Australian Government | eSafetyCommissioner

eSafety submission

Strengthening Victoria's laws against hate speech and hate conduct

October 2023

About eSafety

The eSafety Commissioner (eSafety) welcomes the opportunity to provide feedback on the proposed reforms to strengthen Victoria's anti-vilification protections.

eSafety is Australia's national independent regulator for online safety. Our purpose is to help safeguard Australians from online harms and to promote safer, more positive online experiences.

We recognise that online vilification, and online hate more broadly, can negatively impact a person's mental health, general wellbeing, and online engagement. It can also lead to harassment and physical violence offline, and contribute to broader social harms.

eSafety's enabling legislation, the *Online Safety Act 2021 (Cth)* (OSA), does not define online hate or vilification, or provide eSafety with powers specifically focussed on these terms. However, as this submission explains, there are intersections between the online content that may amount to vilification under relevant state and territory laws and the types of content that individuals can report to eSafety for investigation and removal under the OSA.

In addition, eSafety has used its reporting powers to require service providers to explain how they are enforcing their rules against online hate to keep their users safe. eSafety is also working to better understand and address online hate through a variety of research and other initiatives aimed at protecting voices at risk online.

This submission draws on the lessons eSafety has learned through this work to inform consideration of Victoria's proposals for anti-vilification reform.

Australians' experiences and the scope of the proposals

Online hate can broadly be described as any kind of online communication that attacks, discriminates, insults or uses hateful language against a person or group based on their race, religion, ethnicity, gender, sexual orientation, or disability. It is typically targeted towards members of minority groups, and is intersectional, meaning that multiple attributes of a person's identity can compound the severity of the harm they experience.

eSafety research has found:

- One in ten children have been the target of online hate (Mind the Gap, 2022).
- First Nations adults and the LGBTIQ+ community experience online hate at twice the rate of the general population (Protecting voices at risk online report, 2020).
- Aboriginal and Torres Strait Islander children are almost three times more likely to have experienced online hate than the national average. Around 3 in 10 Aboriginal and Torres Strait Islander children (29%) have had offensive things said to them because of their race, ethnicity, gender, nationality, sexual orientation, religion, age or disability, compared to the national average of 1 in 10 (11%) (First Nations report, 2023).
- Children with disability are more likely to have been the target of online hate than the national average. One in six children with disability have been the subject of online hate (16% compared with the national average of 11%). Among teens with disability (aged 14-17), the likelihood of having experienced online hate increases to just under one in four

(23%), compared with the national average of 14% for the same age cohort (Forthcoming report: 2023).

- LGBTIQ+ teens (aged 14-17) are more than twice as likely to have experienced online hate than the national average. Almost a third of LGBTIQ+ teens have been targeted by online hate (31%) in the past year, double the national average figure of 15% for the 14-17 age category (Forthcoming report: 2024).

In September 2023, eSafety met with representatives from the Victorian Equal Opportunity and Human Rights Commission (VEOHRC) in relation to this research, to investigate opportunities for collaboration between the two organisations.

We note the current *Racial and Religious Tolerance Act 2001 (Vic)* defines vilification as behaviour that incites hatred, serious contempt, revulsion or severe ridicule for a person or group of people because of their race or religion. We understand this definition is under review, and may be expanded to cover vilification based on additional attributes, such as gender and sex, sexual orientation, gender identity and/or expression, sex characteristics and/or intersex status, disability, HIV/AIDS status, and personal association. In line with our research, eSafety supports the expansion of existing definitions to reflect the types of intersectional online hate that Australians are experiencing and to provide protections commensurate to its impacts.

Intersections between eSafety's role and vilification

Resources for those at greater risk of online harms

Building on our research, eSafety works closely with Australian communities most at risk of online hate and other forms of abuse to understand their online experiences, to inform our policies, and to co-design meaningful resources and support so that they can engage confidently online. This includes First Nations people, individuals living with disability, and those from culturally and linguistically diverse (CALD) and LGBTIQ+ communities.

Our resources provide advice on identifying online abuse, ways to prevent further contact from those perpetrating abuse, skills and strategies to deal with online abuse, and how to report different types of abuse to eSafety or other relevant services.

Complaints-based regulatory schemes

eSafety also administers complaints-based regulatory schemes to address:

- cyber abuse material targeted at Australian adults (**adult cyber abuse**)
- cyberbullying material targeted at Australian children (**cyberbullying**)
- the non-consensual sharing of, or threat to share, intimate images (**image-based abuse**)
- illegal and restricted content.

Where the online material reported to us meets the legislated thresholds set out below, eSafety can issue removal notices to the online service on which the content is posted and the hosting service that hosts the content on the service. For adult cyber abuse, cyberbullying and image-based abuse, eSafety can also issue notices to the end-user responsible for posting the material

and require them to remove it. eSafety also has a range of other options available, as set out in our Compliance and Enforcement Policy.

In addition to formal action, we work with online service providers and others to achieve positive outcomes for victims and survivors of online harm. This support may involve facilitating rapid removal of abusive content posted online; referrals to law enforcement, mental health providers or legal services; and providing tips and strategies for how to mitigate further harm.

While the OSA does not expressly provide eSafety with a responsibility to address online hate or vilification, some material which meets Victoria's current definition of vilification (or a future definition following the review) may also meet the legislated threshold under one of our complaints-based schemes.

Adult cyber abuse

The adult cyber abuse scheme covers online material that an ordinary reasonable person would conclude was likely intended to have the effect of causing serious harm to a particular Australian adult (aged 18+), and which an ordinary reasonable person in the position of the Australian adult would regard as being menacing, harassing or offensive in the circumstances. eSafety cannot take formal action against content targeting a group or organisation.

The adult cyber abuse scheme is a safety net to be used when a complaint has been made to an online service provider, but the online service provider has not removed the material. The material must be provided on a social media service, relevant electronic service or designated internet service and can include posts, comments, memes, images or videos.

The scheme is not limited to any specific personal attribute, including abuse relating to an individual's race and religious beliefs and activities. However, hateful elements within a post (for example, the use of ethnic slurs) may be considered as a relevant factor in determining whether the content meets the threshold of adult cyber abuse.

The adult cyber abuse scheme enables eSafety to give formal notices to the service, end-user or hosting service provider to remove the content if it breaches the OSA. In some cases, we may refer a complaint to the online service for consideration if we believe the content is likely to violate that service's terms of service.

Child cyberbullying

eSafety's cyberbullying scheme covers online material that an ordinary reasonable person would conclude was likely intended to have the effect of seriously threatening, seriously intimidating, seriously harassing or seriously humiliating a particular Australian child or young person (under 18). Like the adult cyber abuse scheme, the cyberbullying scheme is a safety net for when a complaint has been made to an online service provider, but the service has not removed the material.

The cyberbullying scheme captures a broader set of online harms which can include online hate or vilification but is not limited to abuse referencing a specific set of personal attributes. It is restricted to behaviour directed towards an individual and does not capture cyberbullying directed towards a group of people.

The cyberbullying scheme also enables eSafety to give end-user notices requiring the end-user to refrain from posting any further cyberbullying material and/or apologise to the child who was the target, in addition to removing to the cyberbullying material.

Image-based abuse

eSafety operates an image-based abuse scheme for Australians of all ages whose intimate images or videos have been (or are threatened to be) shared without their consent. This scheme is also available to people based overseas where an Australian person shares or threatens to share intimate images of them. The image or video can be real, altered or faked to look like a particular individual, or shared in a way that makes people think it is that individual.

This scheme recognises that the non-consensual sharing of intimate images can happen in a variety of contexts and involve different perpetrator motivations. For example, sometimes image-based abuse involves images depicting a person without attire of religious or cultural significance.

The image-based abuse scheme also includes remedial direction powers which allow eSafety to require an end-user who has posted, or threatened to post, an intimate image of another person without their consent, to take specific action aimed at preventing the non-consensual sharing (or further sharing) of intimate images. A remedial direction may be used in conjunction with a removal notice to an online service provider.

Illegal and restricted content

The OSA empowers eSafety to investigate complaints about class 1 and class 2 material that is available to Australians.

Class 1 material includes material that would be refused classification under the National Classification Code. This includes online content that promotes, incites or instructs in matters of crime or violence. eSafety has a graduated series of options to address this type of content, including removal notices. In addition, eSafety recently registered six industry-developed codes to address class 1 material across social media services, app distribution services, hosting services, internet carriage services, equipment, and search engine services.

A subset of class 1 material is material that depicts, promotes, incites, or instructs in 'abhorrent violent conduct' such as murder, torture, rape or violent kidnapping. eSafety has the power to request or require internet service providers to block such material for a limited time if its availability is likely to cause significant harm to the Australian community. The intent of this power is to prevent the rapid distribution of material online, as occurred, for example, after the 2019 terrorist attacks in Christchurch, New Zealand.

Class 2 material includes material that would be classified R18+ or X18+ under the National Classification Code. eSafety can give removal notices in relation to class 2 material to the online service where the content is posted and the hosting service that hosts the content on the service, if the service is provided from Australia. eSafety can also give remedial notices to the online service or the hosting service provider of the service, requiring the recipient to either remove the material or use a restricted access system to prevent children under the age of 18 from accessing it.

Lessons from our operational experience

The Department of Justice and Community Safety (DJCS) notes that in relation to remedies for online vilification specifically, it might be appropriate to clarify that the Victorian Civil and Administrative Tribunal (VCAT) may order a person (against whom a successful claim is made) to take down online material, as this is not stated in the existing remedies provision.

In eSafety's experience facilitating the removal of harmful material, there are several benefits of engaging with online services in the first instance.

- First, our well-established relationships and escalation pathways can often provide the quickest resolution and relief to the targeted person.
- Second, online services are likely to have insight into whether the relevant end-user may be involved in multiple violations. If they are made aware of the issue, they can determine whether further measures beyond removing a particular item of content may be warranted, for example, suspending the person's account.
- Third, the identity, location, and contact details of the responsible end-user are often unclear. The end-user may have posted anonymously or pseudonymously, and/or they may be located in another jurisdiction. eSafety has powers to obtain end-user information from online services where relevant to an investigation. However, this can entail a delay. Further, the nature, extent, and quality of information collected and stored by online services varies. Many online services hold insufficient information to proceed with an end-user removal notice despite this power.

We note that proposed reform 5 would enable the VEOHRC to request information to help people identify who has vilified them. While this could assist in resolving some cases where services do hold sufficient information to identify the relevant end-user, we would encourage caution around the approach of providing this information directly to a complainant. In most cases, privacy and confidentiality considerations prevent eSafety from sharing any identifying information we may obtain through the course of an investigation with others, including the relevant complainant. We believe this is an important protection against the risk of retaliation.

Another important lesson from our consultation and horizon scanning activities on emerging tech trends and challenges is that content removal powers may be of limited utility in immersive online environments, where abuse occurs in real time. Abuse and hate in these environments can cause people lasting harm, but does not necessarily involve lasting content that must be removed. Accordingly, a different set of responses may be needed.

Finally, eSafety's experience shows that complaints mechanisms are an extremely important safeguard for those who have experienced harm, but to truly move the dial, they must be accompanied by more proactive, systemic change across online services.

Powers to promote proactive and systemic change

In addition to our complaints-based schemes for individuals, eSafety promotes proactive and preventative changes with industry to enhance their systems and process to better address the risks of online harms.

Proposed Reform 8 would extend VEOHRC's powers to respond to systemic vilification and investigate anti-vilification matters. While not specific to online vilification, these proposed powers could overlap with eSafety's proactive and systemic change activities, including the Basic Online Safety Expectations and Safety by Design.

Basic Online Safety Expectations

The Basic Online Safety Expectations are designed to improve online service providers' safety standards, transparency and accountability. Services are required to have terms of use, policies and procedures to ensure the safety of users. Services are also required to take steps to ensure that penalties for breaches of their terms of use are enforced against all accounts created by the end-user.

With regard to online hate, the *Online Safety (Basic Online Safety Expectations) Determination 2022* (the Determination) establishes expectations that service providers will:

- take reasonable steps to ensure that end-users are able to use the service in a safe manner (s6(1)) and to proactively minimise the extent to which material or actively on the service is unlawful or harmful (s6(2)).
- ensure they have terms of use, policies, and procedures in relation to the safety of end-users, as well as policies and procedures for dealing with reports and complaints (s14).
- take reasonable steps to ensure that penalties for breaches of their terms of use are enforced against all accounts held by those end-users (s14(2)).

Additionally, the Explanatory Statement to the Determination states that services should use their terms, policies and procedures to address harmful material that is not necessarily unlawful or explicitly referenced in the OSA, such as online hate against a person or group of people on the basis of race, ethnicity, disability, religious affiliation, caste, sexual orientation, sex, gender identity, serious disease, disability, asylum seeker/refugee status, or age.

The OSA provides eSafety with powers to require online service providers to report on the reasonable steps they are taking to comply with any or all of the Basic Online Safety Expectations. When deciding which providers to give a notice to, the OSA requires eSafety to have regard to specified criteria. These criteria and other considerations are summarised in the BOSE Regulatory Guidance. The obligation for services to respond to a reporting requirement is enforceable and backed by civil penalties and other enforcement mechanisms. Information obtained from the reporting notices are published in transparency summaries where appropriate.

On 21 June 2023, eSafety issued a non-periodic reporting notice to Twitter (subsequently X), requiring it to explain what it is doing to minimise online hate, including how it is enforcing its terms of use and hateful conduct policy.

eSafety intends to publish information received in response to the notice to improve transparency and accountability, and would be happy to provide this information to DJCS at the appropriate time.

eSafety will continue to issue further notices on a broader range of harms to online service providers in the coming months.

Safety by Design

eSafety also aims to produce positive outcomes for Australians by guiding and supporting the online industry to enhance safety measures through our Safety by Design initiative.

Safety by Design encourages industry to anticipate potential harms and implement risk-mitigating and transparency measures throughout the design, development, and deployment of a product or service. This approach seeks to minimise any existing and emerging harms that may occur, rather than retrospectively addressing harms after they occur.

The initiative promotes online safety through three guiding principles:

1. **Service provider responsibility**: The burden of safety should never fall solely upon the user. Every attempt must be made to ensure that online harms are understood, assessed, and addressed in the design and provision of online platforms and services.
2. **User empowerment and autonomy**: The dignity of users is of central importance. Products and services should align with the best interests of users.
3. **Transparency and accountability**: Transparency and accountability are hallmarks of a robust approach to safety. They not only provide assurances that platforms and services are operating according to their published safety objectives, but also assist in educating and empowering users about steps they can take to address safety concerns.

A Safety by Design approach can seek to address online hate by promoting a proactive approach to user safety that includes measures such as:

- having individuals or teams accountable for online hate and vilification policies
- putting tools and processes in place for detecting and removing online hate
- ensuring that community guidelines and processes about online hate are accessible and easy to understand
- carrying out open engagement with a wide user base including independent experts and key stakeholders, on the development, interpretation and application of online hate standards and their effectiveness or appropriateness
- committing to consistently innovate and invest in safety-enhancing technologies
- publishing information about safety tools, policies, and processes, and their impact and effectiveness
- ensuring that design features and functionality preserve fundamental user and human rights.

Practical resources are provided via the Safety by Design assessment tools, including educative content on intersectional risk factors, insights into perpetrator motives, and exploration of human rights in the digital context.

We have focused our Safety by Design work on diverse, marginalised, and at-risk groups to make sure their needs are effectively considered, incorporated, and actioned in the design of online

products and services. We also consider that education and empowering people will always form the basis for addressing the social and behavioural issues that manifest online.

Clarifying powers to remove online material

eSafety would welcome further information about VCAT's current and potential future powers and processes for ordering removal of online vilification and how this works in practice.

For example, it is not clear if the powers are limited to orders against end-users responsible for posting online vilification, or if this extends to orders against service providers on whose platforms the content is posted. As outlined above, in our experience, engaging with service providers can facilitate quicker removal of the harmful material. There are also difficulties when exercising powers on end-users, including difficulty in confirming who the end-user is (particularly if a pseudonymous or anonymous account) and where they are located.

We also welcome information about how VCAT enforces removal orders, what the outcome is for individuals or online service providers who do not comply, and VCAT's access to the relationships and communication channels with online services it might need to gather necessary information relating to offending material, such as the extent of its spread across a service, or details of the end-user's account(s). This can be vital information when contextualising a dispute and deciding whether it is appropriate to exercise take down powers.

eSafety would welcome referrals of complaints about online content that is likely to meet our legislated thresholds, to ensure complainants have access to all options available to them. In relation to adult cyber abuse, cyberbullying and image-based abuse, eSafety can investigate complaints made by the person targeted by or depicted in the material, by a person authorised to make the complaint on behalf of the targeted person, or by other persons such as parents/guardians where the material targets a child. As such, for eSafety to act on referred matters, the targeted person would need to authorise the referring person to lodge a complaint with eSafety (or make the complaint directly themselves).

We strongly recognise the importance of clear communication with the public about where they can go for help with online issues, including online hate. This includes having minimal overlapping avenues for support and clear expectations about what is involved in different complaints processes and outcomes. Where there is an overlap between online hate or vilification and a type of harm within eSafety's remit, it may be easier and faster to utilise eSafety's complaints-based scheme. Recognising that being the recipient of online hate can be extremely distressing and having multiple escalation paths could prove to be confusing, coordinating efforts wherever possible may be in the best interests of the target.

Review of the Online Safety Act 2021

The Australian Government is expected to conduct an independent review of the OSA throughout 2024. eSafety will be closely working with the Australian Government to ensure the OSA remains fit for purpose and adequately reflects Australians' needs and expectations.