harms and need to support partnerships with civil society organisations working in the online safety space and addressing issues not covered, or less well covered, by eSafety itself.

See also recommendation 4.

The section on online safety promotion (page 31) discussed engagement with police; the Australian Centre to Counter Child Exploitation; industry and technology companies; cooperation with government, Catholic, and independent school education bodies; but it does not mention civil society in a general manner, one which would encompass charities working on other aspects of online safety. The issues paper notes, "While community engagement with eSafety is increasing, there are still opportunities to increase Australians' awareness of the support eSafety provides",[23] but it then focuses on parents, teachers, carers and supervisors. The focus is entirely on children's online safety. This is a reflection of the culture in eSafety and strikes a very different tone to that of the introductory letter from the Minister.

See recommendation 3.

8. Are the thresholds that are set for each complaints scheme appropriate?

The threshold should take into account the impact of the sanction. Criminal sanctions that deprive people of liberty require a very high standard. Civil sanctions which may result in

[22] Page 30.
[23] Page 31.

monetary penalties have a lower standard. When it comes to removal of online content, the impact is even lower than speeding fines or parking fines. In light of this it appears the threshold for adult cyberbullying is unnecessarily high. The standard should be lower than that of defamation claims (which require a loss) a person seeking only removal of the content should be able to avail themselves of the adult cyberbullying scheme in cases which are currently rejected. The threshold for the scheme related cyberbullying of a child appears more appropriate for both adults and children.

In the case of content which might form the basis of a defamation claim:

- If the poster is outside of Australia and the target is Australian, and there is no public interest case for the content to be available, eSafety should be able to request that a platform block the content in Australia.
- If the poster is in Australia, and there is no public interest case for the content to remain available, eSafety might request the poster or platform to remove the content.

Where there is a public interest case to be made for the content remaining online, eSafety may leave the matter to be resolved through a concerns notice and potentially a definition of the courts through a defamation proceeding.

Recommendation 14: The threshold for cyberbullying of an adult should be lowered to be similar to that of cyberbullying against a child. It should be appropriate to use eSafety to block defamatory provided there is no public interest in the content being public. If there is a potential public interest in the content being public, the issue should be decided through a defamation claim.

9. Are the complaints schemes accessible, easy to understand and effective for complainants?

The complaint schemes seem effective for the content they cover, however, limitations in their coverage are concerning. This is particularly the case around hate speech targeting groups, which may impact individuals in that group but fall outside the act as an individual isn't specifically targeted.

The harm from online content targeting a group was discussed in the very first case related to online racism in Australia, Jones v Toben.[24] In that case it was held that the Adelaide Institute website, which promoted Holocaust denial, would more likely than not "engender feelings of hurt and pain" and a "sense of being treated contemptuously, disrespectfully and offensively" in Jewish Australians.[25] The court also noted how the presence of such material "would cause damage to the pride and self-respect of vulnerable members of the Australian Jewish community, such as, for example, the young and the impressionable", how some "might well experience, whether consciously or unconsciously, pressure to renounce the cultural differences that identify them as part of the Jewish community", and "that it is more probable than not that there are members of the Australian Jewish community who will become fearful of accessing the World Wide Web to search for information touching on their Jewish culture because of the risk of insult from the material".[26] These are all good reasons

[24] *Jones v Toben* [2002] FCA 1150.
[25] Ibid [93].
[26] Ibid [96].

why Holocaust denial material should be removed, as was ultimately ordered by the court. The case, however, took ten years and a huge expense. It is not a reasonable process for removing content in today's online world. The eSafety Commissioner provides a faster and more effective process and it is unconscionable that it is still unable to be used in such circumstances.

See recommendation 1.

10. Does more need to be done to make sure vulnerable Australians at the highest risk of abuse have access to corrective action through the Act?

As just noted, the courts found a risk of harm from online content to vulnerable Australians back in 2002, yet even if such content is demonstrated to be causing harm to one or more children, even then it doesn't fall within eSafety's powers as the act currently stands.

See recommendation 1.

11. Does the Commissioner have the right powers to address access to violent pornography?

Yes, but we are concerned with reliance on the classification scheme or abhorrent violent content scheme. There may be a category of content that can be explicitly proscribed (rather than relying on the classification scheme), yet falls below abhorrent violent content.

12. What role should the Act play in helping to restrict children's access to age inappropriate content (including through the application of age assurance)?

The Act should allow the Commissioner to, on request, classify websites / applications as suitable for very young children if they meet certain requirements. This should be a higher standard than for online services for the general public. We favour the approach of YouTube which has a separate site (www.youtubekids.com) with different functionality that caters for younger children and allows parents to make decisions on the configuration. This acknowledges that very young children are now using the internet, sometimes even before they can read, and this should be possible in an appropriately safe manner. It also acknowledges that more content such as children's television, children's books and audio books, and children's music are delivered digitally.

The classification of specific sites as requiring age restriction should not be a function of the eSafety Commissioner but of ACMA / the Classification Board.

eSafety should determine (in a platform neutral manner) the safety requirements that are needed for sites that have users below a certain age. This would essentially mean dividing the Basic Online Safety Expectations (BOSE) into multiple schemas so there are stricter requirements for platforms that allow younger users. The Minister may then set expectations for different age ranges and the eSafety Commissioner would set acceptable ways for platforms to meet those requirements. E.g. the required responsiveness to bullying complaints for a platform restricted to those over 16 might be lower (slower, lower accuracy) than for a platform that allows anyone over 13 to use it.

Systems that are designed for a general audience including children:

- Must take a more responsible approach and use business models that avoid incentivising harm.
- Be built using safety by designed principles to address known risks more completely than currently required by the Basic Online Safety Expectations. Advanced Online Safety Expectations would need to be created to provide the expected requirements.
- Require regular certification that platforms still meet the requirements for exemption.

Advanced Online Safety Expectation would, among other requirements, need to:

- Outlines specific risks to be avoided or mitigated
- Prescribe certain mitigation mechanisms for some risks
- Proscribe tolerance levels for harms that cannot be entirely avoided and requirements to ensure this is measured, transparently reported, and independently verifiable. Environmental protection mechanisms provide a useful model.
- Require regular transparency reports that are specific to the Australian experience of using the platform, and specific in terms of the type of harm. For example, hate speech is not a specific enough category, but hate speech targeting First Nations Australians is. The intersection of multiple categories of hate should also be reported when over a specified level, for example, hate speech targeting First Nations Australians that involved gendered hate.

The eSafety Commissioner might also provide policy advice on what constitutes an acceptable mechanism for enforcing age restriction and might undertake research to determine if such mechanisms, once implemented, are working effectively.

Recommendation 15: Different standards ought to apply to platforms that cater to different age ranges. The standards should be set by the Minister in the BOSE, while acceptable mitigations to meet those standards should be advised by eSafety.

13. Does the Commissioner have sufficient powers to address social media posts that boast about crimes or is something more needed?

Addressing these posts is primarily a law enforcement role, not an eSafety one. The role of eSafety in such cases may involve securing the content's removal, e.g. on referral from police, after ensuring police have archived it for evidence. The process for securing identity information from online platforms is one carried out by police using the Mutual Legal Assistance Treaty (MLAT) process, not by eSafety.

A general power allowing eSafety to seek removal of unlawful content on referral from a competent Federal Government, State Government or Territory Government Agency would facilitate this. Such agencies may be human rights agencies, police, courts, or other law enforcement agencies (e.g. Fisheries Investigators have had cases of social media being used in the illegal sale of abalone and lobster[27])

See recommendation 2.

[27] https://vfa.vic.gov.au/about/news/st-albans-duo-arrested-for-abalone-trafficking

14. Should the Act empower 'bystanders', or members of the general public who may not be directly affected by illegal or seriously harmful material, to report this material to the Commissioner?

Australia has a habit of focusing too strongly on government and not enough on partnership with civil society. The trusted flagger system in Europe allows authorised civil society organisations to notify platforms of breaches as a first step. Trusted flaggers can take complaints from the public, verify the breach, then report it. This reduces the burden on regulators and allows platforms to prioritise reports from the trusted flaggers. Such trusted flaggers could also be given a role as third party reporters to eSafety (reporting on behalf of others, potentially even while maintaining anonymity of the original reporter). This would avoid opening a floodgate, but would allow a significant increase in resources. In addition to civil society, education departments responsible for government schools, and sector bodies for private schools, might employ someone in a position where they act as a trusted flagger.

It would be appropriate for community representative bodies and civil society organisations focused on human rights to be trusted flaggers, as occurs in Europe. Members of parliament are also trusted flaggers, allowing elected representatives (or their staff) to assist their community. We note that in the past some platforms introduced trusted flaggers themselves, but removed this status from an organisation if the organisations reported content that was within the terms of service of the platform too often. This effectively forced trusted flaggers to only enforce the existing rules set by the platforms and not hold the platforms to account for content that should have been against the rules, and might even be against the law, but was not covered by the platforms terms of service. This approach is inappropriate and it would be better if trusted flaggers were determined by eSafety and could then claim this status with any eSafety regulated platform with which they engaged. Claiming the status may require registering an account with the platform and notifying the platform that the account belongs to an Australian trusted flagger.

Recommendation 16: A trusted flagger system for approved Civil Society Organisations should be added to the legislation. It should be eSafety rather than platforms that determine who qualifies as a trusted flagged.

There may be some kinds of content, such as cyber bullying and hate speech, where trusted flaggers are more appropriate, and other types of illegal and seriously harmful material where anonymous reporting from the public either to eSafety or police is more appropriate. The National Security Hotline[28] is an example of such a service which anyone can use. The Commissioner's office may not be the appropriate frontline for reports from the public at large, at least not without substantially increased resources, which may in some ways duplicate services in other areas of government. Allowing those who are personally impacted to report directly to eSafety allows faster responses. With limited resourcing, protecting this capacity is important. Expanding it slightly to allow reports from approved organisations that can support and advocate for victims, as discussed above, is a reasonable compromise.

[28] https://www.afp.gov.au/crimes/terrorism

There is no concept of a specialist civil society organisation, civil society partners, or trusted flaggers in the legislation. In practice such partners do exist, and by virtue of having contact points within eSafety and other relevant agencies, the way they operate is different to that of the public. It may be worth incorporating in the legislation in the manner discussed above. I note that Harm Prevention Charities are particularly relevant as they by definition seek to mitigate or prevent harm to human beings. The barriers for becoming such a charity have been substantially relaxed as part of recent reforms, so relying on this status alone may no longer be enough and an approval process managed by either eSafety or perhaps the Minister may be needed.

Recommendation 17: Anonymous reporting to eSafety should be possible for some kinds of online safety violations and the anonymity of users of such services should be protected by the Act.

I note that the Online Hate Prevention Institute currently shares certain information with eSafety along with other government agencies, including law enforcement agencies, in the form of our confidential reports. These are not “reports” or “complaints” but a sharing of information. When sharing such information some agencies thank us for the information but inform us it would need to be lodged in a different way to count as a report. We seldom engage in those extra steps as it places an additional cost and burden on us. We are not funded by the government to cover the cost of meeting such burdens. We trust that having made the information available, relevant agencies will ensure it is used appropriately. We have at times been asked if the information may be passed on to different agencies, for example from AFP to ASIO, and always consent to this. We also list the agencies, companies, and civil society organisations that have received a confidential report in order to facilitate collaboration.

15. Does the Commissioner have sufficient powers to address harmful material that depicts abhorrent violent conduct? Other than blocking access, what measures could eSafety take to reduce access to this material?

We have succeeded in removing both videos and manifestos that constitute abhorrent violent content on many occasions. We have asked eSafety to ensure the content is referred to the classification board where it would be given an RC rating. This process is important as it enables groups like ours to point those we find hosting (or in the case of search engines linking) to such material to the listing when we ask for it to be removed. Having a description and a declaration that the content is illegal makes the process for us much smoother.

If eSafety will not refer to the classifications board (a requirement that I believe may have been removed from the legislation), and even if it is does, it would be useful for the Commissioner’s office to maintain its own public list describing content that has been declared to be abhorrent violent content. As far as we are aware, the Christchurch video and manifesto are the only two instances of abhorrent violent content given an actual RC rating and appearing in the list.

Recommendation 18: A description of proscribed abhorrent violent content should be maintained either with the Classifications Board or by eSafety, or both, in order to facilitate

voluntary removal (potential on advice from online safety focused civil society organisations) by companies who find they are hosting such content.

16. What more could be done to promote the safety of Australians online, including through research, educational resources and awareness raising?

The Online Hate Prevention Institute is a charity dedicated to online safety in an area that is largely overlooked by eSafety. While eSafety has some webpages on online hate that link to our resources, and those of others, and has carried out some relevant research, this remains a real deficiency.

Here is an example of projects we have worked on just this year in the online safety space and which largely fall within the broad remit of online safety, but outside of the specific focus of eSafety, or which eSafety mich currently support through its grants:

Project 1: Online Antisemitism in Australia, a partnership with the Executive Council of Australian Jewry, has been running for 18 months and monitors 11 online platforms on an on-going basis. It has resulted in over 5,500 items of data being collected and three reports published. The data is classified into 27 categories, the prevalence of different categories and the effectiveness of platforms in removing each category varies. The project incorporates data from intensive monitoring in the Moment Project described below.

Project 2: Building regional and national capacity in civil society to counter extremism is a project funded under a grant from Home Affairs. The project supports the training and employment of three young people from remote communities to monitor and analyse harmful online content, create publications, provide training to local organisations, and support OHPI Exit's training for those supporting people vulnerable to extremism.

Project 3: New Zealand Online Antisemitism Project sees 4 young people in New Zealand being employed in New Zealand and seconded to OHPI for training and management to monitor and write articles about online antisemitism in New Zealand. 514 items of antisemitism have been collected and 6 articles produced in the last 2 months.

Project 4: The Moment Project documents both online antisemitism and online anti-Muslim hate between October 2023 and February 2024. The project involves a total of 320 hours of intensive monitoring on 10 different social media platforms, with equal time spent on each platform and that time evenly divided between work on antisemitism and anti-Muslim hate on the platfrom. The monitoring distinguishes between 27 types of antisemitism and 11 types of anti-Muslim hate. A resulting report Online Antisemitism After October 7 shows a dramatic rise of online antisemitism, varying from 350% to 1000% by platform. The main type of antisemitism is traditional antisemitism, such as conspiracy theories, deicide, and blood libels. This traditional antisemitism is also used in relation to Israel and accounts for most of the Israel related antisemitism, e.g. accusing Israel of controlling the media or other national governments. Takedown rates a month or more after content was first reported varied by platform from 4% to 36%. A forthcoming report on anti-Muslim hate (which also covers anti-Palestinian racism and anti-Arab racism) notes this hate has also risen and is particularly prevelant on X (Twitter) and platforms used by the far right. The most common categories of anti-Muslim hate are demonising / dehumanising Muslims, presenting Muslims

as a cultural threat, and presenting Muslims as a security risk. Take down rates a month or more after data was gathered vary from 2.9% to 38%, again, far from acceptable. Takedown rates on hate against Jews and Muslims both improved when we reexamined the same data in June 2024 (three months after last checking it), but they remained below what the public would expect.

Project 5: Our Referendum Project identified 161 media articles that were shared in 528 social media posts during the campaign. Each post had 10+ comments, and we collected and analysed a total of 37,785 comments that were made across these posts. The posts were analysed for disinformation about The Voice, campaigns and the referendum process. We also looked for racism against First Nations People and others, cyberbullying and other kinds of hate. We found that 2249 of the total comments contained misinformation about Campaigns, and 1004 contained anti-Indigenous racism. These numbers are after removal by both platforms and media companies, so the initial numbers would likely have been significantly higher.

Project 6: International police training. In partnership with a European Civil Society Organisation, the Online Hate Task Force, and the Brussels South Police district, and with the participation of the European Commission, we ran six hours of training over two days focusing on online hate targeting Jews and Muslims. Australian police from the AFP, NSW Police, and Queensland Police attended along with police from Europol and from Brazil, Brussels, UK, Italy, The Netherlands, Switzerland, Sri Lanka, Pakistan, Philippines and more.

Penalties, and investigation and information gathering powers

We have recommended a number of times that a small civil penalty be introduced for end users who engage in harmful conduct that would, at present, simply see the platform required to remove the content. Higher penalties, and criminal penalties, make it less likely most harmful online conduct will be discouraged. Removal of content is not a sufficient deterrent.

Recommendation 19: eSafety should be given the ability to level small fines to users who engage in harmful online conduct, with larger fines for repeated or more serious offences where criminal proceedings are not being pursued.

17. Does the Act need stronger investigation, information gathering and enforcement powers?

The transparency reporting under the Basic Online Safety Expectations are not fit for purpose. We need transparency reports related to Australian content, and Australian reports of content (whether the content originated) and how they were addressed. We also need this data broken down in greater detail. Hate speech alone is not a useful category, it needs to be broken down by the group(s) targeted.

18. Are Australia's penalties adequate and if not, what forms should they take?

The penalties for platforms are not keeping up with international norms and need to be increased to similar levels to those in other countries. Without this, online safety in Australia will be a secondary concern with companies failing to invest in local staffing in Australia and prioritising complaints from countries where the risk they face is higher. In the EU for instance they have a new reporting facility within platforms that are tied to EU law and reports made using this system are getting priority responses.

Recommendation 20: Penalties should be increased to be in-line with those overseas.

19. What more could be done to enforce action against service providers who do not comply, especially those based overseas?

Bilateral or multinational treaties with countries with larger markets and a greater ability to enforce compliance would be helpful. Ideally an international system needs to be developed so the international rule based system can extend to harms in cyberspace. In the meantime, perhaps it is possible to aim for compatibility with the EU and to work towards a system where their infrastructure might be used more widely. This may be a task for DFAT rather than eSafety.

20. Should the Commissioner have powers to impose sanctions such as business disruption sanctions?

Yes, particularly for repeated violations or non-compliance. Such powers would also be useful for smaller platforms based overseas that simply refuse to engage.

International approaches to address online harms

21. Should the Act incorporate any of the international approaches identified above? If so, what should this look like?

The shift away from incident based responses to systemic responses is the only viable approach in the long term, unless eSafety can operate on a cost recovery basis so platforms cover the costs of the incident base work (both valid and invalid complaints). This would essentially be the platforms outsourcing what should be an internal function to the government.

22. Should Australia place additional statutory duties on online services to make online services safer and minimise online harms?

The BOSE does this to an extent, but without being enforceable. Such a duty, would be useful provided the expectations on what meets the duty are clear and regularly updated so

new best practices are continually shifting to become the new expected norm and standards continue to rise, only dropping when there are technological disruptions (for example generative AI is creating new challenges).

23. Is the current level of transparency around decision-making by industry and the Commissioner appropriate? If not, what improvements are needed?

Industry transparency reports need improvement, as previously discussed. The lack of local contact points for many companies make it difficult to discuss decision making with them. Meta is an exception with an Australian advisory group, which we serve on, and where key civil society organisations are briefed and given the opportunity to ask questions of company experts from Australia and around the world. A similar investment and engagement from other platforms would be welcome. Greater engagement by eSafety with civil society organisations engaged in online safety work outside of the schemes eSafety administers would also be welcomed. The group Meta has assembled would provide a good start.

See recommendation 3.

24. Should there be a mechanism in place to provide researchers and eSafety with access to data? Are there other things they should be allowed access to?

Cambridge Analytica highlighted how such access can be seriously abused. We not convinced a pipeline should be provided to researchers for general research, as Twitter did in the past, though it is clearly useful to researchers. Data for approved online safety projects (reviewed and approved by eSafety) may be useful.

The provisions of the EU's Digital Services Act, including allowing access for auditing and compliance monitoring, would be useful for Australia. Such access could be provided by civil society under contract to the government. Allowing for multiple compliance monitoring efforts by different researchers using different methods, perhaps specialised in different forms of online harms, would be best.

A legal protection for data scraping for approved online safety projects (whether research or operational) combined with an exception from the Privacy Act, and subject to some limitations on the use of the data, would be useful to online safety organisations like the Online Hate Prevention Institute.

Recommendation 21: Civil society organisations working in online safety should, one approved by eSafety, have a legal right to scrape data and an exception from the Privacy Act, subject to certain obligations.

We are also particularly frustrated with platforms that do not have a mechanism to extract a URL directly to reportable content. For example, on TikTok you can report a comment, but you cannot get a URL that will take you directly to that comment. This makes is very difficult to check if reported comments have been removed, or to share them in a useful manner with relevant agencies.

Recommendation 22: Platforms should be required to provide URLs that directly link to any user generated content that may be used to breach online safety.

25. To what extent do industry's current dispute resolution processes support Australians to have a safe online experience? Is an alternative dispute resolution mechanism such as an Ombuds scheme required? If so, how should the roles of the Ombuds and Commissioner interact?

Many platforms are largely uncontactable by the public. There simply isn't a mechanism to speak to a person. We often serve as an intermediary passing on concerns of specific users and civil society organisations to our contacts at the platforms we maintain relationships with. We are careful to ensure we only pass on strong cases, and these concerns are almost always rapidly addressed.

I suspect an Ombuds will be rapidly overwhelmed if they end up being the gateway to have platforms address concerns people are simply unable to bring to the attention of a platform representative in any other way. From our experience, complaints are likely to range from recovery of compromised or wrongly closed accounts, to appeals on wrongfully removed content or failure to take action on other abusive or harmful content.

26. Are additional safeguards needed to ensure the Act upholds fundamental human rights and supporting principles?

The act does not seek to uphold fundamental human rights in general, outside of the Basic Online Safety Expectations. It is focused on a few specific and enumerated schemes. More general expressions about online safety in the Act tend to be interpreted inline with the specific schemes and the original focus on children's online safety. As stated before, the Minister's letter appears to take a far more holistic focus to online safety than the existing act and the briefing paper.

Regulating the online environment, technology and environmental changes

Three elements must be regulated when addressing harms in the online environment:

- The level / frequency in which harm occurs,
- The response time in addressing reported harms, and
- The effectiveness of the response to reports of harm.

A system that has a very high degree of harm, but responds quickly and accurately when reports are made is a problematic system that is failing to put in place mitigations.

A system that has taken reasonable steps to mitigate harm, reducing it to a more acceptable level, but takes too long to address reports of harm is a problematic system.

A system that has taken reasonable steps to mitigate harm, responds in reasonable time, but gets the response wrong, either rejecting reports of actual harm and refusing to address the harm, or being too quick to remove content and being gamed into removing content based on false report, is also a problematic system.

Regulation needs to set expectations for the degree of harm that can be found on a platform, the response time for reports, and the degree of accuracy expected in reports. These expectations may vary with different types of harms, and for different types of content.

In order to have systemic improvement, these metrics all need to increase over time so that the best practice of today becomes the expected practice of tomorrow. There will be slippage due to innovation as technology changes and new challenges emerge, but outside of this there should be a push towards continual improvement in online safety.

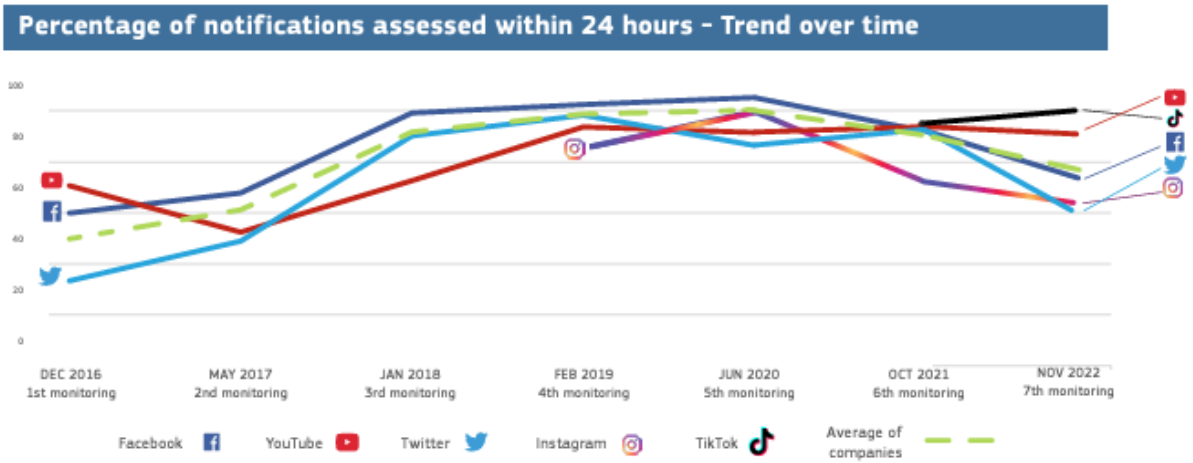
Recommendation 23: Metrics to assess compliance with preventing online harms should measure the level / frequency in which harm occurs, the response time in addressing reported harms, and the effectiveness of the response to reports of harm. Thresholds for acceptable values of reach of these metrics should be set by the government and routinely reviewed and adjusted to ensure there is a continual improvement in online safety.

The longest standing expectation, and measurement of it, comes from the European Commission. In May 2016 the European Commission, Facebook, Microsoft, Twitter and YouTube agreed a “Code of conduct on countering illegal hate speech online”.[29] A key feature was the commitment that, “The IT Companies to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary.”[30] This expectation has been embedded in a range of European legal instruments, most recently the Digital Services Act, and is no longer voluntary. This occurred as a voluntary agreement wasn’t found to be effective. The EU runs periodic evaluations of the code of conduct, the most recent evaluation report was published in November 2022.[31]

The monitoring exercise involves a range of government agencies and civil society organisations that tackling online hate monitoring all the items they report over a particular 6 week period. The number of items that are accessed within 24 hours are then recorded and reported. The Commission also assesses the rate of removal overall and on a per country basis. A graph of the removal rates from 7th reporting exercise is shown below.[32] Since 2020 the percent of reports actioned in 24 hours has been falling. Removal also dropped significantly on some platform, Facebook fell from 87.6% of complaints resulting in removal in 2020 to 69.1% in 2022, TikTok from 80.1% in 2021 to 60.2% in 2022. This mirrors a period in which many of the companies have dramatically cut the number of staff they have working in Trust and Safety, the teams which review user reports.

---

[29] https://commission.europa.eu/document/download/551c44da-baae-4692-9e7d-52d20c04e0e2_en
[30] Ibid.
[31] https://commission.europa.eu/system/files/2022-12/Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf
[32] Ibid.



Data in the monitoring exercise is not ideal as it is an amalgamation of data from different agencies and CSO, which focus on different types of hate, and on unlawful content based on what is unlawful in their jurisdiction. The same level of effort is not put into each type of hate, or across each type of hate in each country. These challenges can be address in the context of Australia, as we did doing a recent three month exercise monitoring antisemitism and anti-Muslim hate, where we ensuring we put the same level of effort into each of these hates on each of 10 platforms. The resulting takedown rates (after at least 4 months since the data was collected, and having shared the data being monitored with Facebook, Instagram and YouTube) were as shown in the table below:

	Antisemitism		Anti-Muslim Hate	
Platform	Items Removed	% Removed	Items Removed	% Removed
Facebook	191	59.5%	54	49.5%
YouTube	73	31.7%	10	14.7%
Twitter / X	124	32.5%	73	27.7%
Instagram	95	48.2%	31	39.2%
TikTok	58	29.6%	34	43.0%
Reddit	61	25.8%	29	31.9%
LinkedIn	123	47.9%	22	32.8%
Gab	109	26.7%	45	28.3%
BitChute	210	65.6%	52	54.2%
Telegram	52	14.8%	30	19.1%

A regular monitoring exercise for Australia would be very useful. This is the sort of exercise best managed in collaboration with civil society, idealling within a network of trusted partners.

Recommendation 24: There should be an Australian monitoring effort similar to that run by the European Commission, with the involvement of appropriate civil society organisations.

27. Should the Commissioner have powers to act against content targeting groups as well as individuals? What type of content would be regulated and how would this interact with the adult cyber-abuse and cyberbullying schemes?

Our view is that the Commissioner should definitely have power to act against content targeting groups in addition to content targeting individuals. We have recommended this since 2013. Australia should also sign the 2003 Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems.[33]

See recommendation 1.

The issue paper states that "hate speech is highly contested and context dependent, and these policies are not always enforced in line with community expectations". We agree that policies against hate speech are not enforced inline with community expectations, but we disagree that the regulation of hate, or the identification of it, is "highly contested" or to use another word controversial.

While there were ideologically driven efforts to overturn S 18C of the Racial Discrimination Act by the previous government, these moves (ultimately rejected by the Senate) ran strongly against public opinion. One survey found that "just 10% of Australians believe people should have the freedom to 'insult' and 'offend' people on the basis of race, culture or religion. Over 75% are opposed."[34] To the extent there is disagreement on what constitutes hate speech, this needs to be addressed with research and public education. Regarding online hate in particular, a representative survey carried out by eSafety found "The overwhelming majority of people support action to check the spread of online hate speech including the introduction of legislation and getting social media companies to do more."[35] The report found 69% agreed or strongly agreed that online hate was spreading in Australia and around the world, 78% felt social media platforms needed to more to stop the spread of hate, and 71% supported the introduction of specific legislation against online hate speech.[36] This is similar to Canadian data showing 80% of Canadians believe hate speech is a problem in social media and 45% believe it is a major problem.[37]

eSafety reports that "over 50% of young people have seen or heard hateful comments about a cultural or religious group online".[38] A much narrower question was used when surveying adults with eSafety's research finding 14% of Australian adults were the targets of hate

---

[33] https://rm.coe.int/168008160f
[34] Andrew Jakubowicz, Kevin Dunn, and Rachel Sharples, "Australians believe 18C protections should stay", The Conversation, 16 February 2017. https://theconversation.com/australians-believe-18c-protections-should-stay-73049
[35] "Online Hate Speech: Finding from Australia, New Zealand and Europe", eSafety, 2020. https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf?v=1720955237919
[36] Ibid p. 7.
[37] https://www.cira.ca/en/resources/documents/state-of-internet/canadians-deserve-a-better-internet-2021/
[38] https://www.esafety.gov.au/young-people/online-hate

speech, based on the question “In the last 12 months..,…have you received a digital communication that offended, discriminated, denigrated, abused and/or disparaged you because of your personal identity/beliefs (e.g. race, ethnicity, gender, nationality, sexual orientation, religion, age, disability, etc.)?”[39] Had the question asked “have you seen digital content that offended, discriminated, denigrated, abused and/or disparaged a community based on its identity/beliefs (e.g. race, ethnicity, gender, nationality, sexual orientation, religion, age, disability, etc.)?” the figure would likely be far higher. Recent US Government research suggests about a third of US internet users experience hate speech online.[40] The number is likely similar in Australia.

Australia is significantly behind other countries and the European Union when it comes to addressing online racism and hate that targets protected characteristics. Meta has some of the most rigorous policies on hate speech and defines protected characteristics as: “race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease”. Australia could use this list, or to the report of the Victorian Parliament’s *Inquiry into anti-vilification protections* which examined the question of what should be protected in depth.[41]

We note that Australia has a special obligation on online racism against our First Nations people. This should be explicitly listed and coordinated with the Australian Human Rights Commission and the National Anti-Racism Strategy. When phrased in terms of the forms of hate, increasingly antisemitism and anti-Muslim hate are explicitly listed and counties have or are appointing coordinators or special envoys to tackle these two forms of hate. Australia is in the process of appointing special envoys to tackle these two forms of hate as well and it would make sense to coordinate with them. The national coordinators on antisemitism meet regularly and have a task force on online antisemitism which the Online Hate Prevention Institute has briefed. We also met recently with the European Commission’s coordinator on anti-Muslim hate. It is concerning that our work is recognised and relied upon globally, but not well integrated with eSafety as it is currently largely outside their remit.

The Issue paper also underplays the extent to which the DSA in Europe and other legislation contains a strongly focused on online hate speech. The DSA, for example, requires platforms to have a consistent reporting system for violations of European Law including hate speech.[42] The Online Hate Prevention Institute has attended hearings on the topic of online hate speech in the US Congress and testified to an inter-Parliamentary hearing in the European parliament. This is a very significant topic in the online space and a real blindspot in the Act. This is despite a significant amount of global thought leadership coming from Australian civil society.

In 2020 while serving on a panel for the UN Forum on Minority Issues I explained that to address hate speech, “Counter speech is not enough. Empowering young people is not

[39] Ibid, p. 4, 6.
[40] “Online extremism: More Complete Information Needed about Hate Crimes that Occur on the Internet”, United States Government Accountability Office, January 2024. p 37. https://www.gao.gov/assets/d24105553.pdf
[41] “Inquiry into anti-vilification protections”, Legislative Assembly Legal and Social Issues Committee, Parliament of Victoria, March 2021. https://www.parliament.vic.gov.au/get-involved/inquiries/inquiry-into-anti-vilification-protections/reports
[42] https://digital-strategy.ec.europa.eu/en/policies/safer-online

enough. These are the dominant approaches of recent years, they are what we have promoted even as the hate continued to rise. If we want to get on top of this problem, we need to take it more seriously. We need to invest both economically and through political capital in real solutions."[43] The remarks included some suggestions on how we can move forward globally. Australia still has a long way to go in this area of online safety, starting with a political commitment and including an expansion of the Online Safety Act to include group based hate.

28. What considerations are important in balancing innovation, privacy, security, and safety?

The idea that technological advancement, particularly through the internet, would be stifled by regulation was the dominant paradigm in the early years of the internet. Early technologists saw the internet as a domain beyond the control of governments, as was most clearly expressed in the poem "Declaration of Freedom of Cyberspace" by Electronic Frontier Foundation co-founder John Perry Barlow.[44] As the internet matured and became an essential part of daily life, this view has changed. Not only technologists and digital ethicists, but technology companies themselves have called for regulation.

In 2018 for example, the founder of the World Wide Web, Sir Tim Berners-Lee, warned how "conspiracy theories trend on social media platforms, fake Twitter and Facebook accounts stoke social tensions, external actors interfere in elections, and criminals steal troves of personal data" and that in responding to these issues technology companies aim to "maximise profit more than maximise social good" and he recommended that "a legal or regulatory framework that accounts for social objectives may help ease those tensions".[45]

In 2020 Mark Zuckerberg, CEO of Meta, called for greater regulation by governments of social media saying, "even if I'm not going to agree with every regulation in the near term, I do think it's going to be the thing that helps creates trust and better governance of the internet and will benefit everyone, including us over the long term.[46] In 2021 YouTube's CEO Susan Wojcicki said that "[New regulation] makes sense in many places, but in other places could have a number of unintended consequences. And we just want to make sure that as we work with regulators, that it achieves what they are looking for, as opposed to doing something that could be harmful, ultimately, to the creator ecosystem".[47]

As regulation is put in place, there is a need to balance innovation, privacy, security, and safety. The days of releasing products that are knowingly harmful to the public are long past - regardless of the industry. The use of the internet, like the use of public roads, involves some unavoidable risk. The question is how much risk the public can be expected to bear, and what are reasonable expectations of industry to prevent or mitigate that risk.

---

[43] https://ohpi.org.au/address-to-the-united-nations/
[44] https://www.eff.org/cyberspace-independence
[45] https://www.theguardian.com/technology/2018/mar/11/tim-berners-lee-tech-companies-regulations
[46] https://www.cnbc.com/2020/02/15/facebook-ceo-zuckerberg-calls-for-more-government-regulation-online-content.html
[47] https://www.marketplace.org/shows/marketplace-tech/youtube-ceo-susan-wojcicki-on-tech-regulation-and-transparency/

The term “gross negligence” has been used in Australian courts, but is not a term of art but rather one of degree.[48] As Justice Tottle noted,[49] Australian courts have followed the approach of Mance J in the English Case of The 'Hellespont Ardent' in which Mance J ruled that there was no subjective mental element of appreciation of risk needed for gross negligence to occur, but rather that it included “conduct which a reasonable person would perceive to entail a high degree of risk of injury to others coupled with heedlessness or indifference to or disregard of the consequences”.[50] Justice Tottle noted that this “heedlessness or disregard need not be conscious”.[51] It seems that while a duty of care obligation might be avoided by implementing recommended best practices to mitigate risk, gross negligence, as described above, should remain an exception where liability applies and a regulator is empowered to take action resulting in large fines.

The idea that innovation should be protected at the cost of privacy, security, and safety seems out of step with today’s expectations of technology. Innovation will result in unforeseen risks, but should be no excuse when it comes to foreseeable risks. In an unpublished paper on Pokemon Go we noted how the company advertised physical devices to use with the game on the basis that they removed the requirement for children to be looking at the screen while walking around looking for pokemon and that this would improve safety around traffic. This statement was made prior to the game being released, indicating knowledge of a safety risk to the public which rather than fixing, the company chose to exploit to increase revenues. Such action should in our view be regarded as gross negligence, particular when the foreseen fatalities later resulted.

Technology is constantly evolving. Innovation means some risks will not be foreseeable. For example, the 2019 Christchurch terrorist attack exposed a new risk with Facebook’s live streaming capability. Later in the year The 2019 Halle terrorist attack was live streamed using Twitch. The traditional use of Twitch (primarily used to screen share digital content while gaming), was significantly different to Facebook’s live streaming (primarily focused on live streaming real video) so this live streaming may not have been reasonably foreseeable despite the Christchurch incident. There needs to be processes to try to envisage risks to privacy, security, and safety before release, but also an ongoing process to address unforeseen risks once platforms are operating.

Companies should have:

- Reasonable reviews of privacy, security, and safety before products or new features are released
- Where applicable, compliance with published standards e.g. IEEE 7002-2022 (“IEEE Standard for Data Privacy Process”).[52]
    - There are a wide range of standards that companies can use and contribute to and they are at the cutting edge of technology, for example IEEE P7018 is developing a “Standard for Security and Trustworthiness Requirements in

[48] *GR Engineering Services Ltd v Investment Ltd* [2019] WASC 439, [65], [70].
[49] Ibid [68].
[50] *Red Sea Tankers Ltd v Papachristidis (The 'Hellespont Ardent')* [1997] 2 Lloyd's Rep 547, 587.
[51] *GR Engineering Services Ltd v Investment Ltd* [2019] WASC 439, [68].
[52] https://standards.ieee.org/ieee/7002/6898/

Generative Pretrained Artificial Intelligence (AI) Models".[53] New standards can be initiated where there are gaps, and this being international expertise together.

- A means for users to report risks to privacy, security, and safety risks and auditing to ensure reported risks are assessed.
- Crisis response process including:
    - The ability to declare and respond to a crisis with a surge in capacity and emergency protocols, including for communications and disclosure of the crisis.
        - These need to be focused on mitigating public harm, not corporate liability.
    - Sector collaboration as well as cooperation with governments and civil society in a crisis.
        - We work with GIFCT[54] (as does eSafety) which has evolved from a committee of the major technology companies into an independent pro-for-profit that coordinates responses to terrorist attacks and helps multiple companies detect and remove the proliferation of terrorist content after an attack.

29. Should the Act address risks raised by specific technologies or remain technology neutral? How would the introduction of a statutory duty of care or Safety by Design obligations change your response?

The Act should be platform neutral, but address different types of online engagement. Live streaming poses different risks to sharing private messages, social media posts are different to comments on posts by professional publishers like news media.

The concept of safety by design has been somewhat abridged in the issues paper. It includes designing systems to remove the potential for harm to occur, reduce the risk of harm occurring, and enable better mitigation when harm does occur. How this is implemented will be different for each platform, but some design patterns can be applied. For example, ensuring that all the content a user can edit can be reported. When we started in 2012 it was possible to edit a YouTube profile to include a username or profile description that e.g. incited violence, but it was only possible to flag a video (not a profile). This was a design flaw we had YouTube correct. When Facebook first introduced pages they could be used exactly the same as a user profile, so abuse happened in the name of a page which could then be abandoned without risking the users account. This undermined the real names policy and facilitated pseudonymous use of Facebook with no accountability. A design change we recommend led to pages either being treated more strictly if they remained anonymous, or owners could opt to list their names as the own of the page and take responsibility for it. In short, to enjoy freedom of expression someone must take responsibility for content. The mechanisms are platform specific, but the issues to be addressed and the design feature that addresses it can be designed in platform neutral terms.

[53] https://standards.ieee.org/ieee/7018/11306/
[54] https://gifct.org/

30. To what extent is the Act achieving its object of improving and promoting online safety for Australians?

It is promoting online safety in the limited areas of its schemes. Safer Internet Day is also a positive contribution. The Act has many aspirational goals which in practice fall outside the specific schemes and are therefore minimally realised.

31. What features of the Act are working well, or should be expanded?

The notice mechanism appears to be working well, for those harms where a notice can be issued. It should be expanded to cover other harms.

32. Does Australia have the appropriate governance structures in place to administer Australia's online safety laws?

We believe the culture of eSafety is too focused on Children's online safety. Fixing this would mean having multiple divisions to eSafety, with one division focused on children's safety and at least one that is broadly focused on everything else. This would prevent the children's online safety agenda, important as it is, leading to other issues being ignored or delayed.

A civil society advisory group focused on issues outside of children's online safety (and education in schools about children's online safety) is needed. Meta currently has such an advisory group in Australia that has functioned very effectively for some years.

See recommendation 5.

33. Should Australia consider introducing a cost recovery mechanism on online service providers for regulating online safety functions? If so, what could this look like?

As a fundamental principle, companies that produce online harms, which could be considered digital pollution, should not be able to outsource the job of cleaning up those harms (that they create and profit from) to the taxpayer. If their business is only viable by causing serious harm without sufficient mitigation or prevention, or by requiring government assistance to reduce the harm, then they are not viable companies. Where they are foreign corporations, which is usually the case, and they extract value from Australia (money or content) while requiring the Australian government to clean up behind them, Australia is being exploited in much the same way as irresponsible mining companies do when they go into remote locations and destroy the local environment and society by having no mitigation or restoration strategies. Australia has a duty to prevent this both by requiring harm prevention and mitigation by companies, at their expense, and by providing government oversight at the systems level.

A supervisory fee, such as that in the European Union, the UK, and proposed by Canada, may be appropriate to fund some of the supervisory functions of eSafety. This should in our view not cover the individual complaints system, but rather the systemic work and monitoring of compliance.

A per-incident based cost recovery model might also be possible where a person has reported a breach to a company, then required the support of eSafety before the company agrees to take action. The fees for such complaints should be based on actual costs (so would be very low per incident) and it should be explicitly illegal for companies to seek to recover these costs as additional fees to platform users. For convenience such fees might be due in aggregate each quarter. Where a platform seeks to challenge a takedown notice, there should be no fees with costs instead awarded by the court that heard the challenge if eSafety prevails.

A third funding requirement should be for civil society programs that mitigate online harms. This could be administered by eSafety with companies adding money to a funding pot, or could be provided directly by the companies, subject to eSafety supervision.It is becoming increasingly difficult for online safety information, and promotion of organisations working on online safety, to be seen online unless they pay for advertising. There is a perverseness to a company with a poor state of online safety profiting from advertising that aims to raise awareness and discourage poor behaviour. Such public service announcements, both the advertising cost and the production costs, should be at the platform's expense.

Recommendation 25: Cost recovery is appropriate for the systemic work of eSafety. A fee per report may be appropriate for individual items that have been reported to a platform by a user, which the platform has not acted on within reasonable time, and which eSafety then needs to investigate and address. A third component should be funding towards civil society work addressing online safety, ensuring this funding supports online safety generally and is not overly focused in one area (such as education in schools).

Summary of recommendations

Recommendation 1: The remit of eSafety should explicitly include the safety of individuals and groups within society that are impacted by online hate.

Recommendation 2: The power to issue notices and takedown orders should be extended beyond the current schemes into a general power that allow eSafety to act, on referral from a relevant government authority, in response to any content likely to be unlawful under Commonwealth, state, or territory laws. The existing scheme provides sufficient protections to allow such notices to be challenged through the courts where a company feels content is likely to be legal.

Recommendation 3: eSafety needs to have a wider and more regular engagement with civil society organisations working across different issues related to online safety.

Recommendation 4: eSafety grants need to cover a wider scope of online safety, not just the work related to the dedicated schemes eSafety manages.

Recommendation 5: eSafety should be restructured to add one or more Deputy Commissioners, and related support staff, who can maintain a focus on different areas of

online safety with at least one focus on general online safety beyond the current specific schemes.

Recommendation 6: An additional objective of the Online Safety Act should be to fulfil Australia's international human rights obligations, particularly in addressing Australian based or generated content that is causing or contributing to harms overseas.

Recommendation 7: The Online Safety Act should make reference to eSafety or the Minister maintaining a list of functionalities of online services and expectations in relation to them. The Act could then apply to any online services having any of those functionalities, rather than categories such as "search" and "social media" which may become less applicable as technologies change.

Recommendation 8: The Act should maintain a distinction between connectivity providers that respect net neutrality, which should have immunity in relation to content they carry, and services which relate to content and which should have obligations to mitigate or prevent harms from that content.

Recommendation 9: At least some Basic Online Safety Expectations should be enforceable and some of these should be expressed in a form that can be measured, with acceptable limits indicated. The measurements must be meaningful, appropriate, verifiable, and not prohibitively expensive to obtain.

Recommendation 10: To protect the public interest, civil society organisations, including community representative bodies, charities with an online safety or human rights objective, professional bodies with relevant expertise, and academics should have a means to provide feedback on industry codes and their appropriateness, or the need for changes to them.

Recommendation 11: Enforceable expectations should be set by government and exist outside of the terms of service of any one company, and should be consistent for different providers with respect to equivalent services.

Recommendation 12: Regulatory obligations should depend on risk, taking into account reach, functionality, and mitigations.

Recommendation 13: Grants from eSafety need to be more widely available to address a greater variety of online harms and need to support partnerships with civil society organisations working in the online safety space and addressing issues not covered, or less well covered, by eSafety itself.

Recommendation 14: The threshold for cyberbullying of an adult should be lowered to be similar to that of cyberbullying against a child. It should be appropriate to use eSafety to block defamatory provided there is no public interest in the content being public. If there is a potential public interest in the content being public, the issue should be decided through a defamation claim.

Recommendation 15: Different standards ought to apply to platforms that cater to different age ranges. The standards should be set by the Minister in the BOSE, while acceptable mitigations to meet those standards should be advised by eSafety.

Recommendation 16: A trusted flagger system for approved Civil Society Organisations should be added to the legislation. It should be eSafety rather than platforms that determine who qualifies as a trusted flagged.

Recommendation 17: Anonymous reporting to eSafety should be possible for some kinds of online safety violations and the anonymity of users of such services should be protected by the Act.

Recommendation 18: A description of proscribed abhorrent violent content should be maintained either with the Classifications Board or by eSafety, or both, in order to facilitate voluntary removal (potential on advice from online safety focused civil society organisations) by companies who find they are hosting such content.

Recommendation 19: eSafety should be given the ability to level small fines to users who engage in harmful online conduct, with larger fines for repeated or more serious offences where criminal proceedings are not being pursued.

Recommendation 20: Penalties should be increased to be in-line with those overseas.

Recommendation 21: Civil society organisations working in online safety should, one approved by eSafety, have a legal right to scrape data and an exception from the Privacy Act, subject to certain obligations.

Recommendation 22: Platforms should be required to provide URLs that directly link to any user generated content that may be used to breach online safety.

Recommendation 23: Metrics to assess compliance with preventing online harms should measure the level / frequency in which harm occurs, the response time in addressing reported harms, and the effectiveness of the response to reports of harm. Thresholds for acceptable values of reach of these metrics should be set by the government and routinely reviewed and adjusted to ensure there is a continual improvement in online safety.

Recommendation 24: There should be an Australian monitoring effort similar to that run by the European Commission, with the involvement of appropriate civil society organisations.

Recommendation 25: Cost recovery is appropriate for the systemic work of eSafety. A fee per report may be appropriate for individual items that have been reported to a platform by a user, which the platform has not acted on within reasonable time, and which eSafety then needs to investigate and address. A third component should be funding towards civil society work addressing online safety, ensuring this funding supports online safety generally and is not overly focused in one area (such as education in schools).