THE ONLINE HATE PREVENTION INSTITUTE

Empowering communities, organisations and agencies in the fight against hate.

ONLINE HATE PREVENTION INSTITUTE SUBMISSION

CONSULTATION ON THE ONLINE SAFETY (BASIC ONLINE SAFETY EXPECTATIONS) DETERMINATION 2021

SUBMISSION TO THE DEPARTMENT OF INFRASTRUCTURE, TRANSPORT, REGIONAL DEVELOPMENT AND COMMUNICATIONS

8 OCTOBER 2021

[redacted]

ABOUT THE ONLINE HATE PREVENTION INSTITUTE

The Online Hate Prevention Institute (OHPI) is Australia's only harm prevention charity dedicated to tackling online hate and extremism. It has been doing so since January 2012. We thank the Government for the opportunity to provide this submission.

Since our establishment in 2012, we have worked with various parts of Government to enhance online safety. We have worked with, *inter alia*, the Australian Federal Police ("AFP") and through them the Australian Intelligence Community, the Attorney General's Department ("AGD"), the Department of Foreign Affairs and Trade ("DFAT"), the Department of Veterans Affairs, the Australian Human Rights Commission ("AHRC"), and various state government departments and police forces.

OHPI has been involved in the consultations which led to the current online safety system, including the consultation on the Coalition's 2012 *Enhancing Online Safety for Children Discussion Paper* and the Government's 2014 *Enhancing Online Safety for Children* consultation, and the 2020 and 2021 departmental and parliamentary consultations on the *New Online Safety Act.*

We have engaged directly with eSafety, and regularly find ourselves at the same meetings and conferences, both within Australia and internationally, as eSafety staff, representatives from AFP, DFAT, and AGD. We are often the only Australian civil society group at these meetings, usually invited by overseas organisers who recognise us as a global leader in this space. It is a matter of regret that local organisers seldom engage sufficiently with civil society.

Our focus on online hate and extremism covers hate against individuals (e.g. cyberbullying, serious trolling, RIP trolling etc), hate against specific groups within society (e.g. not only antisemitism, Islamophobia, homophobia, misogyny, racism, transphobia but also serious attacks on other groups such as ANZAC veterans, First Nations' people, politicians, etc.), and hate targeting our society as a whole; for example, white supremacy and other form of violent extremism. In recent years, our focus on combating violent extremism has increased as online radicalisation has grown. We have been actively involved in removing terrorist manifestos and abhorrent violent content videos, and we monitor social media for threats.

ABOUT THE AUTHORS

DR ANDRE OBOLER

Dr Andre Oboler is the CEO & Managing Director of the Online Hate Prevention Institute. He is an Honorary Associate at La Trobe Law School, a global Vice President of the IEEE Computer Society, Vice Chair of the Global Public Policy Committee of the IEEE, an expert member of the Australian Government's Delegation to the International Holocaust Remembrance Alliance, and a consultant on online hate and extremism for the American Jewish Congress.

Andre was formerly a Senior Lecturer in Cyber Security at the La Trobe Law School, intercultural liaison for the Victorian Education Department's independent inquiry into antisemitism, co-chair of the Online Antisemitism working group of the Global Forum to Combat Antisemitism, an expert member of the Inter-Parliamentary Coalition to Combatting Antisemitism and served for two terms with the board of the UK's higher education regulator the QAA. His research interests include online regulation, hate speech and extremism in social media, and the impacts of technology on society.

He holds a PhD in Computer Science from Lancaster University, and a B. Comp. Sci. (Hons) & LLM(Juris Doctor) from Monash University. He is a Senior Member of the IEEE, a Graduate Member of the Australian Institute of Company Directors, and a Member of the Victorian Society of Computers & Law.

PROF. DAVID WISHART

Dr David Wishart is a Director of the Online Hate Prevention Institute and an Adjunct Professor with the Law School at La Trobe University. He recently completed a term as Acting Dean of the Law School. His expertise includes Competition Policy, Corporations Law, Constitutional Law, Electronic Evidence, the law as to citizenship, and issues relating to law and Indigenous peoples. His *Curriculum Vitae* includes more than 50 refereed articles and a number of self-authored books.

He holds a Bachelor of Commerce, LLB(Hons), and an LLM from Melbourne University, and a PhD from the Australian National University.

ALETTE DE KOKER

Alette de Koker is an intern at the Online Hate Prevention Institute. She is a penultimate year Law (Hons.) and International Relations student at La Trobe University. She has previously interned at the Victorian Parliament and has served as a consultant for a leading Australian online dispute resolution lawtech company. Her research covers a range of international law and human rights projects. She is currently assisting the Online Hate Prevention Institute in the areas of law reform, countering extremism, and combating anti-Asian hate in social media.

EXECUTIVE SUMMARY

We welcome the passage of the *Online Safety Act 2021* (Cth) and this consultation opportunity to the draft *Online Safety (Basic Online Safety Expectations) Determination 2021*. We note that the core Basic Online Safety Expectations are set by the legislation and cannot be amended in the *determination* but that the Department is seeking feedback on:

1. A set of additional expectations for social media services, relevant electronic services and designated internet services
2. Reasonable steps that could be taken to meet certain expectations

We believe the Government has correctly set the roles of Government, Social Media platforms, and members of the public but that there is a missing fourth critical stakeholder in online safety. Civil society should be part of the solution, as we have expressed in previous submissions on online safety. The Government has an opportunity, through the *Online Safety Determination,* to ensure basic standard are met to enable Civil Society to do its part to enhance online safety.

This submission outlines how civil society should be included as a critical stakeholder in online safety. It also sets out the way the various expectations could be strengthened which, in the experience of the Online Hate Prevention Institute, would be beneficial to achieving the control of online hate.

EXPECTATIONS REGARDING DEALINGS WITH AUSTRALIAN CIVIL SOCIETY (NEW DIVISION)

There are limits to a regulator's capacities and government cannot sufficiently resource a regulator to fully meet the community's need. Civil society can assist in bridging this gap and has an essential role to play. To enable it to do so, the work of civil society, like many other areas, needs to be supported by government. Of most immediate concern is the need to ensure that relevant civil society organisations have points of contact with the technology companies so they can raise issues that the usual processes are failing to address.

We understand the burden that regular points of contact could potentially place on service providers. To limit that burden, we recommend the number of groups able to utilize these points of contact be limited. The existing Federal Register of Harm Prevention Charities, which is strictly controlled, provided a pre-vetted list of charities working to prevent harm to human beings. We suggest reusing this list, plus additional charities authorised by eSafety.

Definition:

(1) Relevant civil society organisations, a list of which will be made public by eSafety, will include those that are:

(a) charities on the federal register of harm prevention charities; or

(b) charities recognised by the eSafety Commissioner as representing a segment of the community that faces particular online safety risks.

Additional expectations—provider will be able to be contacted by civil society

(2) The provider of the service will ensure there are employees or agents of the provider who are designated as the service's liaisons with Australian civil society (the liaison).

(3) Relevant civil society organisations will have a means of contacting and or meeting with the service's liaisons to raise online safety concerns that are not being sufficiently addressed through other processes.

Additional expectations—provider will consult with civil society in Australia

(4) Where the provider intends carrying out significant changes that impact online safety in Australia and:

(a) The change is specific to Australia or a specific group of countries including Australia, or

(b) The change is global, and the provider intends consulting civil society overseas on the changes,

The provider shall provide a consultation process for relevant civil society organisations in Australia.

EXPECTATIONS REGARDING SAFE USE (DIVISION 2)

S 6(2) The provider of the service will take reasonable steps to proactively minimise the extent to which material or activity on the service is or may be unlawful or harmful.

We welcome the inclusion of S 6(2) but recommend:

- Rewording the provision and separating the provision about unlawful content from the provision about harmful content.
- Aiming for the removal, rather than the minimisation, of unlawful content / cessation on unlawful conduct. Provisions which permit unlawful content / conduct by advocating a response less than following the law would run counter to public policy.
- Aiming for the minimisation of harm in the case of harmful content. Removal is one – but not the only – tool to achieve this, and other approaches include limiting who can see the content. Some of these options include: (a) optional safety filters to screen out certain content, (b) forced limitations so content is only viewable to the poster or to them and their connections, (c) preventing posts being advertised, (d) limiting or preventing content being served to those not directly looking for it, (e) preventing the sharing of content, (f) preventing content showing in search results, etc. The public policy considerations are significantly different where the content is lawful, even if it is harmful, as opposed to where the content is unlawful.

- Being explicit that unlawful content includes content which is unlawful under Commonwealth, State or Territory laws and able to be seen by people in the relevant jurisdiction

As currently worded, with respect to unlawful content, the provision can be read as urging platforms to make such content lawful, rather than to act on the content – which is presumably not the intention. When applied to harmful content, this provision encourages platforms to make the content less harmful – which is appropriate and welcome.

We recommend replacing this expectation with:

S 6(2) The provider of the service will take reasonable steps to proactively

(a) minimise the presence of material or activity on the service which is unlawful under a law of the Commonwealth, a State or a Territory and is visible to people in that jurisdiction.

(b) minimise the extent to which material or activity on the service is harmful.

S 6(3)

We welcome the provisions as listed in S 6(3) but recommend extending them.

The reasonable steps in the case of our proposed S 6(2)(a) (as outlined above) are covered by the existing S 6(3)(a). Additional reasonable steps which could cover our proposed S6(2)(b) might include:

S 6(3)(f) In the case of harmful activity, limiting the content's visibility to those opting-in to seeing this user's content (e.g. friends, followers, connections, or subscribers depending on the platforms), preventing the content appearing in search functions, hiding the content from children and users who turn on safety filters.

Engaging with external assistance, particularly from civil society, could also be added to the reasonable steps listed in S 6(3). The Online Hate Prevention Institute is the leading Australian charity dedicated to this space and we invite the government to recognise our role, and to recommend platforms work with us to identify and address gaps in their technology, processes, and training. We propose the following:

S 6(3)(g) collaborating with relevant civil society groups (such as the Online Hate Prevention Institute) to:

a) review the effectiveness of processes to detect, moderate, report and remove (as applicable) material or activity on the service that is or may be unlawful or harmful, and;

b) to solicit recommendations for improvements, and the identification of areas of concern.

S 9(1) If the service permits the use of anonymous accounts, the provider of the service will take reasonable steps to prevent those accounts being used to deal with material, or for activity, that is or may be unlawful or harmful.

S 9(1) may be superfluous if the field is covered by S 6(2). As currently worded, the two provisions means only anonymously posted unlawful content would require removal (under S 9(1)), highlighting the weakness identified above regarding S 6(2). At the same time, S 9(1) may also require the removal (rather than harm minimisation) of harmful but lawful content / conduct (when it occurs anonymously). This poses a significant interference with free speech and may lead to negative consequences for whistle blowers, political activists, members of oppressed minorities, and those acting in the public interest to hold extremists (such as neo-Nazis) accountable.

One example of legal, but harmful, online speech would be the promotion of Covid misinformation. This misinformation is very harmful as it can encourage risky behaviours such as non-compliance with public health directions and increase vaccine hesitancy among the Australian population. Under one reading of S 6(2), a platform might be compliant if they allowed the material to remain online, but seriously limited its spread by not sharing in the feeds of other users / recommending it. If the poster was anonymous, however, S 9(1) would not be satisfied unless the material was removed. One could take it a step further and argue that even a post that simply says "people should not get vaccinated" would need to be removed under S 9(1). To be clear, this is only to provide an example of how the provision may be interpreted. The Online Hate Prevention Institute fully supports the COVID-19 vaccination program and we have worked to combat certain COVID misinformation. S 9(1), however, appears to be too broad.

A second example of legal, but harmful, online speech would be activists countering extremism. Such activists often anonymously expose the extremist activity they monitor. This may well cause harm to the extremist by holding them accountable in public for what they have said or done. We don't take this approach ourselves, as we prefer to share the information with the relevant platforms, agencies, and police, but we understand and support the rights of those who do take such action. Shutting them down due to an expectation under S9(1) would permit a greater harm.

We do not believe that speech that would otherwise be permitted should be prohibited simply because it is posted anonymously. There are often valid reasons for anonymous speech. Limiting this provision to material, or for activity, that is or may be unlawful would be an improvement, but if doing that it should apply to all unlawful content regardless of whether it is posted anonymously. We note in the case of harmful content, a "minimise" requirement (as under S 6(1)) leaves more room for discretion in particular cases.

Some special measurers in the case of anonymous or pseudonymous speech are justified, provided they take account of all the circumstances. We recommend:

S 9(1) If the service permits the use of anonymous accounts, or pseudonymous account that cannot be traced to a real person, the provider of the service will take reasonable steps to:

(a) protect the identity of those users, disclosing their identity only as authorised by law or as necessary to ensure the person's safety*

(b) prevent these accounts from being used in an unlawful manner

(c) prevent these accounts being used in a manner which, while lawful, is both harmful and against the public interest (having due regard to the importance of free speech)

* A special provision may be need to allow disclosure to parents, guardians, and school authorities in the case the anonymous user is a child. A similar provision may be needed in the case vulnerable adults.

EXPECTATIONS REGARDING CERTAIN MATERIAL AND ACTIVITY (DIVISION 3)

We recommend the inclusion of a section 11(2) that outlines "*Reasonable steps that could be taken*".

S 11(2) Reasonable Steps that could be taken to minimise the provision of material include:

(a) Review by staff, in reasonable time, of reports made about this content

(b) The use of Artificial Intelligence to remove such material

(c) The use of Artificial Intelligence to hide it such material pending a manual review

(d) Cooperation with all levels of government, and relevant civil society organisations, to facilitate rapid action on material identified by experts

We recommend including an addition provision regarding unlawful content.

Additional Expectation – provider will take reasonable steps to minimise provision of certain material and activities

The provider of the service will take reasonable steps to minimise the extent to which the following material is provided on the service, or activity occurs on the service:

(1) Material or conduct which is unlaw under a law of the Commonwealth, a State, or a Territory

(2) Material which incites unlawful conduct or the posting of unlawful material

(3) Material which attack or incite hate against a person or group of people on the basis of race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, serious disease, disability, asylum seeker / refugee status, or age.

Additional Expectation – provider will consult the International Holocaust Remembrance Alliance's Working Definition of Antisemitism when seeking to identify antisemitic content

Antisemitism is a form of racism and religious vilification which targets the Jewish community. In identifying what qualifies as antisemitism, providers shall have regard to the International Holocaust Remembrance Alliance's Working Definition of Antisemitism, including the need it expresses to consider all the circumstances.

EXPECTATIONS REGARDING REPORTS AND COMPLAINTS (DIVISION 4)

We have no additional recommendations for this section.

EXPECTATIONS REGARDING MAKING CERTAIN INFORMATION ACCESSIBLE (DIVISION 5)

We recommend an additional requirement to ensure the material which is made available is also archived in a manner that is available.

S 17(1)(e) archived in a publicly accessible location with details of when each version was in force

This may be necessary both to identify changes over time and to understand what the requirements were at a given point in time.

EXPECTATIONS REGARDING RECORD KEEPING (DIVISION 6)

We have no additional recommendations for this section.

EXPECTATIONS REGARDING DEALINGS WITH THE COMMISSIONER (DIVISION 7)

Section 46(1)(f) of the Act requires that platforms provide a mechanism to "enable end-users to report, and make complaints about, breaches of the service's terms of use". It is unclear if every "report" is to be treated as a "complaint". It seems possible that a valid mechanism could register reports, but only elevate them to the status of a complaint if the reporting user remains dissatisfied after the platform has fully investigated their report. With such a mechanism in place, S 20(1) of the determination would be somewhat defeated as the time and effort to pursue a matter to the point of it becoming a complaint may greatly inhibit the number of reports that turn into disclosable complaints. On the other hand, if one company treats all reports as complaints, and another company does not, the data cannot be readily compared.

As this is a core expectation and cannot be changed, we recommend adding an additional expectation to allow the Commissioner to request the number of *reports*. Further, we believe it would be useful to *disaggregate* these reports for the Commissioner can see specifically where the concerns are in a given platform and effectively ascertain how well the platform is responding.

S 20(4) If the Commissioner, by written notice given to the provider of the service, requests the provider to give the Commissioner a statement that sets out the number of reports (either in total, or of a particular type) made to the provider during a specified period (not shorter than 6 months) about breaches of the service's terms of use, the provider will comply with the request within 30 days after the notice of request is given.

An additional limitation which could be considered is to limit the requests to breaches of the terms of use which would correlate with content or conduct which is unlawful under a Commonwealth, State or Territory law.

Alternative S 20(4) If the Commissioner, by written notice given to the provider of the service, requests the provider to give the Commissioner a statement that sets out the number of reports (either in total, or of a particular type) made to the provider during a specified period (not shorter than 6 months) about breaches of the service's terms of use (that correlates to material or conduct which is unlawful under a Commonwealth, State or Territory law), the provider will comply with the request within 30 days after the notice of request is given.

We also note the requirement for a reporting mechanism for nine types of unlawful activity under Section 46(1)(e) of the Act could lead to two paths of reporting, one of which (building on the argument above) might lead to complaints being avoided. This could lead to reports which fall outside of S20(1) / our S20(4) (as they are not reports / complaints of a breach of the terms of service, but rather for breaches of Australian law) and outside of the Determination's S 20(2) as the platform may deal with the matter (or not) without it triggering a removal notice. This data is vital to understanding what share of complaints are dealt with without a removal notice, what shared remain online and never progress to getting a removal notice, what are removed after a removal notice, and how many (if any) remain online despite a removal notice.

S 20(5) If the Commissioner, by written notice given to the provider of the service, requests the provider to give the Commissioner a statement that sets out the number of reports made to the provider during a specified period (not shorter than 6 months) for one or more type of complaint covered in S46(1)(e) of the Act (these being, (i) cyber-bullying material targeted at an Australian child; (ii) cyber-abuse material targeted at an Australian adult; (iii) a non-consensual intimate image of a person; (iv) class 1 material; (v) class 2 material; (vi) material that promotes abhorrent violent conduct; (vii) material that incites abhorrent violent conduct; (viii) material that instructs in abhorrent violent conduct; and (ix) material that depicts abhorrent violent conduct), the provider will comply with the request within 30 days after the notice of request is given.

This would make it is possible for the Commissioner to learn which of the nine categories are causing particularly high levels of reporting on particular platforms.