REPORT OF THE STATUTORY REVIEW OF THE *Online Safety Act 2021*

Delia Rickard PSM

October 2024

Acknowledgement of Country

I acknowledge the Traditional Custodians of Country throughout Australia and acknowledge their continuing connection to land, waters and community. I pay my respects to the people, the cultures and the Elders past, present and emerging.

Content warning

This report discusses themes and case studies related to the prevention of online harms that may cause distress in readers. Themes include online and technology-facilitated abuse, sexual assault, child sexual exploitation and abuse, self-harm, suicide, disordered eating, pornographic content, and hateful language directed at groups of people.

Contact us

Director—Online Safety Branch
Department of Infrastructure, Transport, Regional Development, Communications and the Arts
GPO Box 594
Canberra ACT 2601 Australia

Email: creative.design@infrastructure.gov.au
Website: www.infrastructure.gov.au

Cover image: ©Adobe Stock.

CONTENTS

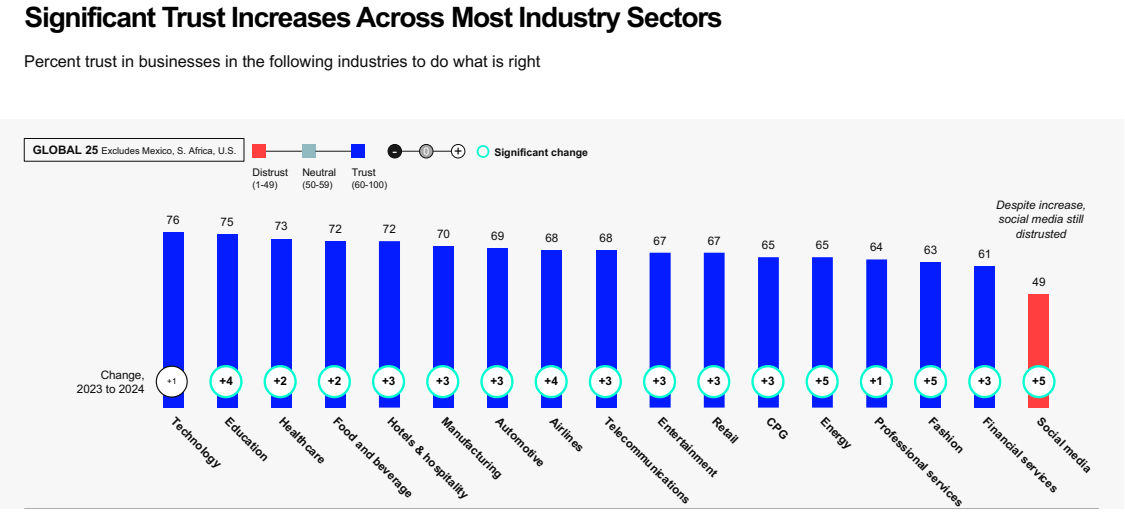
INTRODUCTORY REMARKS

Today there is no escaping the need to be online. Government services, banking, news, making doctors' appointments, homework assignments and even parking meters are all moving online, with more and more areas now only accessible online. This creates multiple new challenges. It also demands that we do all that we can to ensure that the online world is a safe one.

As has been noted many times before, the online world has brought with it many positives. These range from being able to find the answer to almost any question within seconds, connecting with family and friends old and new, access to games and entertainment and connecting with others who share our interests. It also enabled schooling to continue during the COVID years, for many of us to work from home, the ability to shop remotely almost anywhere in the world and significant time saving. All who read this will no doubt have their own examples.

Unfortunately, it has also brought enormous problems that are causing our society and individuals huge harms. These include the proliferation of child sexual exploitation and abuse material and the bullying and abuse of individuals and groups. People say things online, often under the cloak of anonymity, that most of us would never say to a person's face. We are also seeing the promotion of terrorism, ever increasing misogyny, people withdrawing from public life for fear of abuse and the constant promotion of beauty standards that are unattainable for most of us with resulting disordered eating. Mental health issues are on the rise, many are falling prey to the addictive features of services, image-based abuse and deepfakes are proliferating and matters that are outside of the scope of this review, such as disinformation, threats to our democracy and the proliferation of scams, are wreaking havoc globally. That said, most of us regularly choose to spend time on social media despite reports that it is the least trusted of all the main business sectors.

According to the 2024 Edelman Trust Barometer[1] social media is the least trusted sector globally:

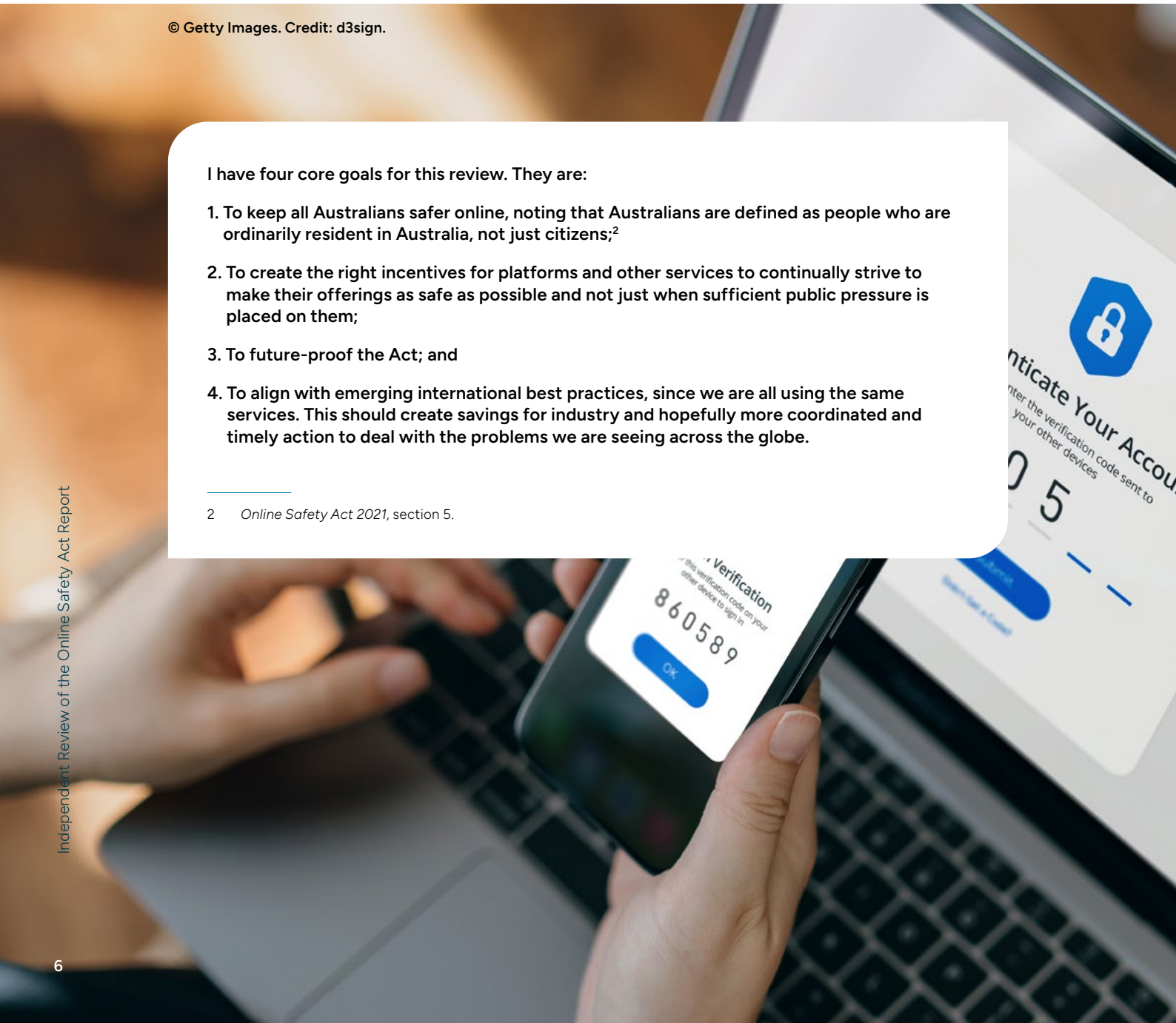


1 Edelman Trust Institute, 2024 Edelman Trust Barometer: Global Report, 45. 2024 Edelman Trust Barometer Global Report, accessed 30 October 2024.

Things need to change!

While some online services do more to help with user safety than others, initiatives often come way too late, don't go far enough and occur only in response to huge amounts of public pressure. For example, Meta, to its credit, has just announced an important range of responses to the plague of sextortion. This is terrific but it is an issue that has been with us and growing for around a decade, and its reforms still don't cover all of its services. It shouldn't have taken until this year, a year where the media has had an enormous focus on harms like sextortion and whether to limit online access for young people, for action to be taken.

Not all of us will be exposed to such harms. Unfortunately, those who are the most marginalised in our society are also the people who experience the greatest harm online. This includes our First Nations people, people from our culturally and linguistically diverse communities, women, children, people with disability and people from the LGBTQIA+ communities. We need to do more to prevent this and make the online world a safer place for all. Sadly, the universal wisdoms that so many of the world's grandmothers taught us, such as treat others as you would like to be treated and if you can't say something nice about a person don't say anything at all, are too frequently ignored in the online world.



I have four core goals for this review. They are:

1. **To keep all Australians safer online, noting that Australians are defined as people who are ordinarily resident in Australia, not just citizens;[2]**
2. **To create the right incentives for platforms and other services to continually strive to make their offerings as safe as possible and not just when sufficient public pressure is placed on them;**
3. **To future-proof the Act; and**
4. **To align with emerging international best practices, since we are all using the same services. This should create savings for industry and hopefully more coordinated and timely action to deal with the problems we are seeing across the globe.**

2 *Online Safety Act 2021*, section 5.

Australia's online safety laws were world leading when introduced in 2015, as was the concept of having a Commissioner and team to enforce the *Online Safety Act 2021* (the Act), conduct much needed research and provide education about staying safe online.

However, since the laws were first introduced, and even since they were updated with the introduction of the Act in 2021, we have been overtaken. While the current approach of taking down harmful material, setting expectations for industry through the unenforceable Basic Online Safety Expectations and enforceable codes and standards has helped many, it has not been able to cope with the scale of problems in the online world.

Some jurisdictions, such as the United Kingdom and the European Union, have moved to systems approaches. Others, such as Canada, are looking to do so. A systems approach puts responsibility on the online services to keep users safe. I think that it is essential for Australia to go in a similar direction. With 5.45 billion internet users around the world as of July 2024,[3] and 96.2 per cent of Australians using the internet with 81 per cent using social media[4], a systems approach is really the only way we can have a meaningful impact on online safety.

Unfortunately, the natural incentives of online platforms and other industry participants don't always align with safety. They make their money by keeping people online and exposed to advertising. Disappointingly, it is often the sensational and extreme content that drives attention and keeps people online so the incentives for content moderation and limiting the time spent online just aren't strong enough. This needs to change.

This report recommends the introduction of an overarching **duty of care** and a **due diligence** approach. These are similar to the concept of duty of care that we are used to in our work, health and safety regime. This would require all services to **take reasonable steps to prevent foreseeable harms.** It would also bring with it a requirement that online services **apply safety by design principles to the design of all new services and to any significant changes to existing ones.** For the largest and riskiest services, this requires at least annual risk assessments, mitigation of risks, measurement of success (or otherwise) and strong transparency reporting. The regulator, eSafety, would have the ability to identify services that are sufficiently risky to be included in this group even if they have less reach.

The Act is technology neutral, so these requirements should also help future-proof the Act as they would apply to all new technologies such as the metaverse and generative artificial intelligence as well as services yet to be imagined.

There would be a provision for eSafety to make mandatory codes of conduct to spell out what services should do in particular areas. These would be drafted by eSafety in close consultation with industry, academics, civil society and others. eSafety would not have to wait for these codes to be in place before it could take action for a breach of the duty of care.

I am also recommending that the core **enduring harms** to be covered by the legislation are those that relate to:

- Harms to young people
- Harms to people's mental and physical wellbeing
- Instruction or promotion of harmful practices
- Threats to national security and social cohesion; and
- Other illegal content, conduct and activity.

I have tried to test whether these overarching harms would cover all of the harms we are concerned about within eSafety's remit – I believe they will.

There is a need to **simplify the Act**, especially the way that it currently breaks the online world into eight categories for regulatory purposes with names that won't resonate with most people and functions that have changed over time. I'm proposing that the eight categories be condensed to four:

- Online platforms
- Online search and app distribution services
- Online infrastructure services; and
- Equipment and operating system services.

I am recommending that Australia **retain the takedown schemes** in the current Act. I have heard too many heartbreaking stories of suicides and real fear resulting from online harms to think we are at the point where quick takedown schemes are no longer needed, though hopefully that time will come.

3 Statista, July 2024.

4 Law, D. Red Search, Australian Internet Statistics 2024 (updated 5 March 2024), Australian Internet Statistics & Facts (2024) - Red Search.

The schemes dealing with cyberbullying of children; adult cyber abuse; image-based abuse and the Online Content Scheme would be retained and, in some instances, strengthened and made more consistent. The Act would also confirm that the informal approach eSafety uses, of contacting the platforms and asking them to take down material, is legitimate. It is clear to me that this approach is the quickest way to get harmful material removed and, as long as statistics about its use are kept, is totally legitimate.

One of the issues that industry and eSafety are united on is that the National Classification Scheme, which is central to our current Online Content Scheme, is no longer fit for purpose as it relates to eSafety's responsibilities. **I am proposing that the Act be decoupled from the scheme** and a harms focus be introduced to replace it.

Human rights are fundamental to the Act, including the right to freedom of opinion and expression. At times though, these rights can be in conflict and a balancing act is required as acknowledged in Article 19 of the International Covenant on Civil and Political Rights. Article 19 (3) provides that freedom of expression may be limited where those limitations are demonstrated to be necessary for ensuring 'respect for the rights and reputations of others.' Our right not to be discriminated against, freedom from cruel, inhuman or degrading treatment, and the right to freedom from arbitrary interference with home and family can be damaged by online hate. For freedom of speech to flourish online, the 'digital town square' in which discourse occurs should be a safe place for expression. If not, the voices of marginalised groups may be silenced out of fear in engaging in hostile online spaces.[5] **I am recommending that services use systems to remove online hate** targeted at individuals and groups with defined protected characteristics. Many services' terms of use already include this. As the editorial in The Australian newspaper said recently: "There should be no room for hate speech, vilification, bullying or abuse online or in public debate."[6]

The report also looks at ways to address a number of **wicked problems** such as the increasing use of end-to-end encryption which hinders platforms' and authorities' ability to detect child sexual exploitation and abuse material, and problems of technology-facilitated abuse and sextortion. I am recommending the use of **fusion cells**, which bring together the smartest people on an issue, to find solutions for some wicked problems where competition incentives fail to do so.

As the Australian Competition and Consumer Commission (ACCC) has done in its reports on digital platforms, I recommend that services have easy ways to report problems, including for non-members, and **internal dispute resolution** schemes that meet criteria set out in a mandatory code. Likewise, I am recommending that a **Digital Ombuds** scheme be introduced.

Another area where Australia is now out of step with other jurisdictions is **penalties**. Our current highest penalty is $782,500. Internationally, penalties in this area are now taking the form of a percentage of a service's global turnover. I am recommending a maximum penalty of the greater of 5 per cent of global turnover or $50 million. In keeping with my goal to create the right incentives for platforms to do a better job of keeping users safe, going to court should normally be a last resort. The approach I've taken has been two-fold. First, to try and get services to fix problems through, for example, persuasion, remedial directions or enforceable undertakings. But, if that doesn't lead to positive change, then eSafety should not hesitate to litigate.

In my experience as a regulator, most major overseas entities comply with Australia's legal system when a regulator seeks to take action on an issue. That said, when they have no presence in Australia they could choose not to comply. The report looks at a number of ways to ensure that this does not become an issue, including **requiring services to have an Australian place for service, business disruption powers** where this can be done within our Constitution, and **licensing services**.

Another important issue this report focuses on is governance. The work of eSafety and the eSafety Commissioner has grown enormously over the years, as has the difficulty of many of the decisions that must be made. There is a good case to be made that eSafety should become a **standalone Commission with at least a Chair, Deputy Chair and a Commissioner**. Collective decision-making helps ensure that all issues are considered from every angle and, in my experience, leads to better decision making. The Commission must also be adequately funded to have the staff and technology needed by an agency with the breadth of responsibilities being recommended.

5 Review submission 135 – Australian Human Rights Commission, 71.

6 The Australian, Editorials, 'Silencing Free Speech is a Bad Idea', 27 September 2024.

My great hope is that the recommendations made in this report lead to a safer, less toxic environment online for all and a more cohesive society. As I once heard an ethicist from a New South Wales University say, **"I know you can but should you?"** This is a question that both services and users should ask themselves regularly.[7]

Acknowledgements

There are many people I need to thank for their assistance in producing this report.

First up is the wonderful Secretariat that the Department kindly provided to help me. They have been of invaluable assistance to me. Their clear and clever thinking, patience and hard work has been instrumental in putting this report together. They have been an excellent sounding board and I have also very much enjoyed their company and devotion to getting the report right. I will miss working with them.

I am also hugely indebted to the staff at eSafety and the eSafety Commissioner. They have been extremely generous in sharing their time and thoughts with me and I very much hope that this report leads to them having all the tools and resources they need to do an even better job at keeping Australians safe online. The work they all do is difficult and challenging. They are exposed to things most of us would hope never to see. We should all be grateful to them and their indefatigable leader Julie Inman Grant. She and her team truly care about making our online world a safer place. I wish them all only the best for the future.

Finally, I want to thank the 168 organisations, industry members and individuals who took the time to send in submissions. I read every one of them, and they helped me with my thinking enormously. I would especially like to thank the victims of online harm who were so honest in sharing their experiences. I'd also like to thank the more than 2,100 individuals who took the time to send in comments.

Likewise, I'd like to thank all the individuals, organisations and industry members I met with during our consultations. These meetings have all helped with putting together the recommendations in this report.

7 Unfortunately, I am unable to recall the person's name, but their comment has stayed with me.

EXECUTIVE SUMMARY

Overview

The statutory review of the *Online Safety Act 2021* (the Act) was announced on 22 November 2023 by the Hon Michelle Rowland MP, the Minister for Communications. As part of the announcement, Minister Rowland acknowledged the review had been brought forward by one year to make sure Australia's online safety laws keep pace with the evolving online environment. Ms Delia Rickard PSM was appointed to conduct the review.

As set out in the Terms of Reference, the purpose of this review was to undertake a broad-ranging examination of the operation and effectiveness of the Act, including key consideration of:

- The existing complaints schemes
- Whether the Act should be amended to include a duty of care
- Ensuring that industry acts in the best interests of the child
- Whether additional arrangements are needed to capture harms not explicitly covered by the Act
- Whether penalties and enforcement are adequate; and
- Whether the existing powers are sufficient or changes are needed to strengthen the Act.

The review has drawn on evidence from a range of sources, including extensive stakeholder consultation, public submissions and available research. The public submission process yielded more than 2,270 responses with 169 substantive submissions. The review also involved 72 meetings and roundtables, with civil society organisations, government and law enforcement agencies, the tech and digital platforms industry and international stakeholders.

The Online Safety Act 2021

The review found that the Act has been world leading and has provided support to individuals through a range of complaint and removal schemes as well as strengthened powers to address illegal and harmful material. It has also established trail-blazing transparency requirements on industry and enabled eSafety to engage in a range of educational activity to help Australians stay safe online.

The review considered the current objects of the Act which are to improve and promote online safety for Australians and found that more descriptive objects would better serve the Act. It has recommended that the objects of the Act would be to enhance the online safety of Australians and Australia by:

- Promoting human rights and safety
- Promoting and protecting the best interests of the child
- Building an evidence base around online safety and existing and emerging online harms
- Preventing and alleviating online harm present in Australia; and
- Improving online safety for all in Australia by advancing service provider responsibility for preventing harms and mitigating the damage done along with user empowerment and transparency.

Defining the online industry

The review found that the eight sections currently specified for the purposes of the Act are no longer fit for purpose. The sections are currently defined as social media services, relevant electronic services, designated internet services, internet search engine services, app distribution services, hosting services, internet carriage services and those who manufacture, supply, maintain or install relevant equipment. The review heard that the sections created uncertainty, are complicated and confusing.

The review has recommended new categories of industry sections to better reflect a risk-based and proportionate regulatory approach. In defining new categories of industry, the goal is to capture all parts of industry which may facilitate online harm. The following industry sections have been recommended:

1. Online Platforms (services providing online interaction and online content)
2. Online Search and App Distribution Services (services which gate-keep access to online platforms)
3. Online Infrastructure Services; and
4. Equipment and Operating System Services (including manufacturers, suppliers, maintenance and installers).

The review recommends a proportionate approach to the application of obligations on industry, particularly in relation to reporting. Obligations should be tiered according to their reach (that is the extent of their Australian user base) and the level of risk associated with use of the service. This is to capture services in addition to the most popular platforms, which may inherently pose a greater level of risk to some or all users - due to their features, use by vulnerable groups (including children), or past behaviour or policies with respect to safety.

A duty of care

A common theme heard throughout consultation was the need for a more systemic and preventative approach to online harms. Australia's current laws and regulatory settings are not good enough to address the volume of online harm that is occurring. **The review recommends that Australia adopt a singular and overarching duty of care that encompasses due diligence, and is underpinned by safety by design principles, risk assessment, mitigation and measurement.** An overarching duty of care would place responsibility on service providers to take reasonable steps to address and prevent foreseeable harms on their services. It shifts much of the burden for remaining safe online away from individual users and onto those most capable of identifying and addressing harms – the service providers themselves.

Our major international counterparts in Europe, the United Kingdom and North America are almost all moving towards a systems-based, proactive approach. While there are differences in approaches, the overall objective remains the same: services must take reasonable steps to keep their users safe. Greater consistency across our respective national regimes would simplify compliance for service providers, reducing costs and regulatory burdens. This would also provide economies of scale and more coordinated and efficient investments in safety. Measures which maximise convergence of regulation between countries can also help to maximise the potential for securing and enforcing extra-territorial compliance.

Enduring categories of harm

To complement the overarching duty of care provisions and framework for the performance of due diligence, the review recommends establishing **enduring categories of harm** within the Act. While the examples within the categories may change over time, the following broad categories should be included in a reformed Act:

- Harms to young people
- Harms to people's mental and physical wellbeing
- Instruction or promotion of harmful practices
- Threats to national security and social cohesion; and
- Other illegal content, conduct and activity.

Risk assessment

An essential part of meeting the duty of care for online service providers is a requirement to undertake regular risk assessments of their services. Risk assessment requirements are a core feature under both the European Union's Digital Services Act and the United Kingdom's *Online Safety Act 2023*, and are built into the first phase of Australia's industry codes and standards. They are at the heart of a preventative and systemic approach to making the online world a safer place by design, and by working to prevent harms rather than merely responding after the fact. As the saying goes, 'it is better to put a fence at the top of the cliff, than an ambulance at the bottom of it.'

All service providers should diligently perform risk assessments and implement mitigations (which include safety by design principles), both at regular intervals and when introducing or significantly altering products or features. However, stringent and enforceable risk assessment requirements should particularly be placed on the larger services with high 'reach' and other services posing a high risk.

Risk reporting obligations must capture the whole risk assessment cycle and include the essential components of assessment, mitigation and measurement. The review found that to remain effective, risk assessment must be ongoing, and will require services to regularly repeat the process.

The review has recommended that in introducing a duty of care, the regulator should be empowered to make codes to provide mandatory and enforceable compliance measures to direct them about how to comply with certain aspects of a duty of care. Codes, however, are not intended to create safe harbours and the absence of a code should not prevent the regulator from taking enforcement action under the duty of care.

A considerable amount of time and effort has gone into developing the industry codes and standards under the Online Content Scheme, and the work is still underway on developing a second phase of codes. The report supports the continuation of this work as it will take time to make and implement any legislative changes based on the recommendations in the report. Implementing the current Act should continue so that the protections it provides remain in place. There will need to be transitional arrangements to ensure a continuity of protection under the Act as the new framework is implemented.

Accountability and transparency

The review observed the opaque nature of services, and in particular online platforms and search and app distribution services, and that the term "black box" is often associated with online services. It is recommended that transparency measures are put in place both so that the regulator can properly monitor the safety of services and so that others can make assessments of how much trust to place in the services.

The review found that one of the most useful powers that eSafety has is its ability to require services to provide information related to the Basic Online Safety Expectations. This enables the regulator to ask forensic questions and make an assessment about how much services are, or are not, doing to keep users safe. It can deliver broader online safety gains by shedding light on a service's practices.

The review found that more transparency is needed and **recommended that services with the greatest reach or risk prepare and provide an annual transparency report to eSafety**. With transparency reports required annually, it is recommended that the service publish a summary of its report on its website, rather than eSafety producing the public report as is currently the case. Services would not need to reveal matters that are commercial in confidence or which could be used by bad actors to, for example, circumvent systems.

The regulator should continue to be able to require transparency reports from all services and ask the questions needed to better understand what services are and aren't doing, and the consequences.

Compliance and audits

Ideally all services should have a well-resourced compliance function that reports directly to the audit and risk committee (or equivalent) as needed, but at least quarterly. At a minimum **all services of greatest reach or risk must have a compliance function**. The compliance function should be independent from other areas of the service and staff should have training in compliance. Only the board should be able to dismiss the head of the compliance function.

eSafety should have the discretion to require a service to be audited at their own expense and provide the audit report to eSafety.

Providing researchers with information

Research contributes greatly to society's ability to meet current and future changes and can directly benefit the wellbeing of citizens. A scheme that provides accredited independent researchers access to data would encourage more research and more detailed consideration of the many complex problems in the online world and help decision makers. The report has recommended that **those services designated as having the greatest reach or risk should be required to be involved in sharing data** for research purposes, though clearly other services could voluntarily do so.

The scheme would be targeted towards research for the purposes of determining compliance with a duty of care model, the takedown schemes and research into emerging problems and harms. Services should only be able to refuse access if they do not have the data, if giving access to the data will lead to significant vulnerabilities in the security of their service, or if it would compromise confidential information (including trade secrets). Researchers would need to be authorised to participate in the scheme and the scheme would need to be designed to minimise the administrative burden for all involved, ensuring the projects are of genuine value to the advancement of online safety and to **have regard to the *Privacy Act 1988* (Privacy Act).**

Safety nets – supporting online users

No government can completely protect its people from online harms. Systems-based regulation, such as a duty of care and due diligence, aims to prevent harms from occurring, whereas complaint-based removal schemes focus on minimising the impact of harms once they have occurred. Investigating individual complaints is resource intensive but necessary, at least for now, to protect individuals and limit the harm they experience.

The review found that the removal schemes are recognised as a strong, world-leading model of regulation that have been successful in addressing impacts on individual users. Civil society and industry representatives noted that the existing schemes are highly valued and mostly perceived to be working well. In addition to the benefits that a duty of care will bring, **there are changes needed to address inconsistencies across the four complaint and content-based removal schemes and changes to better support those making a complaint**.

Reducing the wait time to issue a notice

The review found that the complaint scheme rules allow seriously harmful content to remain online for too long. Under current arrangements, the online service must have failed to act on a complaint for 48 hours before eSafety can issue a formal removal notice. It **is recommended that this is reduced to 24 hours**. However, there are circumstances such as where no clear complaint mechanisms exist on the online service, or where reporting would lead to a reasonably foreseeable risk of further harm to the user experiencing the abuse, where eSafety should be empowered to waive the statutory delay.

Adult cyber abuse scheme

The **adult cyber abuse scheme should be amended by lowering the threshold**. The new threshold should require that an ordinary reasonable person would conclude that 'it is likely the material was intended to have an effect on a particular Australian adult', and that an ordinary reasonable person would 'regard the material as being, in all the circumstances, menacing, harassing or seriously offensive'.

The Act should also include additional powers to require an end-user to stop posting cyber abuse about an Australian adult in an end-user notice, subject to a civil penalty for non-compliance.

Harmful patterns of behaviour

The review found that the schemes' current focus on specific items of content can limit the regulatory response to harmful material that is reposted after it has been taken down. Even though the content has previously met the regulatory threshold for action, the affected individual would need to make a new complaint to eSafety before the reposted material could be removed. It is recommended that the Act be amended to **enable eSafety to issue a removal notice for material that has met the regulatory threshold for removal under a prior complaint, where eSafety becomes aware that the material has been reposted.**

The review also found there are unnecessary inconsistencies between some of the schemes and recommends that eSafety be empowered to issue an end-user notice requiring a user to stop posting cyber abuse about an Australian adult. This would **better align the adult cyber abuse scheme with the child cyberbullying and image-based abuse schemes.**

Online hate

The review acknowledges that online hate is not new, but its prevalence online and its ability to spread at a magnitude and order not seen before is a major concern. Online hate has the potential to cause significant harm to individuals and impact community safety more broadly. The review heard about many experiences of individuals and community groups experiencing online hate and it is clear that further regulatory intervention is needed to address these harms. It is recommended that a **definition of online hate material be included in the Act**, that a systems approach is adopted to stop online hate against individuals and groups, and that when interpreting the threshold of harm for adult cyber abuse, that online hate material is considered.

Volumetric attacks

The review considered volumetric (pile-on) attacks and heard many individual experiences of online abuse which included volumetric or 'pile-on' attacks. Where the harm of individual comments can be damaging to the targeted user's wellbeing, the impact of a volumetric attack done at scale can magnify and compound the harm. Often the content is shared with an accelerating level of outrage and toxicity, and ultimately a high volume of abuse. These attacks can be among the most serious forms of online abuse. It is recommended that the Act **defines a 'volumetric attack' or 'pile-on' attack**, which is currently not defined.

The distribution of harmful content by various individual users and across different platforms means there is no single point for regulatory action. The Act should also be amended to provide eSafety with the ability to issue a notice to services in relation to a suspected 'volumetric attack', which may require information related to the attack, specify remedial actions to be taken, and require the service to report back on steps taken.

No wrong door

The review also explored ways to better support individuals who are seeking help for a harm experienced online, and acknowledged that this is usually when they have suffered something significant, such as abuse, threats, reputational damage or financial losses. There are many places that a person can complain to, including eSafety, the police and anti-discrimination bodies and other regulators. In light of this, the report has recommended that **the Australian Government should develop a whole of government 'no wrong door' approach to support individuals**

seeking help to address online harms. This will require cooperation and information sharing across portfolios, including law enforcement, to address a range of issues such as online safety, child safety, privacy and scams, among others.

Striking a balance between protections and freedoms

During the review, protecting freedom of speech or expression was a key concern raised. Some raised concerns that content moderation limits freedom of speech, while others described the silencing effects of online abuse and adverse impacts on their work, health, relationships and personal security. The review acknowledges that all human rights are indivisible and afforded equal status, but that freedom of expression requires specific consideration in online spaces because of the opportunities digital platforms provide for realising the benefits of free speech. A broad range of human rights interact with freedom of speech, including: freedom of opinion and expression, freedom of thought, conscience, and religion or belief, right to take part in public affairs and elections, right of privacy and reputation, right to health, rights of equality and non-discrimination and the rights of the child. As the right to freedom of speech is not absolute, a balancing act between competing rights is required. **The proposal to amend the objects of the Act makes it clear that online safety regulation needs to be centred around all human rights, and not just the right to free speech**.

Dispute resolution

Access to good dispute resolution mechanisms is an important part of how we protect people in Australian society. The eSafety takedown schemes don't catch all types of bad conduct and even world class systems for platforms are not 100 per cent foolproof. The review considered how to better support individuals who need somewhere to go to resolve disputes. This includes people whose posts have been removed who believe they have been taken down unfairly as well as people who have failed to have posts that harm them or their group taken down.

All services should be required to have an **easily accessible, simple and user-friendly way to make a complaint** and internal complaint handling processes that are in line with a code on internal dispute resolution. In particular, this should include a way for non-users to report issues such as when their intimate images have been posted without consent. Services should also be required to respond to reports within a reasonable time. The review also recommends that, in line with the ACCC's Digital Platform Services Inquiry, the Australian Government should develop and implement an **Ombuds scheme** that covers digital platforms and online search and app distribution services.

Wicked problems

The recommendations in the report to introduce a duty of care obligation on services and strengthening the complaints schemes are expected to go a long way to addressing many online harms and result in a significant uplift in online safety for all Australians. That said, the review found that some serious harms are likely to require considerably more work to move the dial. Some examples provided are the complex issue of targeted technology facilitated abuse; the increasing use of end-to-end encryption and the implications for being able to deal with child sexual exploitation and abuse material and other illegal material, and sextortion, which is often perpetrated offshore along with many other online scams.

The review recommends that the Government seek to prohibit 'nudify' apps and services and undetectable cyberstalking apps. It notes that the benefits of end-to-end encryption have been recognised by services, governments and others (though many disagree) and that services must rise to the challenge of preventing and detecting child sexual exploitation and abuse material despite the existence of encryption. Technology facilitated abuse and addressing child sexual exploitation and abuse material despite end-to-end encryption are problems that will require a multi-dimensional and multi-stakeholder approach if we are to make a real difference.

It is recommended that the Australian Government and the regulator should both be able to convene **multi-stakeholder 'fusion cells'**, that involve the smartest people on the issues to analyse 'wicked problems' (such as the implications of end-to-end encryption for combatting child sexual exploitation and abuse and technology facilitated abuse and gender-based violence) and develop coordinated multi-agency and multi-stakeholder solutions.

The review acknowledges the recent introduction by Snap, Meta and Apple of new and enhanced safety features to combat sextortion, but it is likely that more still needs to be done and applied to all relevant services. If competition doesn't spur comprehensive responses then combatting sextortion may also benefit from a fusion cell approach, especially as sextortion often involves multiple platforms in a single sextortion attempt.

Links to the National Classification Scheme

During the review, industry and eSafety raised the issue of needing to decouple the regulation of Class 1 and Class 2 material from the National Classification Scheme (Classification Scheme). The review found that using these borrowed thresholds, which entail applying a range of considerations under the classification scheme, is not fit for purpose. Instead, a framework that supports efficient decision making of dynamic and potentially high-volume online content and allows for rapid responses to illegal and harmful content is needed.

The review recommends that the Act be decoupled from the National Classification Scheme with new Class 1 and Class 2 definitions and thresholds to be specified in the Act and, as far as possible, should align with equivalent standards in the National Classification Scheme. In addition, it is recommended that regulatory remit of eSafety and the National Classification Scheme is more clearly defined, and in particular that content that is currently classified should not also be subject to the Act, with the exception of social media enabled user generated and interactive features such as chat features in gaming.

Penalties and enforcement

The review found that **more significant penalties are needed** to act as a deterrent and to take appropriate enforcement action, especially for those online services which are among the richest global corporations in the world. Should new obligations be placed on services under a duty of care, appropriate and persuasive penalties must be in place. Coupled with stronger penalties, there needs to be a range of enforcement options available to the regulator, including those with a remedial focus.

The review recommends a stronger maximum civil penalty. **Maximum penalties for a breach of the duty of care should be increased to the greater of 5 per cent of global annual turnover or $50 million. The civil penalties for non-compliance with removal notices should be increased to a maximum of $10 million for companies**.

The review supports broader application of remedial powers and improved alignment in penalty provisions across the complaints schemes under the Act including that immediate link-deletion powers should be extended to all of the content removal schemes, not

just the Online Content Scheme, to limit the discoverability of harmful material. The regulator should also have streamlined and more effective powers to deal with individuals who continually harass and abuse others online.

The review acknowledges the **fundamental challenges of extra-territorial enforcement** that apply to regulating the online world, and recommends exploring a number of options to ensure that services submit to the jurisdiction of our courts and their rulings. These include:

- Requiring major platforms to have a local presence in Australia
- Exploring the feasibility of requiring at least large services to be licensed to operate
- At a minimum, a point of service should be established by major platforms in Australia
- Broader access restrictions; and
- Business disruption powers.

It is also suggested that Australia's enforcement of online safety laws will be most effective if it is 'interoperable' and coordinated with like action by our international partners, such as the United Kingdom and the European Union.

Investigation and information gathering powers

The majority of the eSafety Commissioner's current investigative work is focused on complaints made under the Act's removal schemes, and the codes and standards. However, a duty of care would involve proactive and systemic obligations and more general and robust investigations powers are needed.

To support its investigations authority under an expanded Act, the regulator will need the right powers to conduct investigations, monitor compliance and to inspect, audit and validate information provided by services. Specifically, this includes **providing the regulator with flexibility and the right technological tools** to assist with investigations, content removal and the use of sock-puppet accounts.

The review found that changes are also needed to eSafety's information gathering and disclosure powers. Where the Commissioner believes on reasonable grounds that an online service has information about the identity or contact details of the end-user, and it relates to the operation of the Act, it may be necessary to obtain end-user information (basic subscriber information). For example, in the circumstances of issuing an end-user removal notice for child cyberbullying, image-based abuse or adult cyber abuse material, the Commissioner would be empowered to unmask the anonymity of users where an investigation or the exercise of regulatory powers requires this.

The ability to share information about investigations of online services and online harms more broadly where it relates to the operations of the Act can deliver better regulatory outcomes. The review has recommended that the Act be amended to **allow eSafety to disclose information** to any head of a Commonwealth agency or department or an international authority. There is also a need to be able to disclose to teachers, school principals, parents or guardians regarding complaints about image-based abuse to bring it in line with the child cyberbullying scheme. Where non-government organisations have an approved role in assisting eSafety with enforcement activities, eSafety should also be able disclose certain information.

Promotion, education and research

A core function of eSafety, which has been in place from the very start, is that of promotion and education. Teaching the community about online safety, supporting others to deliver online safety education and promoting the supports available to those who are experiencing online harm is crucial. The review found that these functions (awareness raising about eSafety, education and capacity building to prevent online harms, strategic partnerships, and research and evaluation) are as important as ever.

However, throughout the review, a common theme raised was the **need for more to be done to boost awareness of eSafety** and online safety more broadly, particularly in harder to reach groups such as First Nations and remote communities. The review acknowledges the significant work done to promote eSafety and educate the community and notes that there are encouraging trends in these areas. It supports the continuation of these efforts and continued leveraging of media opportunities and strategic partnerships with sporting organisations as well as education and community sectors.

© Getty Images. Credit: GCShutter.

Governance – a future-proofed regulator

In considering the current governance arrangements, the review acknowledged that the functions and powers of the eSafety Commissioner have increased substantially since the creation of the role in 2015. Coupled with an increasingly complicated and contested operating environment, a new governance structure—a **Commission model of governance**—is recommended. A Commission model of governance would support better decision-making and would include a Chair, Deputy Chair and a Commissioner, with the potential for there to be up to nine Commission members. The new Commission should be known as the **Online Safety Commission.**

Ultimately, the ideal end state is a standalone, independent regulator to support eSafety's growing functions and responsibilities.

In any event, the **regulator must be appropriately resourced and have the right regulatory infrastructure** in place to carry out its functions. This includes an ongoing dedicated and appropriately resourced legal team, appropriate corporate management and the right IT in place to do its job well. In determining what may be appropriate in the eSafety context, consideration should be given to how other regulators operate. The review has also recommended that a cost recovery mechanism be developed to fund the cost of regulating industry.

A reform pathway

Should the key recommendations in this report be adopted, their development and implementation will take time to get right. However, this does not detract from the urgency of implementing the recommendations as soon as practicable and, if required, prioritising those changes that provide the most immediate benefits to Australians.

The report recommends that **implementing a duty of care and supporting eSafety are the first priority**. A duty of care is a priority as it will be the most effective and immediate means of improving online safety for Australians, and online services will require a reasonable time to adapt to the new regulatory model. It is also a priority to move to a multi-Commissioner model of governance. Improving the operation of some or all of the Act's four complaints schemes (child cyberbullying, adult cyber abuse, non-consensual sharing of intimate images, and the Online Content Scheme) will have an additional direct benefit to those Australians who experience online harms.

The review has also highlighted one of the enduring challenges of attempting to regulate the online world. That is, it is continuously evolving and governments all over the world are constantly playing catch-up. To address this, the review recommends that an updated Act should be subject to an independent review three years after the commencement of the key reforms to the Act, or by 2029, whichever is earliest. In addition, the Australian Government should consider how its existing administrative arrangements relating to online harms are operating and consider the merits of an overarching Digital Services Commission.

RECOMMENDATIONS

Recommendation 1: That the objects of the Act should be amended to include more descriptive objectives that are linked to the various functions covered by the Act.

Recommendation 2: That current definitions of the online industry sections should be simplified to online platforms, online search and app distribution services, online infrastructure services and equipment and operating system services. These should be included in the Act to better reflect online safety risks and future proof the Act.

Recommendation 3: That the Government consider options to recognise the role of providers of online safety related services and technology in helping to identify and stop the distribution of child sexual exploitation and abuse material.

Recommendation 4: That Australia adopt a singular and overarching duty of care that encompasses due diligence, and is underpinned by safety by design principles, risk assessment, risk mitigation and measurement.

Recommendation 5: The harms that should be highlighted for attention under a duty of care should at a minimum include:

- Harms to young people, including child sexual exploitation and abuse (including grooming), bullying and problematic internet use
- Harms to mental and physical wellbeing, including threats to harm or kill, or attacks based on a person or group of people's protected characteristics, such as sex, gender, sexual orientation, race, ethnicity, disability, age or religion
- Instruction or promotion of harmful practices, such as self-harm/suicide, disordered eating and dares that could lead to grievous harm
- Threats to national security and social cohesion, such as through promotion of terrorism and abhorrent violent extremist content; and
- Other illegal content, conduct and activity.

Recommendation 6: Entities with the greatest reach or risk should be required to complete a risk assessment at least every 12 months and to carry out a risk assessment when significant changes are made to the design and operation of their service. These entities should also be required to provide an annual report detailing their risk assessments, risk mitigations and how successful they have been to the regulator.

Recommendation 7: Services used by more than 10 per cent of the Australian population should be automatically part of the highest tier with additional mandatory responsibilities. The regulator should have a power to deem whether other online services do, or do not, meet the reach or risk requirement, noting that the reach or risk of services may change over time.

Recommendation 8: The best interests of the child should be a primary consideration for online service providers in assessing and mitigating the risks arising from the design and operation of their services, including risks to children who may use the service and risks to children as a result of how the service may be used.

Recommendation 9: The eSafety Commissioner should be empowered to create mandatory rules (in the form of codes) on how entities can comply with certain aspects of the duty of care requirements, including addressing specific online harms. This should not stop services from taking additional steps to protect people. Codes would not create safe harbours.

Recommendation 10: In addition to risk assessments, a service with the greatest reach or risk should be required to provide an annual transparency report and publish a summarised version on its website. This should not replace the broad power for eSafety to require periodic and non-periodic transparency reports from all services.

Recommendation 11: Services with the greatest reach or risk should be required to have a well-resourced compliance function that reports directly to senior management as needed, and at least quarterly to the audit and risk committee and annually to the board. Only the board (or its equivalent) can dismiss the head of the compliance function.

Recommendation 12: The regulator should have the discretion and power to require services to undertake an audit at their own expense.

Recommendation 13: Subject to adequate safeguards, services with the greatest reach or risk should be required to share data with authorised researchers for the purposes of determining compliance with a duty of care model, the takedown schemes and research into emerging problems and harms.

Recommendation 14: For the avoidance of doubt, the legislation should make it clear that informal requests for takedown are legal and legitimate as they lead to quicker results for individuals who are often in severe distress.

Recommendation 15: Users experiencing adult cyber abuse or child cyberbullying should only need to wait 24 hours (not 48 hours) following a complaint to a service before eSafety is able to issue a removal notice.

Recommendation 16: The regulator should be empowered to waive the statutory delay to issue a removal notice for the child cyberbullying and adult cyber abuse schemes where no clear complaint mechanism exists on the online service, or where reporting would lead to a reasonably foreseeable risk of further harm to the user experiencing the abuse.

Recommendation 17: The Government should develop a whole of government 'no wrong door' approach to support individuals seeking help to address online harms. This will require cooperation and information sharing across portfolios, including law enforcement, to address a range of issues such as online safety, child safety, privacy and scams, among others.

Recommendation 18: The adult cyber abuse scheme should be amended by lowering the threshold. The new threshold should require that an ordinary reasonable person would conclude that 'it is likely the material was intended to have an effect on a particular Australian adult', and that an ordinary reasonable person would 'regard the material as being, in all the circumstances, menacing, harassing or seriously offensive.'

Recommendation 19: The Act should enable the regulator to issue a removal notice for material that has met the regulatory threshold for removal under a prior complaint, where the regulator becomes aware that the material has been reposted.

Recommendation 20: The Act should include additional powers to require an end-user to stop posting cyber abuse about an Australian adult in an end-user notice, subject to a civil penalty for non-compliance.

Recommendation 21: The Act should include a definition of online hate material. The definition should acknowledge that online hate involves an attack against a person or people that is based on a protected characteristic and can include dehumanisation. Notably, the definition of online hate material should not include views regarding ideas, concepts or institutions. The definition should also consider potential exclusions (for example where material is posted for artistic, scientific, or journalistic purposes).

Recommendation 22: The Act should be amended to ensure that, in interpreting the threshold of harm for adult cyber abuse, the reasonably proximate cumulative harm caused by online hate material is taken into account.

Recommendation 23: The Act should define a 'volumetric attack' and the regulator should be empowered to issue a notice or notices to multiple platforms based on a single complaint to address volumetric attacks.

Recommendation 24: The Act should be amended to provide the regulator with the ability to issue a notice to services in relation to a suspected 'volumetric attack', which may require information related to the attack, specify remedial actions to be taken, and require the service to report back on steps taken.

Recommendation 25: All services should be required to have an easily accessible, simple and user-friendly way to make a complaint and internal complaint handling processes that are in line with a code on internal dispute resolution. In particular, this should include a way for non-users to report issues such as when intimate images have been posted without consent on a service. Services should also be required to respond to reports within a reasonable time and for some issues within 24 hours.

Recommendation 26: In line with the Australian Competition and Consumer Commission's Digital Platform Services Inquiry, the Government should develop and implement an Ombuds scheme that covers digital platforms and online search and app distribution services.

Recommendation 27: The Government should explore how best to prohibit search engines and app stores from surfacing, selling or distributing 'nudify' apps and undetectable stalking apps.

Recommendation 28: The Government and the regulator should both be able convene multi-stakeholder 'fusion cells' to analyse 'wicked problems' (such as the implications of end-to-end encryption for combatting child sexual exploitation and abuse, and technology-facilitated abuse and gender-based violence) and develop coordinated multi-stakeholder solutions.

Recommendation 29: The Act should be decoupled from the National Classification Scheme with new Class 1 and Class 2 definitions and thresholds specified in the Act and, as far as possible, be based on equivalent standards in the National Classification Scheme.

Recommendation 30: New Class 1 definitions and thresholds should clearly focus on illegal and seriously harmful material and directly correspond to the Criminal Code where appropriate. Sexually explicit material that includes violent and seriously injurious practices, such as choking, should sit under Class 1.

Recommendation 31: New Class 2 definitions and thresholds should include material that is legal but may be harmful, particularly for minors, and consensual sexually explicit material including non-injurious fetish material.

Recommendation 32: Class 2 definitions and thresholds should also capture material dealing with harmful practices such as disordered eating, self-harm and substance use to address their heightened impact, especially on young people, in the context of social media. In the longer term, so that adults are covered, industry should be obliged to prevent dissemination of such content through a broader code dealing with mental and physical wellbeing under duty of care provisions.

Recommendation 33: In reforming the Act and the National Classification Scheme, the regulatory remit of eSafety should be clarified. Content that is subject to the National Classification Scheme should fall outside eSafety's remit (except features that are uniquely social media enabled).

Recommendation 34: The maximum civil penalty that a court can impose should be increased to the greater of 5 per cent of global annual turnover or $50 million.

Recommendation 35: The civil penalties for non-compliance with removal notices should be increased to a maximum of $10 million for companies.

Recommendation 36: The Act should be amended to empower the regulator to use enforceable undertakings or issue remedial directions to services in relation to all relevant penalty provisions, to seek to bring them back into compliance.

Recommendation 37: The Act should allow removal and link-deletion notices to be issued simultaneously under the Online Content Scheme.

Recommendation 38: The Act should empower the regulator to simultaneously issue link removal notices for all harmful content under removal schemes.

Recommendation 39: The finalised duty of care model should include scope to consider repeated non-compliance by services in removing content as evidence of non-compliance with the duty of care.

Recommendation 40: The Act should include consistent powers across the schemes to require end-users to remove content and refrain from posting abuse in the future.

Recommendation 41: The Government should expand access restriction powers against services for seriously harmful non-compliance.

Recommendation 42: The Government should consider options for business disruption powers for seriously harmful non-compliance.

Recommendation 43: The Government should consider the feasibility of requiring major platforms to have a local presence for the purpose of facilitating enforcement action.

Recommendation 44: The Act should require major platforms, that is those designated under the reach or risk criteria under the duty of care requirements, to have a contact point for service in Australia.

Recommendation 45: The Government should consider options for introducing a licensing scheme for major services as a condition for operation.

Recommendation 46: The Act should be amended to empower the regulator with stronger powers in relation to investigations, including to:

- Incorporate the monitoring and investigations provisions of the Regulatory Powers Act into the Act
- Initiate investigations of a service's compliance with the duty of care; and
- Initiate investigations into reposted material that was previously reported and taken down.

Recommendation 47: Amend the Act to provide the regulator with appropriate flexibility to conduct investigations as it thinks fit, including the use of technological tools to assist with investigations and content removal, and the use of sock-puppet accounts.

Recommendation 48: Provide additional powers to the regulator to improve its ability to obtain end-user information under Part 13, including a requirement that prevents services from informing end-users when they have received a notice under Part 13, a requirement for services to collect a user's phone number as a condition for opening an account, and provide a new power to compel the preservation of accounts for investigative purposes.

Recommendation 49: The Act should be amended to empower the regulator with stronger information gathering powers, including to:

- Improve its ability to obtain end-user information under Part 13 of the Act; and
- Set the time period for a written notice to provide evidence under Part 14 of the Act.

Recommendation 50: Section 205 of the Act should be amended to confirm that non-compliance with a requirement to give evidence includes information as requested under section 199 (and other sections in Part 14 of the Act).

Recommendation 51: The Act should be amended to require services to inform the regulator of actions the service has taken in response to the regulator's actions and requests (including informal requests).

Recommendation 52: The Act should be amended to require services to maintain certain records, such as measures taken to comply with obligations under the Act and any actions taken in response to the regulators requests and risk assessments, for the purposes of eSafety's investigations.

Recommendation 53: The Act should be amended to allow the regulator to disclose information to:

- Any head of a Commonwealth agency or department
- International authorities; and
- Teachers, school principals, parents or guardians regarding complaints from a child about image-based abuse (as can be done for child cyberbullying).

Recommendation 54: Allow the regulator to disclose certain information to Non-Government Organisations who have an approved role in assisting the regulator with enforcement activities.

Recommendation 55: The regulator's continued awareness raising activities should include in-person outreach, including in hard to reach communities, and hard copy resources.

Recommendation 56: Educational and promotional material should not only focus on what the regulator does for people experiencing harms, but also include simple messaging about how to make a complaint. Online safety education delivered at schools should focus on awareness of the regulator as a source of help. News media outlets should be encouraged to provide information about the regulator at the end of articles detailing experiences of online harms.

Recommendation 57: If a decision to make structural changes to the regulator includes a change to its name, a major campaign re-launching the regulator should be conducted. The timing of this campaign should be coordinated to align with major changes to the Act.

Recommendation 58: To support collective decision making, the regulator should move to a Commission model of governance and be known as the 'Online Safety Commission'.

Recommendation 59: That the Commission should be comprised of a Chair, Deputy Chair and a Commissioner, with flexibility for the Commission to grow up to nine members as the functions and powers of the regulator increase.

Recommendation 60: That in moving to a Commission, the Act should require Commission members to have an appropriate mix of skills to support informed and robust decision-making.

Recommendation 61: That a newly formed Commission has strong internal governance processes, is transparent in how it does its work, and ensures that it reports meaningfully on its performance.

Recommendation 62: That following consideration of the regulator's functions and responsibilities under a new regulatory framework, the regulator should transition to a standalone, independent regulator to support its growing functions and responsibilities, and to future-proof the regulator.

Recommendation 63: That the regulator should be appropriately resourced to implement the right regulatory infrastructure and carry out its functions. This includes having an ongoing dedicated and appropriately resourced legal team, appropriate corporate management and the information technology it needs to do its job well. Consideration should be given to how other regulators operate to determine what may be appropriate in the regulator's context.

Recommendation 64: A cost recovery mechanism should be developed to fund the cost of regulating industry, with details to be settled by government in consultation with industry.

Recommendation 65: That if required, the Government should prioritise implementation of the key reforms arising from this review that will provide the most substantial and immediate online safety protections for Australians, including in particular the new duty of care and associated reforms. This should coincide with the regulator moving to a Commission model of governance and appropriate resourcing to support the implementation of priority reforms.

Recommendation 66: That the updated Act be subject to independent review within three years of the commencement of the key reforms to the Act, or by 2029, whichever is earliest.

Recommendation 67: That the Government consider how its existing administrative arrangements relating to online harms are operating and whether there is a case for having a central online harms regulator. Given the level of change that needs to happen now to better protect Australians, this consideration may be best left to around the time of the next review.

INTRODUCTION

01

1.1 Terms of reference for the review

Under section 239A of the *Online Safety Act 2021* (the Act), the Minister for Communications is required to initiate an independent review of the Act. This must be done within three years of the Act's commencement and a written report of review must be tabled in each House of the Parliament within 15 sitting days of being received by the Government. On 22 November 2023, the Hon Michelle Rowland MP, Minister for Communications, announced that the review of the Act had been brought forward by one year to make sure Australia's online safety laws keep pace with the evolving online environment.

The Terms of Reference for the Review were published in February 2024, detailing a broad-ranging examination of the operation and effectiveness of the Act. A final report was to be provided to Government by 31 October 2024.

Specifically, the Review was to include consideration of:

- The existing complaints schemes
- Whether the Act should be amended to include a duty of care
- Ensuring that industry acts in the best interests of the child
- Whether additional arrangements are needed to capture harms not explicitly covered by the Act
- Whether penalties and enforcement are adequate; and
- Whether the existing powers are sufficient or changes are needed to strengthen the Act.

The Terms of Reference are reproduced in full at Appendix B.

1.2 An overview of the review process

On 22 November 2023, the Hon Michelle Rowland MP, Minister for Communications, announced the commencement of the independent statutory review and appointed Ms Delia Rickard PSM to conduct the review.

Ms Rickard was the Deputy Chair of the Australian Competition and Consumer Commission (ACCC) for more than a decade and has extensive experience in regulating consumer harms.

The review has drawn on evidence from a range of sources, including extensive stakeholder consultation, public submissions, and available research. Ms Rickard was supported by a Secretariat team from the Department of Infrastructure, Transport, Regional Development, Communications and the Arts (the Department).

On 29 April 2024, an issues paper was released to support discussion and ask the community about protection from online harms, regulating the online environment and investigation and enforcement. The period of public consultation closed 21 June. More than 2,200 responses were received, including: 169 substantive submissions, over 500 short comments and more than 1,600 responses which were part of an online campaign conducted by the Free Speech Union of Australia.

Throughout the course of the review, Ms Rickard met with representatives from industry, civil society, governments and research bodies. Ms Rickard held a total of 72 meetings, seven of which were roundtables. Of those 72 meetings:

- 30 were with community or civil society organisations
- 22 were held with government and law enforcement agencies
- 18 were with tech and digital platforms industry; and
- 2 meetings were held with international government stakeholders (the UK and the EU).

A list of submissions received and stakeholder engagement is provided at Appendix C.

THE ONLINE SAFETY ACT IN 2024

02

The Act commenced in January 2022 to improve and promote the safety of Australians online. Australia has been world-leading in online safety regulation, including establishing the world's first eSafety Commissioner and the first takedown schemes to have seriously harmful content removed.

© Getty Images. Credit: Thurtell.

By way of a brief history, the Office of the Children's eSafety Commissioner was first established in 2015 under the *Enhancing Online Safety for Children Act 2015*. These reforms created the first ever takedown scheme to protect children from serious cyberbullying. In 2017, the remit was expanded to cover online safety for all Australians, the Office of the Children's eSafety Commissioner was changed to the office of the eSafety Commissioner, and the legislation was changed to the *Enhancing Online Safety Act 2015*. These amendments included powers for the Commissioner to lead, coordinate and advise on online safety issues to ensure Australians have safe, positive and empowering experiences online. The eSafety Commissioner also administered Schedules 5 and 7 to the *Broadcasting Services Act 1992* (known as the Online Content Scheme) which focused on specific types of harmful online material.

In 2018, a Statutory Review of the *Enhancing Online Safety Act 2015* and the Online Content Scheme was completed by Ms Lynelle Briggs AO. Following this review, the Australian Government made substantial reforms to online safety, and Parliament enacted the *Online Safety Act 2021*.

These most recent reforms to the Act strengthened the complaints and content-based schemes. An adult cyber abuse complaint scheme was introduced, adding to the child cyberbullying scheme, the non-consensual sharing of intimate images ('image-based abuse') scheme and the Online Content Scheme. The reforms saw the child cyberbullying scheme expanded to provide protections to children bullied in all online environments, not just social media. The image-based abuse scheme was also expanded to address the sharing and threatened sharing of intimate images without the consent of the person shown.

The Act introduced the first of its kind industry-based mechanism, requiring greater transparency from online service providers around efforts to support user safety. The Basic Online Safety Expectations set the Government's minimum safety expectations of online service providers, establishing a benchmark for online service providers to take proactive steps to protect the Australian community. While the Basic Online Safety Expectations do not impose a legally enforceable duty on service providers to implement the expectations, they are an essential part of driving transparency and accountability across online services.[8] These transparency measures drive improvements from industry by enabling the Commissioner to require service providers to report against these expectations through periodic and non-periodic reporting notices and determinations. The reporting requirement aims to boost the transparency of services and provide the Commissioner with a tool to hold services to account for the steps they take to keep Australians safe online.

The Act updated Australia's existing Online Content Scheme,[9] providing new powers to regulate illegal and restricted content, no matter where it is hosted.[10] The Online Content Scheme included in the 2021 Act provides for a broad spectrum of online services to be subject to industry codes or standards. The focus of the codes and standards is limited to illegal or restricted material (by reference to the National Classification Code). Under the Online Content Scheme, the Commissioner can register codes developed by industry bodies or associations representing sections of the online industry. Once registered, industry codes (and standards) are mandatory and enforceable, and take an outcomes-based approach to regulating Class 1 and Class 2 material.

The Act also empowered the eSafety Commissioner to promote and improve online safety for Australians through a range of education, research, coordination and advisory functions. The office of the eSafety Commissioner (eSafety) works with organisations and regulators domestically and around the world,[11] including as a founding member of the Global Online Safety Regulators Network (the only global forum currently dedicated to supporting collaboration between online safety regulators).

The Act is technology-neutral, focusing primarily on specific online harms rather than the means through which the harm was generated. eSafety's schemes and supports have clearly had much success in improving and promoting the safety of Australians online. In 2023-24, eSafety has reported that it received more than 4.7 million unique visitors to its website, more than 13,000 complaints to the child cyberbullying, adult cyber abuse and image-based abuse schemes and more than 13,000 complaints to the Online Content Scheme. However, the technology landscape

8 *Online Safety Act 2021*, section 45.

9 Previously provided for under Schedule 5 and Schedule 7 of the *Broadcasting Services Act 1992*.

10 A removal notice for Class 1 material may be issued no matter where the content is hosted. For Class 2 material, the service must be provided from Australia.

11 For more information refer to International engagement | eSafety Commissioner.

continues to change, with emerging technologies enabling new harms such as the use of smart technology and apps for cyberstalking or coercive control, the use of generative artificial intelligence to generate deepfake images or online bot attacks, and recommender systems exposing users to harmful online content. Changes in technology and the way people use them will continue to present online safety challenges, such as addictive design, self-harm and disordered eating, volumetric attacks, online hate and cyberstalking.

Below are some examples of online harms associated with emerging technologies.

Generative artificial intelligence

'Generative artificial intelligence' describes the process of using machine learning to create digital content such as new text, images, audio, video and multimodal experience simulations. Examples include:

- Text-based chatbots or programs designed to simulate conversations with humans; and
- Image, video or voice generators.

The rapid deployment of generative artificial intelligence and the scale and sophistication of content produced, have the potential to amplify online harms. This could be realised through algorithmic bias resulting from automated decision-making or exposure to discrimination and bias through the outputs of generative artificial intelligence tools. Online harm examples include chatbots providing inappropriate and harmful responses to user prompts, the spread of hyper realistic generative artificial intelligence deepfakes, and the creation of synthetic child sexual exploitation and abuse material.

Immersive technologies

Immersive technologies allow users to engage in a virtual world where users interact with each other in an immersive and interactive computer-generated environment. According to eSafety research in 2022, an estimated 680,000 adults in Australia may be engaging in the metaverse[12], with half of those interacting in these environments at least once a month. No doubt usage has increased since then. More than half of those engaging in the metaverse are using haptic technologies. Haptic technologies transmit tactile information through sensations such as vibration, touch and force feedback to enhance the user experience.

Virtual world interactions, especially when enhanced by haptic technologies, introduce the potential for new online harms that were previously limited to the physical world. In December 2023, the Digital Regulation Co-operation Forum (UK) identified potential issues arising from immersive technologies, including novel forms of harm and the convergence of immersive social media.[13] The potential for horrendous harms was demonstrated in January 2024 when UK police were reported to be investigating a case of a child whose avatar was sexually assaulted in an immersive video game.

Recommender systems and algorithms

Recommender systems prioritise content or make personalised content suggestions to online service users. Recommender systems and their underlying algorithms are built into many online services, sorting through vast amounts of data to present content that the system deems relevant to users.

For example, social media services use recommender algorithms to personalise what is suggested or promoted to users and to increase the reach of prioritised content and accounts. Online services use recommender systems to drive engagement and maintain 'stickiness', encouraging users to spend more time on the service. However, this can also create an incentive to promote content that may be harmful but attention-grabbing, amplifying mis/disinformation, extremist views and reinforcing perspectives the user may already be aligned with, creating echo-chambers.

End-to-end encryption

End-to-end encryption is a means of securing communications from one end point to another and an important defence against security breaches that would otherwise have serious consequences for online users. It transforms standard text, image, audio and video files, and live video streams, into an unreadable format while still on the sender's system or device. The content can only be decrypted and read once it reaches its final destination.

12 eSafety Commissioner (2023), *The Metaverse: a snapshot of experiences in virtual reality*.

13 Digital Regulation Co-operation Forum (DRCF) (2023), *Immersive Technologies Foresight paper*, Immersive Technologies Foresight Paper | DRCF.

End-to-end encryption is increasingly being adopted by services which offer messaging functions to consumers. However, it can also conceal harmful conduct or hinder investigation of the distribution of harmful and illegal online content such as child sexual exploitation and abuse material.[14] This is of great concern and it is imperative that a solution is found sooner rather than later.

Changes to technology models (decentralised platforms)

There is growing interest in developing decentralised online platforms and services (sometimes referred to as Web 3.0 or DWeb). Decentralisation has the potential to provide users with more power online by reducing reliance on mainstream, centralised servers and distributing responsibility for data sharing and storage to communities of users. Existing decentralised services include peer-to-peer services, blockchain-based services, and federated services that run on independent servers. However, decentralisation can also create online safety challenges. Within the current regulatory framework, decentralisation could make it more difficult to hold users responsible for illegal or harmful content and conduct.[15] Decentralised services may not have the size and systems in place to support users to make a complaint or undertake effective moderation of the content provided on the service.

Australians online

The pervasiveness of the online environment and the degree of online harm is a serious concern. The following research findings make this abundantly clear:

- Across the globe there are more than 5 billion active social media user identities[16]
- The average Australian spends about six hours per day online with nearly two hours on social media[17]
- 70 per cent of Australian adults have had at least one negative online experience in the previous 12-months to November 2022[18]
- 71 per cent of children aged 14 to 17 had seen sexual images online, but only 34 per cent of parents had an awareness[19]
- 31 per cent of Australian adults were sent unwanted inappropriate content, such as pornography and violent content[20]
- 72 per cent of Australians who used a dating app or website experienced sexual violence[21]
- Almost two thirds (61.7 per cent) of young Australians aged 12-18 said social media made them feel dissatisfied with their body – a 12 per cent increase since 2022[22]
- A recent large-scale survey of 16,693 adults across 10 countries, found that 14.5 per cent of all respondents had received threats that their intimate images would be shared, with victimisation rates higher among men, younger adults, and LGBTQIA+ adults[23]; and
- In 2023-24, eSafety received 13,824 complaints concerning 33,910 URLs with 82 per cent relating to reports about child sexual abuse, child abuse or paedophile activity.[24]

14 eSafety Commissioner (2023), *End-to-end encryption: Position statement*, End-to-end encryption trends and challenges — position statement | eSafety Commissioner.

15 eSafety Commissioner (2022), Decentralisation – position statement, Decentralisation – position statement | eSafety Commissioner.

16 Meltwater Press Release, 31 January 2024 Report: Global social media users pass 5 billion milestone.

17 Ibid.

18 eSafety Commissioner (2022), Australians' negative online experiences 2022, Infographic – Adults' online experiences | eSafety Commissioner.

19 eSafety Commissioner (2022). Mind the Gap: Parental awareness of children's exposure to risks online, Aussie Kids Online, Melbourne: eSafety Commissioner, 71.

20 eSafety Commissioner (2022), Australians' negative online experiences 2022, Infographic – Adults' online experiences | eSafety Commissioner.

21 eSafety Commissioner (2023), Technology-facilitated abuse: Family, domestic and sexual violence literature scan, Canberra: Australian Government.

22 Butterfly Foundation, Body Kind Youth Survey 2023.

23 Henry, N., & Umbach, R. (2024). Sextortion: Prevalence and correlates in 10 countries. Computers in Human Behavior, 158, 108298.

24 Australian Communications and Media Authority and eSafety Commissioner Annual Report 2023-24 ACMA annual report, 208.

OBJECTS OF THE ACT

03

The Terms of Reference asked me to consider the objects of the Act. The current objects, as set out in section 3 of the Act are:

- To improve online safety for Australians; and
- To promote online safety for Australians.

These have served us well though they don't cover one extremely important group, children based overseas who are the victims of online child sexual exploitation and abuse that is viewed by online users in Australia.[25]

There are a number of ways to approach objectives: Very broad goals, as we have now, or more descriptive objectives that include links to the various functions covered by the Act. On balance, I find the second approach more useful.

I would suggest that the current objectives are replaced as follows:

The objects of this Act are to enhance the online safety of Australians and Australia by:

1. **Promoting human rights and safety.** (This would centre eSafety's work in a human rights framework as occurs in the European Union and United Kingdom.)
2. **Promoting and protecting the best interests of the child.** (This would highlight the importance of protecting the best interests of all children wherever there is a link to Australia.)
3. **Building the evidence base around online safety and existing and emerging online harms.** (This recognises the important work eSafety does in keeping up with what is happening with technology and identifying new and emerging harms with the goal of preventing them occurring. It also helps with setting priorities.)
4. **Preventing and alleviating online harm present in Australia.** (This goes to eSafety's core functions of education, awareness raising and administering its regulatory schemes, including through safety by design, systems focus, and through enforcement action); and
5. **Improving online safety for all in Australia by advancing:**
   a. Service provider responsibility for preventing harms and mitigating the damage done
   b. User empowerment
   c. Transparency.

01

Recommendation 1:

That the objects of the Act should be amended to include more descriptive objectives that are linked to the various functions covered by the Act.

25 Review submission 95 - International Justice Mission, 13.

WHO SHOULD BE REGULATED?

04

The Act specifies sections of the online industry, including for the purposes of industry codes and industry standards. Eight sections of the online industry are currently specified, enabling codes to be developed by, and applied to, relevant sections of the industry.

The Commissioner can direct requests for codes to representatives of relevant sections:

- Social media services
- Relevant electronic services
- Designated internet services
- Internet search engine services
- App distribution services
- Hosting services
- Internet carriage services; and
- Those who manufacture, supply, maintain or install relevant equipment.

During the course of the review, I repeatedly heard that the description and treatment of industry sections in the Act is complex and not fit for purpose. The categorisation of the industry sections suffers from two major issues:

- Categories are too narrow – in practice a service may perform multiple functions, span different categories and what they cover changes; and
- Categories do not support a risk-based and proportionate regulatory approach – a service's size and risk level does not impact how it is categorised.

4.1 Existing sections of the online industry are narrow and inflexible

The existing definitions, particularly social media services, relevant electronic services, and designated internet services do not match the types of services present in the online ecosystem. This creates uncertainty as to which category a service should fall within in order to comply with industry codes, particularly where a service has multiple functions and features. Throughout the review, I have heard that the defined sections of industry are complicated and confusing, making it difficult to understand how the Act applies across the online industry. For example:

- Many social media services incorporate direct-messaging features of relevant electronic services, and are often used for that purpose. This is a problem under the existing definition of social media service, which specifies that enabling online social interaction must be the "sole or primary purpose" of a service
- A number of messaging services (relevant electronic services) now incorporate elements of social media services, such as the ability to create large communities and groups
- Adult content platforms (generally considered designated internet services), increasingly incorporate features for user-generated content, social interaction and messaging commonly seen in social media services and relevant electronic services; and
- Some search engine services incorporate features such as generative artificial intelligence which amount to the direct provision or generation of content appropriate to a designated internet service.

Deciding which definition of the Act to apply is challenging for services with multiple functions and features. Some have argued these decisions can be subjective and imprecise. A number of submissions also highlighted the difficulties this causes in determining a service's place in the regulatory regime and associated compliance obligations.

4.2 Amending industry sections to better reflect a risk-based and proportionate regulatory approach

My preferred approach is to apply broader categories that better allow for risk assessment and mitigation obligations based on a service's specific mix of functions and features. The decision on which entities are captured by the Act and hold associated obligations must serve two fundamental purposes.

All parts of industry which may facilitate harm, or which can play a role in mitigating harm, must be captured by the Act. The Act has evolved primarily to support a content-reporting model of regulation, where harmful content reported to eSafety is subject to removal or remedial powers. Under this model it makes sense to keep services where Australians find harmful material in scope, and to provide flexibility for the regulator to take action. Australians must be able to report material such as image-based abuse to eSafety no matter the service it is posted on, and to have eSafety act, regardless of the size or profile of the service it is posted to. It is also important that the Act cover not merely those services where harmful content and conduct occurs, but also services which may facilitate those harms and have a role to play in mitigating them. This includes the platforms on which the harms occur, and the services providing hosting, access and other functions that provide infrastructure or facilitate access to a service.

There needs to be greater flexibility within categories to better reflect a service's risk and reach, where risk factors may vary and reach affects a service's risk. If a duty of care, imposing stronger proactive systems obligations, is adopted, industry participants will require more guidance on how and to what extent these obligations apply to their services. While some effort has already been made by industry and eSafety to "apply some risk-based differentiation between services" in the development of industry codes and standards, risk-based and proportionate criteria for obligations should be "embedded … at a statutory level".[26] The Act should provide assurance that factors such as a service's size, the risks posed by its functions or features, and its usage by Australians are the critical factors when determining a service's obligations.

4.3 Services can be categorised more clearly and simply

Many submissions have argued for a more 'technology neutral' approach to defining industry sections. A duty of care model would be best supported by a more encompassing and simplified approach that reflects a broader set of categories covering the online industry. The preferred outcome is one that applies to all services that provide regulated material or activity – whilst providing for a flexible and proportionate application in risk assessment obligations and codes.

Coverage of the online industry should at a minimum follow these principles:

- Cover all online products and services that provide, generate or facilitate in-scope content or conduct
- Allow for fair but flexible tiering according to 'reach or risk' so that services are aware of their status and obligations under the Act to allow for regulation that is properly targeted and proportionate; and
- Future-proof and flexible to allow for the emergence of new service types to adapt to change.

The Act should be amended to simplify the categorisation of service types, according to the role they play in the online ecosystem – and their relation to the regulatory powers of the Act.

Four broad categories could be the basis for a simplified structure:

1. **Online platforms (services providing online interaction and online content)**
2. **Online search and app distribution services (services which gate-keep access to online platforms)**
3. **Online infrastructure services; and**
4. **Equipment and operating system services (including manufacturers, suppliers, maintenance and installers).**

These broader categories better support a statutory duty of care model, with online platforms posing the greatest direct risk to users, being the primary services on which harmful content and conduct occurs. However, other categories also have an important role to play, such as when they 'gate-keep' access to and facilitate the operation of services where harm primarily occurs. The categorisation of services would be detailed in legislation (with a requirement for the regulator to set out clear guidance on the risk reporting cycle). The proposed category of online search and app distribution services can also work for the existing powers under the Act, such as the removal of apps or links by the app stores and search engines, or the blocking of websites by internet service providers.

26 Review submission 157 - Communications Alliance, 8.

Online platforms

This first category includes services where the majority of harmful content or conduct occurs. It would capture those services that enable:

- Online social interaction or messaging; and/or
- The provision of online content – including content that is user-generated, directly provided or generated by a service, or recommended by a service (such as by an algorithm).

This category covers services which are currently treated separately under the social media, relevant electronic and designated internet service categories in the Act. However, it also covers other kinds of services to the extent that they incorporate these features. This might include, for example, online search services which, in addition to indexing and ranking search results for other services, provide or generate content on their own service (using generative artificial intelligence, for example). While online search services are also considered separately, to the extent that they facilitate or generate content or conduct on their own services, they should be treated the same as other online platforms.

Online search and app distribution services

The second category consists of services which curate and enable access to specific services in the first category. These are largely gate-keeping functions and include services such as:

- App stores, which curate and facilitate access to social media, messaging, and other relevant services; and
- Search engines, which index online services and content for ranking and recommendation to users upon request.

These services don't generally function as social or content platforms in their own right, but rather allow users to discover and access these platforms. They are an important gateway for accessing services which facilitate online activity and content provision.

From a regulatory perspective, such services have an important 'gate-keeping' function in limiting access to harmful apps or services. For example, through having appropriate controls and filters for accounts or devices used by children, removing dangerous apps from their store, or removing harmful sites from search results.

Online infrastructure services

Similarly, services in this third category would be expected to take actions regarding illegal content when it is reported to them, and to comply with lawful requests and notices from the regulator. All online services depend upon 'infrastructure' services to allow them to be securely located and accessed online, including:

- Hosting services, which provide 'real-estate' (both physical and virtual) for online services – such as the storage of a service's data on physical servers
- Domain name services and registrars, which provide services with an 'address' – enabling users to easily access the service
- Internet service providers, which provide a 'transport' service to users, allowing them to access services online; and
- Other services providing infrastructure support and security. This could include services providing content delivery network services, virtual private networks, distributed denial-of-service (DDoS) attack protection and other cybersecurity support.

Services should undertake proactive and ex ante steps where possible, but may not always be able to proactively monitor or moderate the content or activity on services for which they provide infrastructure support. However, once made aware of illegal or seriously harmful content and activity on a service they enable access to, they are generally capable of removing this material (in the case of hosts) or blocking access to the service (in the case of internet service providers and domain name system services). It is important that these services are defined in the Act, to cover all infrastructures services to the extent they support services which are provided to end-users in Australia. This would cover, for example, hosting service providers which host services provided to end-users in Australia, regardless of where the hosting service is located.

Equipment and operating system services

This category would include devices and equipment that enable access to online services, enhance a user's online experience or are linked to an online service. For example, this could include:

- Devices which enable and facilitate the use of online service, such as phones, tablets, laptops/PCs, Smart TVs and wearables
- Novel equipment with emergent or distinct risk features, such as virtual reality interfaces, augmented reality glasses or haptic suits; and
- Operating systems used by various devices and equipment.

As with the current definition under sections 134(h) and 135(h) of the Act, this industry category would also include services responsible not merely for the manufacture, but also the supply, maintenance or installation of these equipment and operating systems.

In order to future-proof the Act and ensure that services and products are designed with safety at their core, the definitions must be broad enough to capture the full suite of vectors/opportunities where harm can occur. These services should be specifically covered by the regulatory framework.

02

Recommendation 2:

That current definitions of the online industry sections should be simplified to online platforms, online search and app distribution services, online infrastructure services and equipment and operating system services. These should be included in the Act to better reflect online safety risks and future proof the Act.

© Getty Images. Credit: Master.

4.4 Tiering takes a proportionate approach to regulating online services

A duty of care should apply to all services where there is a risk of foreseeable harm and the ability to exercise a degree of control over the environment – including control over the provision of content, facilitation of conduct and contact, and the mitigation of risks and harms relating to these. However, the application of proactive regulatory obligations under the Act (such as risk assessments and transparency reporting) should be tiered and proportioned according to clear risk-based criteria and a transparent process of designation by the regulator. This is an issue that has been raised by many submissions.

The two general criteria recommended for determining a service's level of obligations under the Act are 'reach' and 'risk'. Reach refers to the level of active usage of a service in Australia over a specified period – its level of popularity among Australian users. This is important as the widespread use of a service is a useful proxy for risk, with a large number of Australians potentially exposed to harmful content or activity on that service. Services whose number of Australian active end-users (averaged over a specified period) exceeds a threshold percentage of the population would be presumed to fall under the Act's most stringent level of obligations. The European Union, for example, sets the bar at 10 per cent or more of their population using the service. This would be a useful default threshold to employ in Australia, provided the regulator also has discretion to designate in services with lower reach which nevertheless pose a sufficiently high risk.

4.5 There are risks in only focusing on the size of a service

A service's reach should not be the only indicator of risk to Australian users. There needs to be flexibility to take other risks into account and to designate certain services to the highest level of obligations under the Act or remove safer services from the group with the most requirements placed on them.

Transparency will be critical to this process. The Act should prescribe a transparent process for determining and applying these factors to services or classes of services, and should provide some guidance on what these risk-factors are. However, there should also be flexibility to determine additional factors where evidence for them exists, to keep the Act up to date. In setting out the duty of care regime, certain risk factors should be considered, including:

- **A service's functions and features:** different functions or service features pose different levels of risk. For example:
  - Features which increase the propensity for harmful content to 'go viral' across a platform
  - Algorithms which recommend potentially harmful content
  - Direct (and especially ephemeral) messaging, especially where enabled between users from outside a person's contacts or immediate circle
  - End-to-end encryption
  - Immersive environments
  - Generative artificial intelligence; and
  - Friend recommender systems and user discoverability features which could connect children with paedophiles or sextortion scammers.

- **Whether or not a service is likely to be accessed by children,** determined by factors such as:
  - Whether the content or features of a service are likely to appeal to children (irrespective of whether they are targeted at children); and
  - What systems and processes are in place for preventing access by children to the service.
- **Whether a service especially enables illegal or harmful content and activity,** such as:
  - Child sexual exploitation and abuse material, terrorist and violent extremist content, or otherwise criminal content
  - Pornographic material that can be accessed by children
  - Material that can be accessed by children which encourages or instructs in self-harm, suicide and disordered eating; and
  - Online hate and harmful abuse targeting Australians.
- **Whether a service has deficiencies in its terms of service or enforcement of those terms.**
- **Aspects of a service's governance and ownership** which may raise or lower its risk level or reasonable obligations, such as:
  - Large platforms with opaque or less accountable governance or ownership structures (higher risk)
  - Platforms that make insufficient investment in safety systems, processes and staffing (higher risk)
  - Platforms that lack a complaints system or safety reporting features; and
  - Service providers who have less liability for, or control over their services because they are contracted to other providers (such as government or business enterprises) that hold primary responsibility (lower obligation). Such service providers are still responsible for the safe design of their services, but may have lesser obligations with regards to active monitoring and enforcement of the safe use of their services. The exception of course is where such arrangements are made with a deliberate intent to evade responsibility or obligations.
- **Whether and to what extent the use of a service is primarily governed by other legal frameworks,** such as:
  - Streaming video-on-demand services subject to the National Classification Scheme, which does not allow user-generated content;
  - Services used by government or business enterprises subject to workplace safety laws; and
  - Messaging protocols such as short messaging services (SMS), subject to telecommunications laws.
- **The type and extent of online harm experienced** on a service or through a service's intermediary or gate-keeping function is an indicator that the service presents an increased risk to Australians and should be formally designated in to enforceable obligations. This could be informed by transparency reporting or complaints made to the regulator.

4.6 Designating services and imposing obligations must be transparent and accountable

The regulator should be empowered to make determinations regarding the level of obligations of specific services or classes of services, subject to:

- Clear criteria set out in the Act
- Consultation with the public and affected services; and
- Parliamentary disallowance.

© Getty Images. Credit: Andreajane.

4.7 Providers of online safety technology should be recognised

Providers of online safety related services and technology (such as providers of content-filtering for schools or parental controls) are not currently recognised in their own right in the Act's identification of industry sections responsible for online safety. This is a potential issue, in particular when it comes to the handling of illegal material such as child sexual exploitation and abuse material, which these services may collect in the course of performing their functions. Under section 474.24 of the Criminal Code (Cth), a defence is available for a person who engages with such material if the conduct is in good faith and for the sole purpose of:

- Assisting the eSafety Commissioner to perform functions or exercise powers conferred under Part 9 of the *Online Safety Act;* or
- Manufacturing or developing, or updating content filtering technology in accordance with an industry code or standard under Part 9 of the *Online Safety Act*.

As providers of online safety technology are not defined in this Part, where they are independent of the broader industry sections defined in that Part of the Act, it is not clear whether they are currently able to operate lawfully in helping to filter or control the distribution of child sexual exploitation and abuse material.

03

Recommendation 3:

That the Government consider options to recognise the role of providers of online safety related services and technology in helping to identify and stop the distribution of child sexual exploitation and abuse material.

INDUSTRY'S DUTY OF CARE

05

As noted in my introductory remarks, my goals with this review are to keep Australians safer online, provide services with the incentives to do the right thing, future proof the Act, and align the Act with emerging international best practice. It is clear that the current laws and regulatory settings are not sufficient to deliver on these goals and address the volume of online harm that is occurring.

The phrase I heard most during consultations for this review is 'whack a mole'. This refers to the important work eSafety does in expeditiously removing harmful content from services.

While there is absolutely a place for complaint and takedown schemes as a safety net, the scale of online activity is such that the only way we can achieve real safety improvements is to take a systems approach to regulation.[27] We need service providers to design their systems with safety at their core and quickly identify and remediate problems when they emerge.

A common theme in submissions to this review has been the need for a more systemic and preventative approach to online harms. Something is needed to shift the onus of responsibility for reducing harm from individuals – who are at a disadvantage in terms of their power and available information – onto online service providers who exercise the most control over their design and operation. Systems-based approaches are increasingly being adopted or considered internationally. In the Australian context, this could be achieved in several ways:

- **Making the Basic Online Safety Expectations enforceable.** Currently, services are only required to report on if and how they are meeting these expectations in response to a notice from the Commissioner. A positive and enforceable requirement to meet them would shift the regulatory framework further towards a systemic approach.
- **Instituting a duty of care and due diligence requirements.** These concepts are similar, and can be combined in practice. A duty of care is a positive requirement on service providers to take reasonable steps to prevent foreseeable harms on their services (the approach proposed in and adapted by the United Kingdom). Under this approach a singular and overarching duty of care could be established, or (as in the United Kingdom) multiple duties to address different categories of harm, risks or affected persons. A due diligence approach requires service providers to have robust processes in place to manage the risks of harms on their service (the approach taken in the European Union).

27 In 2024, Statista forecast that the average number of data interactions per connected person would rise to almost five thousand interactions per day by 2025. Statista (2024), Daily digital data interactions per connected person worldwide from 2010 to 2025, Daily data interactions per connected person 2025 | Statista.

5.1 Lifting expectations to obligations

The Basic Online Safety Expectations (the Expectations) along with enforceable requirements established through the Online Content Scheme's codes and standards encompass the Act's current systems-based approach. The Expectations set out a multitude of important expectations that service providers should meet to keep people safe. Examples of the Expectations include that the provider of the service will:

- Take reasonable steps to ensure that end-users are able to use the service in a safe manner.
- Take reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful.
- Take reasonable steps to ensure that the best interests of the child are a primary consideration in the design and operation of any service that is likely to be accessed by children.
- Ensure that the service has terms of use, policies and procedures in relation to the safety of end-users, standards of conduct for end-users, and that these will be enforced; and
- Have clear and readily identifiable mechanisms enabling Australians to report and make complaints about illegal and harmful material under the Act or a service's own terms of use.

These expectations have been instrumental in building transparency over the actions of industry.

There are two clear benefits to the Expectations. First, services know what they are expected to do. Second, the Commissioner can require them to complete transparency reports, with the ability to ask forensic questions about what they are and aren't doing to meet the expectations. This gives the Commissioner, government, experts and the public a very clear understanding about who is and isn't doing enough, and the ability to call out services that are falling short.

There is, however, one major flaw with the Expectations – they are not enforceable. Some submissions have supported making the expectations mandatory and enforceable.[28]

One option for a systemic approach could be to make the Basic Online Safety Expectations enforceable, with strong penalties (see Chapter 10). This would be a considerable improvement, but with new harms and developments in technology constantly emerging there are likely to be gaps in the coverage of the expectations at any point of time. Keeping expectations up to date would depend on the Minister (who makes the expectations by legislative instrument) identifying the new harms and practices and updating expectations – even though it is the service providers that would likely be the first to see the new harms. While some services may take quick action, experience suggests it is unlikely that all will. This would lead to ongoing lags in protection and the need to regularly update Basic Online Safety Expectations.

For this reason, making the Expectations enforceable is not my preferred option. The Expectations were an important first step towards systemic regulation that we can retain and improve on in other, better, ways.

28 For example, submissions from the AFL, Tasmanian Government, the Alannah & Madeleine Foundation, Our Watch, the Uniting Church, Allies for Children, Relationships Australia and the International Justice Mission.

5.2 Australia should adopt a systemic duty of care to prevent online harm

An overarching duty of care approach places responsibility on service providers **to take reasonable steps to address and prevent foreseeable harms on their services**. It shifts much of the burden for remaining safe online away from individual users to those most capable of identifying and addressing harms – the service providers themselves.

The case for a duty of care in the online safety context was first prosecuted by Professor Lorna Woods and William Perrin OBE. They noted that the online world is a space "where so many different things happen that you would be unable to write rules for each one".[29] The specific harms that are foreseeable, and the steps that count as reasonable, will inevitably change over time as services, their underlying technologies, and users change.

Online service providers exercise control over, and have knowledge of, their systems and processes, and must be held responsible for the safety of their services. Through design and moderation decisions, online services "intentionally or not" have "an important impact on content and the risks that materialise on their services".[30] A duty of care recognises that what happens on services is in no small part the result of those design and business decisions. As Reset.Tech argue in their submission, "[f]ocusing on design and operation is important because despite their names 'platforms' are not entirely neutral, passive transmitters when it comes to content. Intentionally or not, their choice architecture has an impact on content. This includes the role of recommender and content moderation systems."[31]

Across the globe, online safety regulation is still relatively new. Australia has largely focused on the removal of specific harmful content or activity that has occurred online through a range of what have been very effective takedown powers. The Commissioner's powers to have harmful material removed have been world leading, and the importance of these powers is discussed in Chapter 7. However, with the pervasiveness of the online world and the emergence of new risks, a different regulatory response is needed.

A duty of care shifts the emphasis of regulation from reactively tackling specific pieces of material to remediate the harm, to taking a preventative and systems-based approach. The historical focus on content and its removal creates significant regulatory gaps in the Australian online safety framework. In the current context, content is only part of the online experience and one of multiple vectors for online harm. There are other vectors of harm not properly captured by a focus on content, including contact and conduct. Contact harms are those which occur "as a result of online interactions with others", and harmful conduct refers to the harmful behaviours that are facilitated by technology.[32] While contact and conduct harms may result in harmful content, a focus on content (as important as it is) may on its own fail to address other harms arising from them or fail to properly prevent the contact and conduct factors which lead to the posting of harmful content.

A duty of care would shift the focus to proactive and systemic measures and, where possible, to prevention. Preventing harm is always preferable to addressing harms once they have occurred. This approach is "risk-based and outcomes-focused"[33] and is expected to have a far greater impact on improving safety than what can be achieved through a reactive model, such as a complaints-based regulatory system.

29 Woods, L. & Perrin, W. (2019), Online harm reduction – a statutory duty of care and regulator, Carnegie UK Trust, April 2019, 28, https://carnegieuktrust.org.uk/publications/online-harm-reduction-a-statutory-duty-of-care-and-regulator/

30 Review submission 70 – Reset.Tech Australia, 4.

31 Ibid.

32 World Economic Forum (2023), *Toolkit for Digital Safety Design Interventions and Innovations Typology of Online Harms, Insight Report*, August 2023, 5.

33 Woods, L. & Perrin, W. (2019), 5.

5.3 Online services must make continuous and ongoing efforts to improve safety

Similar to work health and safety legislation, which has withstood the test of time, compliance with a duty of care would require taking active steps to prevent or mitigate the impact of foreseeable harms. A duty of care would require services to:

- **Engage in an ongoing cycle of risk identification and assessment**
- **Undertake safety by design and risk mitigation; and**
- **Seek to measure the impact of mitigation**
- **Perform transparency reporting.**

This approach combines a duty of care, with a due diligence approach that requires services to consider safety in the design process and at every subsequent stage, as well as in their corporate culture. Embedding this cycle into a service's operations would ensure the consideration of safety in the design and deployment of all services and features. To support this process, an effective transparency and accountability scheme would provide confidence that online services are operating according to their published safety objectives or expose where they are falling down.

It is also an approach that can deal with technologies and harms not yet dreamed of. It can help future proof regulation. Algorithms, recommender systems, addictive design, artificial intelligence, and generative artificial intelligence, business decisions and more are all factors that shape an individual's online experience and have the potential to create significant harm. A duty of care can capture these factors, as well as technology that emerges in the future, to ensure ongoing effective online safety regulation.

© Getty Images. Credit: Paweł Wewiorski.

5.4 Global efforts are now focussed on a systems-based approach

Major online service providers largely deliver the same service to users in multiple countries, but are faced with a fragmented and nation-specific regulatory environment. Greater consistency across our respective national regimes would simplify compliance for service providers, reducing costs and regulatory burden. This would also provide economies of scale and more coordinated and efficient investments in safety.

With most online service providers based outside Australia, there are challenges with compliance and enforcement. While these challenges are shared by many others, jurisdictions with larger markets, like the United Kingdom and European Union, are better positioned to take on the digital industry and enforce their laws. Where appropriate, aligning Australia's approach to other like-minded countries could strengthen Australia's position and lead to better regulatory outcomes in Australia.

The trend is clear, our major international counterparts in Europe, the United Kingdom and North America are almost all moving towards a systems-based, proactive approach. While there are differences in approaches, the overall objective remains the same: services must take reasonable steps to keep their users safe.

Laws that have been enacted by other jurisdictions in recent years include:

- **The European Union's Digital Services Act**, which establishes a due diligence framework to be adopted by all member states. In reality this is very similar to a duty of care. It requires very large online platforms or very large search engines to undertake risk assessments of systemic risks arising from the design or functioning of their services, including algorithmic systems, or from the use of their services. Assessments must include consideration of illegal content, risks to fundamental rights, civic discourse and electoral processes and public security, risks of gender-based violence, to public health, children's wellbeing and serious negative consequences to peoples' physical and mental wellbeing.
- **The United Kingdom's Online Safety Act 2023**, which establishes positive statutory duties of care for a wide range of online services. Providers of regulated services are required to identify, mitigate and manage the risks of harm from illegal content and activity and content and activity that is harmful to children. All providers of regulated services have duties relating to illegal content, including to conduct risk assessments and take proportionate service design and operational measures to prevent or minimise the risk of users encountering illegal content, and to mitigate and manage the risks of harm to individuals. Services likely to be accessed by children, and services that provide pornographic content have additional duties.

Bills currently under consideration in other jurisdictions include:

- **Canada's Online Harms Bill**, which would establish positive statutory duties to act responsibly, protect children and to make certain content inaccessible. Under the duty to act responsibly, a broad range of harmful content is captured, including: intimate content communicated without consent, content that sexually victimises a child or re-victimises a survivor, content that induces a child to harm themselves, content used to bully a child, content that foments hatred, content that incites violence, and content that incites violent extremism or terrorism.
- **The United States' Kids Online Safety Act**, which would establish a duty of care for social media services in relation to children, requiring them to prevent and mitigate harms including violence and harassment, sexual exploitation and abuse, promotion of narcotics, alcohol, tobacco or gambling, promotion of dangerous acts likely to cause serious harm or death, and compulsive usage.

Measures which maximise convergence of regulation between countries help to maximise the potential for securing and enforcing extra-territorial compliance.

5.5 An overarching duty of care is preferable to multiple duties

A singular and overarching duty of care for online safety was originally conceived by Woods and Perrin,[34] and was proposed in the UK government's 2019 Online Harms White paper. In this proposal, the singular duty would be underpinned by codes of conduct that set out how to comply with the duty of care regarding particular harms. Platforms could comply with the codes or explain to the regulator why they were taking a different approach. Importantly, as noted in Carnegie UK's model Online Harm Reduction Bill (which was a major influence on the original UK Online Safety Bill), the "statutory duty of care regime … [does] not critically depend upon the existence of codes of practice, which can take years to formulate and adopt. The regime is in operation from the time that the Act comes into force".[35]

However, by the time the UK Online Safety Act had made its way through the parliament it had been enacted with a set of duties rather than an overarching duty of care. It was a much narrower Act than originally envisaged. The new approach was criticised by Carnegie UK as promising to be only "partially effective", and they strongly recommended a return to a "general duty of care" to "orientate and to give coherence to the regime."[36]

The UK is now in the process of drafting detailed codes for each of these duties, which has added significant time to the implementation of their Act. While time will tell how effective the UK approach is, it is clear that as new harms arise it will challenge the UK regulator to respond swiftly as significant work and legislative amendments will be required to address the harm before new codes come into effect.

I think it is in Australia's best interests to introduce an overarching duty of care rather than multiple duties. As Reset.Tech Australia notes in its submission:

… it is unclear … how the UK OSA [online safety act] is going to address harms arising from overarching abusive designs that do not fall to a particular sort of content, such as dark patterns that deceive users or extended use design techniques deployed at children.[37]

They also note that multiple duties would introduce an unusual paradox. While a singular duty of care "acknowledges that systems are developed and business decisions are made before platforms are populated with content" encouraging them to "safeguard their systems before harm happens",

[I]mplementing duties of care tied to particular sorts of content, requires platforms to risk assess their systems after they are 'populated' with designated content, or after harm has happened. This seems at odds with the sort of 'upstream' and systemic approach that a duty of care enables.

Implementing duties of care rather than a singular duty of care moves the regulation away from a focus on the systems and back to specifying particular types of content. This skews the focus of compliance towards a content-first rather than a systems-first approach.[38]

Taking a proactive and systems-based approach was supported by many of the public submissions provided to the review. While not all submitters commented on whether a duty of care should be singular or multiple, many did such as Reset.Tech Australia, The Victorian Bar and the Law Society of NSW (in the submission of the Law Council of Australia).

34 Woods, L. & Perrin, W. (2019).

35 Carnegie UK Trust (2019), Draft Online Harm Reduction Bill, https://carnegieuktrust.org.uk/publications/draft-online-harm-bill/.

36 Carnegie UK Trust (2021), Evidence to Joint Committee on the Draft Online Safety Bill, September 2021, 2,8.

37 Review submission 70 – Reset.Tech Australia, 5.

38 Ibid.

5.6 Enduring categories of harm to strengthen the attention given to them

As Woods and Perrin note, when Parliaments set out a duty of care they "often set down in the law a series of prominent harms or areas that cause harm that they feel need a particular focus, as a subset of the broad duty of care. They may link the harms to specific groups of persons to whom a duty of care is owed".[39] There is merit in doing this for any overarching duty of care in Australia, in order to highlight for industry and the regulator the categories of harms that require particular attention.

Outlined below is a recommended list of harms to focus on. While there will inevitably be some overlap, categories will help with clarity. The examples may change over time but the headline areas, sadly, are unlikely to. As well as covering users of a service, these must also cover people who are not users of the service, but who are impacted by the service, such as people whose intimate images are shared on a service without their consent. The harms that should be highlighted for attention in reforms to the Act should at a minimum include:

- **Harms to young people** including child sexual exploitation and abuse (including grooming), bullying and problematic internet use
- **Harms to mental and physical wellbeing** including threats to harm or kill, or attacks based on a person or group of people's protected characteristics, such as sex, gender, sexual orientation, race, ethnicity, disability, age or religion[40]
- **Instruction or promotion of harmful practices** such as self-harm/suicide, disordered eating and dares that could lead to grievous harm
- **Threats to national security and social cohesion**, such as through promotion of terrorism and abhorrent violent extremist content; and
- **Other illegal content, conduct and activity.**

It is worth noting that many of the harms falling under these categories are already illegal. However, law enforcement agencies often do not have the resources or sometimes the expertise to police these harms on their own when they proliferate online. Incorporating them under a duty of care would help to enlist the resources and expertise available to online service providers in the effort to better contro the spread of illegal online harms. It would also enable the eSafety Commissioner's civil-based schemes to continue working as a complementary measure to criminal justice responses.

04

Recommendation 4:

That Australia adopt a singular and overarching duty of care that encompasses due diligence, and is underpinned by safety by design principles, risk assessment and mitigation.

39 Woods, L. & Perrin, W. (2019), 35.

40 Review submission 70 – Reset.Tech Australia, 7.

Case Study: Disordered eating content on Instagram

'Sparked by schoolyard bullying and fanned by Instagram. That's how Robb Evans sees his late daughter's battle with anorexia nervosa, which ended in tragedy in April 2023 … While the Victorian dad still searches for clarity on what his daughter was exposed to, the posts she was comfortable showing him included advice on masking illness with water and clothing. "It got more sinister in how few calories could you consume in a day to live," he said. "She was searching for this content and then being presented with more and more of it." Through his grief, Mr Evans has thrown his energy behind a campaign to force teenagers under the age of 16 off social media. Meta and TikTok allow children as young as 13 onto their platforms, though each cannot easily verify ages … Eating disorder experts say teenagers need more time to develop without the influence of social media and want a ban for under-16s among other changes … "Due to the loop of content reinforcing appearance ideals, control of eating etc. the algorithm can reinforce challenges related to the development of an eating disorder and treatment seeking and recovery." … Teens who search for content related to eating disorders or body image issues now see a pop-up with tips and an easy way to connect with support organisations such as the Butterfly Foundation … '

The Canberra Times, Thursday 15 August, 2024[41]

41 Father of late teen leads call for social media change' – The Canberra Times, Thursday 15 August 2024

05

Recommendation 5:

The harms that should be highlighted for attention under a duty of care should at a minimum include:

- Harms to young people including child sexual exploitation and abuse (including grooming), bullying and problematic internet use
- Harms to mental and physical wellbeing including threats to harm or kill, or attacks based on a person or group of people's protected characteristics, such as sex, gender, sexual orientation, race, ethnicity, disability, age or religion
- Instruction or promotion of harmful practices such as self-harm/suicide, disordered eating and dares that could lead to grievous harm
- Threats to national security and social cohesion, such as through promotion of terrorism and abhorrent violent extremist content; and
- Other illegal content, conduct and activity.

A comprehensive approach to online harms

In the European Union and the United Kingdom, the types of harms that are captured include financial harms (such as scams), societal harms (such as contributing to misinformation by diluting the availability of public interest journalism) and national security harms (such as misinformation or disinformation campaigns during elections).

While these harms all sit well within a duty of care model, under Australia's current administrative arrangements they are dealt with by other existing or proposed legislation and by other regulators.

When the Online Safety Act is next reviewed it would be worth considering if all of these areas should be brought together under an expanded Online Safety Commission with a much broader mandate.[42] However, given the scope of change proposed within this report, and that a duty of care is a significant change, now is not the time to consider whether such sweeping changes should be made.

42 I note the second interim report of the Joint Select Committee on Social Media and Australian Society recommends the Australian Government establish a Digital Affairs Ministry which would have overarching responsibility for the coordination of regulation to address the challenges and risks presented by digital platforms, including matters such as privacy and consumer protection, competition, online safety and scams.

5.7 Risk assessment

An essential part of meeting the duty of care for online service providers would be to undertake regular risk assessments of their services. Risk assessment requirements are a core feature under both the European Union's Digital Services Act and the United Kingdom's Online Safety Act, and are built into the first phase of Australia's industry codes and standards. They are at the heart of a preventative and systemic approach to making the online world a safer place by design, and by working to prevent harms rather than merely responding after the fact. As the saying goes, 'it is better to put a fence at the top of the cliff, than an ambulance at the bottom of it.'

In taking reasonable steps to prevent foreseeable harms on their services, service providers would be required to:

- Identify the relevant online harms and their practices that may contribute to harms
- Assess the risk of those harms arising on their service, and how their design and operational decisions affects this risk
- Decide and implement the measures they need to take to mitigate or repair these risks; and
- Measure, review and report on the effectiveness of these steps.

As each service differs in its design, operation and usage, the risks and required measures will also be different. Each service provider must proactively assess the risk factors and contributors on their services.

All service providers should diligently perform risk assessments and implement mitigations (which include safety by design principles), both at regular intervals and when introducing or significantly altering products or features. However, stringent and enforceable risk assessment requirements should particularly be placed on the large services with high 'reach', and other services designated by the regulator as posing a high risk – according to reach and risk criteria. Consideration should be given to alignment with other relevant risk assessment frameworks to reduce regulatory burden on entities.

Risk assessments are necessary in many areas of life, however they are particularly important for online services where key decisions and design choices are not always transparent. This is especially the case as technologies like artificial intelligence – which are powerful but opaque in many ways – increasingly influence how services operate.

06

Recommendation 6:

Entities with the greatest reach or risk should be required to complete a risk assessment at least every 12 months and to carry out a risk assessment when significant changes are made to the design and operation of their service. These entities should also be required to provide an annual report detailing their risk assessments, risk mitigations and how successful they have been to the regulator.

07

Recommendation 7:

Services used by more than 10 per cent of the Australian population should be automatically part of the highest tier with additional mandatory responsibilities. The regulator should have a power to deem whether other online services do, or do not, meet the reach or risk requirement, noting that the reach or risk of services may change over time.

Risk reporting obligations must capture the whole risk assessment cycle

Two essential components of any risk assessment process are risk assessment and risk mitigation: to assess the risks and their potential impacts, and to develop and implement measures which mitigate or repair these risks. This is reflected, for example, in Articles 34 and 35 of the European Union's Digital Services Act, in the Phase 1 codes and standards under Part 9 of the Act, and in reasonable steps for meeting expectations in the Basic Online Safety Expectations. In addition to this, a risk assessment obligation should also mandate review and evaluation of all significant mitigation measures and reporting on the process and outcomes to the eSafety Commissioner once a cycle is complete.

To remain effective, risk assessment must be ongoing and will require services to regularly repeat the process. In a rapidly changing environment, regular assessments and mitigations will make sure services are responding to any changes such as shifts in user behaviour, evolving harms, or improvements in mitigation. If risk assessments are not regularly undertaken, there is a risk that they will no longer be up to date and that mitigations won't be addressing risks on a service as they currently exist or will be inferior to current best practice. Creating a rhythm will help to generate a "virtuous cycle", where continuous efforts to improve safety drive down levels of harm over time.[43] This would allow for a progressive reduction of risk and harms on a service over time.

A risk assessment cycle should be completed at least on an annual basis, similar to Article 34 of the EU Digital Services Act. This period of time allows for sufficient time for services to conduct the process, while being sufficiently regular to capture relevant changes on a service.

Identifying and assessing risk

The first stage in the risk assessment cycle is to **identify** the harms relevant to a service which needed to be assessed, and to **assess** the systemic risks relating to the presence or prevalence of those harms arising from the service's design, operation and governance. As provided for in Article 34 of the EU Digital Services Act, service providers would be required to "diligently identify, analyse and assess any systemic risks … stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the uses made of their services".

The identification and assessment of risk should be made with reference to the general categories of online harms outlined at section 5.6 of this report, with service providers analysing and assessing how their service's design and operation affects the presence and prevalence of harms within these domains on the service. In particular, service providers should consider how systemic risks are affected by factors such as those identified by the EU Digital Services Act at section 34.2:

- The design of recommender systems and any other relevant algorithmic systems
- Content moderation systems
- Applicable terms and conditions and their enforcement; and
- Data related practices.

This list is not exhaustive, and should also include other factors to services specifically or in general. These might include:

- Internal complaints and dispute resolution processes
- Staffing and resourcing, such as the number, distribution and training of trust and safety personnel; or
- Systems for verifying age and identity of users or account applicants.

Other factors to consider may be those identified in industry codes and standards, such as those in section 5(d) of the Social Media Services Online Safety Code (Class 1A and Class 1B Material) for determining the risk profile of a social media service. New requirements can build upon many existing elements in the current framework and from regulatory guidance made by the eSafety Commissioner.

Reducing, mitigating and repairing risk

Of course, identifying and assessing risks is only the first step. Once risks have been identified, measures must be adopted and implemented whereby these risks are **reduced, mitigated** and **repaired**:[44]

43 Woods, L. & Perrin, W. (2019), 45.

44 World Economic Forum (2023), *Digital Safety Risk Assessment in Action: A Framework and Bank of Case Studies*, Insight Report, May 2023.

- Risk of harm must be reduced to prevent harms occurring or proliferating in the first place, through the embedding of appropriate design decisions and safety mechanisms
- Harms must be mitigated through measures to detect harmful content, contact, or conduct, and then to remove, suspend, restrict or otherwise reduce exposure to it; and
- Harms must be repaired through robust and appropriate measures for identifying, escalating and prioritising problems, addressing and resolving complaints or appeals, providing support and guidance to those affected, and appropriately consulting affected stakeholders.

Reduction, mitigation and repair are often described under the broad category of mitigation, however they do capture different aspects of the process. In particular, they address the difference between measures taken to prevent harms from occurring (reduction) and measures taken to address the impact of harms when they occur (mitigation and repair). Where possible and as far as possible, services should incorporate 'safety by design' measures for "anticipating, detecting and eliminating online harms before they occur".[45] But where proactive and preventative measures fall short, measures should be in place to mitigate and repair the harms which occur.

Article 35 of the EU Digital Services Act provides a non-exhaustive list of measures service providers can implement to address the risk and impact of harms on their services. This list could be the basis of similar provisions in the Online Safety Act. These measures include:

- Adapting the design, features or functioning of services
- Adapting their terms and conditions and their enforcement
- Adapting content and conduct moderation processes, including the speed and quality of processing reports related to harmful content and conduct and, where appropriate, the expeditious removal or restriction of the content or accounts
- Adapting any relevant decision-making processes and resources for content and conduct moderation
- Testing and adapting algorithmic systems, including recommender systems
- Adapting advertising systems
- Reinforcing the internal processes, resources, testing, documentation or supervision of activities in relation to the detection of systemic risk
- Taking awareness-raising measures and adapting online interfaces to provide more information; and
- Taking measures to protect the rights of the child, including age assurance, complementary default safety measures and parental controls and tools to help minors signal abuse and obtain support.

Monitoring, measuring, reviewing and reporting

Finally, a robust risk assessment process requires effective measuring and evaluating the accuracy of assessments and effectiveness of the measures undertaken, and the full and accurate reporting of the process for accountability.

This element of the process requires:

- Capabilities and practices for effectively and accurately monitoring the occurrence and impact of harms and the factors contributing to them
- Effective and accurate measurement and review of the impact of measures undertaken to address the occurrence and impact of harms; and
- Full and frank reporting on the process and progress of the risk assessment, to the regulator for accountability purposes and for the purposes of feeding back into the next iteration of risk assessment through application of lessons learned.

To provide for external reviews or audits of risk assessments, there needs to be a requirement on service providers subject to the risk assessment obligation to preserve records and supporting documentation for their risk assessments for a period of five years, as expected for certain records under the Basic Online Safety Expectations. Providers should be required to supply these records and documents to the regulator on request. Keeping records is also provided for in Article 34(3) of the EU Digital Services Act, which requires providers to "preserve the supporting documents of the risk assessments for at least three years after the performance of the assessments". The UK's Online Safety Act similarly requires services to "make and keep a written record" of "all aspects of every risk assessment" (section 23.2).

45 eSafety Commissioner, *Safety by Design,* https://www.esafety.gov.au/industry/safety-by-design.

Risk assessments to also be undertaken when significant changes are made to a service

It is not sufficient that risk assessments be conducted on an established regular basis. When significant changes are made to the design and operation of a service, these changes may have an effect on the level of systemic risk in a service which can be sufficient to render much of the existing cycle's risk assessment obsolete. Such changes could include the introduction of new features and products (or significant changes to existing ones), or the making of significant changes to the resources, architecture, rules, terms of use and policies governing a service.

Examples might include:

- The introduction of end-to-end encryption in a service's direct messaging function
- Changes to the algorithms and recommender systems affecting contacts, conduct and content delivery on the service; and
- Changes in the staffing, technology, processes or policies relating to content moderation on a service.

All of these kinds of changes affect the calculus of risk[46] on a service and may produce increases in harms such as the spread of illegal or harmful content. When such changes are made, existing assessments of the systemic prevalence or potential for harms, and the existing suite of mitigations and repairs, may no longer be sufficient or effective – and new mitigations and repairs will need to be considered.

This would reflect the reporting approach already adopted by eSafety in the Online Safety (Designated Internet Services—Class 1A and Class 1B Material) Industry Standard 2024 (sections 33 and 34) as well as the Online Safety (Relevant Electronic Services—Class 1A and Class 1B Material) Industry Standard 2024 (section 34).[47] This requirement would also be in line with the EU Digital Services Act's requirement that risk assessments be conducted "in any event prior to deploying functionalities that are likely to have a critical impact" on risks (section 34.1), or the UK Online Safety Act's requirement for further assessments before "making any significant change to any aspect of a service's design or operation" (sections 9.4 and 11.4).

Risk assessment obligations to be applied on a risk-based and proportionate basis

While all services should be required to undertake risk assessments, this requirement to conduct and report on risk assessments as outlined above should only be applied to services on a risk-based and proportionate basis. The application of full risk assessment obligations would therefore be limited according to reach and risk criteria:

- **The reach of a service in Australia.** Services which are used by a significant number of Australians would be presumed to pose a sufficient risk of harm to be subject to risk assessment obligations, unless their functionality is so limited as to clearly pose no relevant risk. This would be similar to such obligations under the EU Digital Services Act, which apply to "Very Large Online Platforms" with monthly active users in the EU amounting to 10 per cent of the population or above.
- **The inherent risk of harm posed by a service, in the absence of mitigating measures.** Some types of service may, independently of their reach, pose a significant risk of harm to Australians, sufficient to rule them in to a risk assessment obligation (see Chapter 4 on the sections of the online industry).

Beyond these criteria, the regulator should have discretion to require services to conduct the risk assessments and produce reports described above. Importantly, services which don't meet the threshold for formal risk assessment and reporting should still be subject to an overarching duty of care and safety by design, and would be expected to conduct some level of risk assessment and mitigation, retaining records that could be subject to audit.

46 Treasury describes the calculus of risk as having four components: (a) the probability that the harm would occur if care was not taken; (b) the likely seriousness of that harm; (c) the burden of taking precautions to avoid the harm; and (d) the social utility of the risk-creating activity. Treasury.gov.au 2019 Foreseeability.

47 These standards are intended to commence in December 2024.

Risk assessment and mitigation measures to incorporate the best interests of the child as a primary consideration

Under a duty of care, online services would be responsible for the safety of all their users, and all persons affected by the use of their services. However, in the online environment as elsewhere, children are particularly vulnerable and susceptible to harms and require special protection of their rights and safety. Article 3(1) of the Convention of the Rights of the Child provides that "[i]n all actions concerning children, the best interests of the child shall be a primary consideration." I firmly believe this should be a guiding principle which online service providers should follow in the design and operation of their services.

Currently, the Basic Online Safety Expectations provides the expectation that online service providers will "take reasonable steps to ensure that the best interests of the child are a primary consideration in the design and operation of any service that is likely to be accessed by children", with one of the reasonable steps outlined being that risks to child safety are assessed and appropriately mitigated. I recommend that this expectation be an essential basis of meeting the duty of care, and in particular of the conduct of risk assessment and mitigation obligations. However, in keeping with the broader scope of a duty of care, beyond the end-users of a service, online services should not only consider the risks to children insofar as their service is "likely to be accessed by children", but how the best interests of children may be impacted by the use of their services. As International Justice Mission note in their submission:

> *It is critically important that the OSA regime takes into consideration non-users who are harmed … through of services and platforms … In some of the worst forms of online child sexual abuse – such as livestreamed child sexual abuse – children who are non-users undergo severe harm and trauma.*[48]

08

Recommendation 8:

The best interests of the child should be a primary consideration for online service providers in assessing and mitigating the risks arising from the design and operation of their services, including risks to children who may use the service and risks to children as a result of how the service may be used.

48 Review Submission 95 – International Justice Mission, 11.

5.8 Codes

The Act should provide for the regulator to make codes where there is a need to provide mandatory and enforceable compliance measures for regulated entities and to direct them about how to comply with certain aspects of a duty of care. Codes would not be intended to cover the field of the duty of care obligation and should be developed as needed. Enforcement action under the duty of care can be taken even if there is no code. The regulator should have sufficient flexibility to determine their scope. The decision to make a code could be made based on considerations, such as the identification of a poor industry practice or a request from industry, but in any event would be an independent exercise of the regulator's powers.

The regulator should 'hold the pen' drafting the codes in the first instance but should conduct a robust public consultation with relevant stakeholder groups like industry participants, civil society, academics, and other relevant regulators in drafting. This will be a more efficient process, with code-making expected to take less time. It would also relieve industry associations of the greater impact on time and resources that comes with an industry-led approach. As mandatory and enforceable instruments, all final codes would be subject to: public comment and consultation, the normal scrutiny processes of Parliament for delegated legislation, registration on the Federal Register of Legislation, and a 15 sitting day period of notice for motions of disallowance.

Codes could be established in relation to the broad harm domains set out in the Act, providing flexibility for the regulator to address specific harms within a specific code. For example, self-harm could be addressed in a code relating to harms to mental and physical wellbeing. A code could elaborate on or identify specific harms to be addressed under a broad harm domain where necessary or when harms emerge. However, a duty of care recognises that services themselves have a primary responsibility, having the greatest power over and knowledge of their systems, for identifying harms before (where possible) or as they emerge.

Codes could also be established by the regulator to specify mandatory and enforceable compliance measures for service providers in meeting the duty of care or to address other deficiencies identified through the administration of the Act. This might include mandatory requirements on conducting risk assessments, transparency, or in relation to an Australian context where offshore service providers may not have the relevant local knowledge (such as in relation to protecting First Nations People).

Codes, however, should not create safe harbours. Inevitably once a code is made, there will be a period of time before it is reviewed. In the meantime new and better ways to protect people will be found and I do not want to disincentivise this happening. Some new ways may not amount to a breach of the code but others may. Where the service is concerned that they could be in breach of the code, they should consult with the regulator and the regulator should have the ability to approve their changes. The regulator should also continue to be able to publish non-binding guidance.

Providing maximum flexibility in relation to code-making powers is not without risk. There is a possibility that the number of codes will grow over time and it will become an impossible task to keep them up to date. This risk can be mitigated by ensuring codes are developed through a public consultation process, subject to scrutiny and have a review mechanism built in.

09

Recommendation 9:

The eSafety Commissioner should be empowered to create mandatory rules (in the form of codes) on how entities can comply with certain aspects of the duty of care requirements, including addressing specific online harms. This should not stop services from taking additional steps to protect people. Codes would not create safe harbours.

5.9 Transitioning industry codes and standards under the current Act

A considerable amount of time and effort has gone into developing the industry codes and standards under the Online Content Scheme, and that work is still underway on developing a second phase of codes. It is not my intention for this work to go to waste, and I encourage this work to continue. It will take time to make and implement any legislative changes based on the report's recommendations, so it is important to continue implementing the current Act so that the protections it provides for may be effective in the meantime. Transitional arrangements would at any rate be needed to ensure a continuity of protection under the Act as the new framework is implemented.

5.10 Micro sites and decentralised platforms

While the global regulatory environment is trending toward harm mitigation through the largest or highest risk online service providers, decentralised online platforms and services (including Web 3.0 technologies or DWeb) could introduce new regulatory challenges in online safety.

Decentralisation offers potential benefits by providing greater control to users by reducing their reliance on mainstream, centralised services and distributing responsibility for data sharing and storage to communities of users.

In its submission, Digital Rights Watch asserts that decentralised platforms should not be framed as a regulatory challenge but as "an example of how online communities can self-manage and moderate their communities according to specific contextual and cultural rules and norms," and that such sites can free users from the cultural monopoly of mainstream services.[49]

While acknowledging the potential benefits, it has been estimated that most major decentralised platforms do not have the necessary tools to manage harmful content and conduct or enforce their own rules, with cases of administrators and moderators having to remove content manually screen by screen.[50] With concerns around the ability to moderate or regulate decentralised services or platforms, you can easily see the potential risks for an increase in illegal material or harmful content and conduct and creating more 'online cesspools.' Decentralisation makes it more difficult to hold users responsible and creates challenges for the current online safety framework.

eSafety has suggested a range of ways that decentralised services could work to keep users safe. Suggested strategies include implementing a community moderation policy based on agreed rules, opt-in governance established in blockchain networks, ability to establish trusted pseudonyms to enable users to remain anonymous as long as they engage appropriately, and enabling third party content moderation tools to prevent the most harmful content.[51] Under recommendations in this report, these platforms would be subject to the duty of care and would need to consider measures such as these, especially if they are designated for heightened obligations under reach and risk criteria.

Working to support decentralised platforms and arm them with appropriate safety tools may be appropriate for the time being, but this is an area that needs to be watched, including looking at whether the type or frequency of harm is changing over time and whether decentralised services are doing enough to keep users safe and act within the laws of Australia.

49 Review submission 112 - Digital Rights Watch, 17.

50 Article, Samanthal Lai, Yoel Roth Online Safety and the "Great Decentralization" – The Perils and Promises of Federated Social Media | TechPolicy.Press. Also reference original study: Findings Report: Governance on Fediverse Microblogging Servers.

51 eSafety Commissioner (2022), Decentralisation – position statement | eSafety Commissioner eSafety Commissioner.

ACCOUNTABILITY AND TRANSPARENCY

06

A term often associated with online services is the “black box”. This term symbolises the lack of transparency around how services and the technologies they use, such as recommender systems, artificial intelligence, moderation systems and more, lead to what we do and don’t see when we access them. The opaque nature of services, and in particular online platforms and search and app distribution services, demand that transparency measures are put in place so the regulator can properly monitor the safety of services and enable others to assess how much trust to place in a service.

As I have said elsewhere in this report, there will be a need to ensure alignment with existing frameworks and reforms across Government. In the case of transparency obligations, I understand the Safe and Responsible Artificial Intelligence agenda, which is currently in development, also contains proposed transparency measures.

6.1 Transparency reporting

One of the most useful powers eSafety has is its ability to require services to provide information related to the Basic Online Safety Expectations. It enables the regulator to ask forensic questions and make an assessment about how much services are (or are not) doing to keep users safe. It can deliver broader online safety gains by shedding light on a service's practices. This power should be retained. However, in order to implement a duty of care model, greater transparency is needed in respect of the decisions and processes of services to ensure compliance with the duty of care and due diligence obligations.

The Act should continue to provide eSafety with the ability to issue transparency reporting notices to services. There should be a broad ability for eSafety to require information, in the manner and form specified, about any element of how the service, management and users conduct themselves. This would include requiring the provision of information in response to specific questions.

Services with the greatest reach or risk should also be required to prepare and publish annual transparency notices. This is in line with requirements of the EU's transparency reporting obligations in Articles 15 and 42 of the Digital Services Act. Australian transparency reports should cover, among other things:

- **Proactive content moderation,** including automated moderation, the amount and type of content removed as well as the human resources devoted to content moderation
- **The number of complaints and types of complaints received,** how they were resolved including the number not dealt with, and the average time to resolve complaints
- **The number of Australian end-user accounts suspended or removed** permanently from the service and why, as well as the number of challenges to these decisions and the outcomes
- The number of average monthly users, broken down into children and adults
- **The results of risk assessments**
- **The mitigation measures** put in place, or to be put in place, as a result of risk assessment
- **Any measurement of the impact mitigation measures** (an evolving and underdeveloped area); and
- **Audit reports** if the regulator has used its discretion to require one.

To the extent that we can align with questions asked by other regulators such as the EU or UK, this will reduce the burden on industry though, of course, the responses must relate to the online safety of Australians.

This annual transparency report would provide eSafety with important information needed to better understand what is happening on each service, where problems are, and how to best prioritise use of eSafety's resources.

The full transparency report should be provided to eSafety. eSafety currently publishes summaries of reporting notice responses. With transparency reports required annually, continuing to do this would be a large resource drain on eSafety. Instead we should follow the EU Digital Services Act example and require the service to publish a summary on its website. They would not need to reveal matters that are commercial in confidence or which could be used by bad actors to, for example, circumvent systems.

10

Recommendation 10:

In addition to risk assessments, a service with the greatest reach or risk should be required to provide an annual transparency report and publish a summarised version on its website. This should not replace the broad power for eSafety to require periodic and non-periodic transparency reports from all services.

6.2 Providing individuals with information about decisions taken that affect them

All services should be required to have a clearly accessible, simple, and user-friendly way to report problems to the service – whether those problems relate to harms to the user or action taken against a user. All such contacts must be responded to within a reasonable time. Where a contact relates to threats of physical harm or image-based abuse, responses should be provided within 24 hours of notification.

6.3 Compliance function

Ideally all services should have a well-resourced compliance function that reports directly to senior management as needed, but at least quarterly. At a minimum all services of greatest reach or risk must have a compliance function. The compliance function should be independent from other areas of the service, be adequately resourced and staff should have training in compliance. Reporting should be at least quarterly to the audit and risk committee of the board and at least annually to the board. Only the board (or its equivalent) should be able to dismiss the head of the compliance unit.

11

Recommendation 11:

Services with the greatest reach or risk should be required to have a well-resourced compliance function that reports directly to senior management as needed, and at least quarterly to the audit and risk committee and annually to the board. Only the board (or its equivalent) can dismiss the head of the compliance function.

6.4 Audits

The regulator, at its discretion, should be able to require any service that is designated within the highest reach or risk group to undertake an audit. To the extent that the requirements of the Act are similar to those of the EU Digital Services Act, the requirements for the audit report should be aligned as much as possible to reduce the burden on industry.

Any audit should cover duty of care and due diligence obligations recommended in this report as well as content moderation, algorithms, compliance with codes and the takedown schemes and other matters eSafety has concerns about. The audit report should be provided to eSafety.

The auditor undertaking the report must be independent, and must not have done work for the service in the previous year and not be presently performing any other work for the service. If eSafety wants regular audits, auditors should be changed at least every five years to avoid capture.

While auditing online platforms, search engines and app distribution services is an emerging field of expertise and will no doubt continue to mature, any chosen auditor should meet certain criteria. They should have a track record in risk management, technical competence and expertise, and adhere to an appropriate professional ethics code of practice as required by the EU Digital Services Act in Article 37.

12

Recommendation 12:

The regulator should have the discretion and power to require services to undertake an audit at their own expense.

6.5 Providing researchers with information that can be analysed and shared with the community

Research contributes greatly to society's ability to meet current and future changes and can directly benefit the wellbeing of citizens. A scheme that provides accredited independent researchers with access to data would encourage more research and more detailed consideration of the many complex problems in the online world and help decision makers.

Some online platforms, such as Meta, provide conditional access for researchers but not all do.

The United Kingdom's regulator OfCom is currently considering the best way to provide access to data for research, with a report expected by May 2025. Data access is also provided for in Article 40 of the EU's Digital Services Act, though I understand they are still working out the best way to do this, with regulations recently released for consultation. The Australian Government's September 2023 Response to the Privacy Act Review Report also agreed in-principle that entities regulated by the Privacy Act should provide information to online users about the use of targeting systems, including clear information about the use of algorithms and profiling to recommend content to individuals.[52] The Report noted that information about targeting systems could be requested by the Information Commissioner to monitor compliance and should be made available to the public to facilitate research into emerging risks.

When finalising Australia's approach, we should consider where the United Kingdom and European Union end up and why. We should also consider advancements in Australian privacy reforms to ensure alignment and reduce regulatory overlap. The ability to provide data access to accredited researchers should be included in any revisions to the Act, even if the provision has a delayed activation date.

At least initially, only services designated as having the greatest reach or risk should be required to be involved in sharing data for research purposes, though clearly other services could voluntarily participate.

Any research approved under this scheme should be for the purposes of determining compliance with a duty of care model, the takedown schemes, or for research into emerging harms.

In line with the EU Digital Services Act's Article 40A, services should only be able to refuse access if they do not have the data or if giving access to the data will lead to significant vulnerabilities in the security of their service or the protection of confidential information, including trade secrets. In designing this provision, compatibility with the Privacy Act should also be considered.

An independent body with an appropriate level of expertise would be responsible for authorising researchers. The Australian Research Council strikes me as a possible candidate, though I have not had the chance to canvass this with them.

To reduce the administrative burden for all involved, consideration could be given to establishing a panel of approved researchers, with a call for applications every three years and requirements to notify the regulator about any conflicts of interest that arise or any change in affiliations. Alternatively, the number of research project approvals could be capped each year.

It will also be necessary to apply certain criteria to the application process, to ensure they are of genuine value to the advancement of online safety.

52 Australian Government (2023), Government Response to the Privacy Act Review Report, 12.

13

Recommendation 13:

Subject to adequate safeguards, services with the greatest reach or risk should be required to share data with authorised researchers for the purposes of determining compliance with a duty of care model, the takedown schemes and research into emerging problems and harms.

Managing risks associated with a scheme for third-party data access

There are key risks which would require careful management, including user privacy, cybersecurity and misuse of sensitive data, misrepresentation and distortion of data for nefarious purposes, and potential conflicts of interest based on commercial or political interests.

Using the EU Digital Services Act model as a guide, the following criteria could be considered in a future Australian scheme:

- Researchers or at least the lead researcher should be an Australian resident
- Affiliation with a scientific research organisation
- Independence from commercial interests
- Disclosure of the funding of their research
- Capability to fulfil specific confidentiality and data security requirements in relation to protecting personal information
- Research is for the purpose of the detection, identification and understanding of specific risks or the assessment of the adequacy, efficiency and impacts of the risk mitigation measures of services with the greatest reach or risk
- Access to the data is necessary for the purposes of the research; and
- An undertaking to publish research results, free of charge, within a reasonable period after the research is completed, subject to appropriate privacy concerns and safety of end-users.

6.6 International collaboration could deliver greater transparency

Aligning transparency reporting and working with our international partners to build better repositories of information would also deliver gains.

Consideration could be given to providing public access to data portals such as ad repositories and databases of moderation decisions, noting some services already provide some portal access.

We should also let people in Australia know about the resources that everyone can access. For instance, the EU's terms and conditions database[53] and statement of reasons database[54] which contains 10 million statement of reasons from the past six months alone as of October 2024. These data bases are all publicly available and are meant to inform the public, researchers and other actors.

eSafety should also let its international counterparts know what public data, research and educational materials they have that others may find helpful – as I am sure they already do. eSafety's educational resources and protection programs are excellent and I'm confident that jurisdictions that have only limited resources for creating such materials would find there is much they can adapt and use for their own communities. Most of the services operate globally or near globally so it makes sense for regulators to cooperate and share, to maximise the global efforts to protect all people from online harms.

And of course, bodies such as the Global Online Safety Regulators Network (GOSRN), that Australia helped establish, are essential. We are all dealing with global entities, so it is essential that we all learn from each other and collaborate for maximum impact.

53 European Commission Online Platforms Terms and Conditions Database

54 European Commission Home - DSA Transparency Database

SAFETY NETS – SUPPORTING ONLINE USERS

07

Although the move towards systems-based regulation aims to limit the online harms that users may experience, it is crucial that users can seek immediate support when they do experience harms. eSafety's complaint-based schemes are powerful tools, enabling quick remediation of harmful online content. It is my hope that strengthening and streamlining existing schemes will enable eSafety to better support people online and provide additional protections where the existing schemes fall short.

7.1 Systems-based regulation can prevent online harms, but safety nets are needed when harms occur

Harmful online content can be seriously damaging, especially for those most at risk, such as children and young people. As eSafety has noted, the social, emotional, psychological, and physical impact resulting from the production, distribution and consumption of harmful content is felt both immediately and over time.[55]

The potential for online harms is endless. Harms arise in many ways: through contact with harmful content, as a result of contact with others online, or from harmful behaviour enabled by a specific technology or a service.[56] The online environment can amplify these harms, spreading content fast and to a wide audience. Where the harm is significant, individuals need appropriate and effective actions to have harmful material and accounts removed.

Although many Australians have demonstrated resilience in the face of an increasingly toxic digital communication environment, the impact of harms on individuals and the degradation of what is now our key communication framework has the unfortunate capacity to leave many individuals under-supported and to shape social interaction in ways broadly undesirable, adversarial and hostile.[57]

No government can completely protect its people from online harms. Systems-based regulation such as duty of care and due diligence obligations aim to prevent harms from occurring, whereas complaints-based removal schemes focus on minimising the impact of harms that have occurred. Investigating individual complaints can be resource intensive but is necessary, at least for now, to protect targeted individuals and to limit the harm they experience.

Throughout the review, I heard a range of concerns about the existing complaint schemes, including concerns about relying on removal schemes that only take effect after a harm has occurred, their ability to deal with constantly changing technology and harms and their ability to address the increasing volume of harmful online content, particularly with content now created by generative artificial intelligence. Communities highlighted that complaint schemes place the burden of addressing harms on targeted individuals and have a greater impact on vulnerable communities.

Many of these concerns would be addressed by introducing a statutory duty of care. A systems-based approach is better suited to adapt to emerging harms and harm types, including interactions with others (contact risks such as grooming, recruitment or radicalisation), behaviour facilitated by technology (conduct risks such as technology facilitated abuse or technology facilitated gender-based violence) and contract risks (commercialisation and datafication of online users). However, a very substantial uplift in safety efforts is needed by online services before consideration could be given to limiting the availability of takedown powers.

It is my hope that by enhancing systems regulation through a statutory duty of care we will eventually remove the need for case-based regulation. It is clear though that, for the time being, the Act's complaint-based removal schemes must remain in place.

Notice and takedown regimes provide a useful role, especially for specific types of content or harms like non-consensual intimate imagery and adult cyberbullying, in complementing systems-based regulatory frameworks.[58]

55 eSafety Commissioner (2024), 'Impact Analysis: Online Safety (Relevant Electronic Services – Class 1A and 1B Material) Industry Standard 2024 and Online Safety (Designated Internet Services – Class 1A and 1B Material) Industry Standard 2024', 7.

56 World Economic Forum (2023) Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms, August 2023, https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf, 5.

57 Review submission 106 - RMIT Digital Ethnography Research Centre, 2.

58 Review submission 166 – Meta, 19.

7.2 Complaint and content-based removal schemes are effective and valued

Throughout the review I repeatedly heard that the removal schemes are recognised as a strong, world-leading model of regulation that have been successful in addressing impacts on individual users. Civil society and industry representatives noted that existing schemes are highly valued and perceived to be working well.

The public facing complaints mechanism in the complaints and content-based removal notices schemes are world leading, and for those who have been harmed in specific ways as covered by the Act, it can be life changing.[59]

Young people in particular are relieved to learn that there is somewhere to go to seek help. The review heard that these services are worth preserving.

The presence of the child cyberbullying and image-based abuse schemes are very valuable and allow education providers to spread community awareness and empower people with the knowledge that this option exists.

The numbers of complaints reported to eSafety make it clear that the existing schemes are valued by the Australian community, and enable eSafety to rapidly respond to complaints that meet the thresholds for regulatory action. eSafety consistently receives feedback that users are extremely grateful for the schemes and for eSafety's swift actions.

eSafety's annual reports give us an insight into the effectiveness of the cyberbullying, adult cyber abuse and image-based abuse takedown schemes, though there are inconsistencies with how this information is presented year to year.[60] This needs to be addressed. Year on year comparisons are essential to assess the efficacy of these schemes (see Chapter 13).

59 Review submission 70 – Reset.Tech, 29.

60 Annual report 2022-23 Australian Communications and Media Authority and eSafety Commissioner; Annual report 2023-24 Australian Communications and Media Authority and eSafety Commissioner.

Child cyberbullying scheme

In 2023-24 eSafety:

- Received 2,693 complaints – a 37 per cent increase on the previous year
- Made 821 informal removal requests
- Were successful in removing 82 per cent of child cyberbullying content.

In 2022-23 eSafety:

- Received 1,969 child cyberbullying complaints
- Made 636 informal removal requests and were successful in having 84 per cent of the child cyberbullying content removed
- Issued 13 formal end-user notices that required individuals responsible to remove the cyberbullying material and cease cyberbullying the target.

Adult cyber abuse scheme

In 2023-24 eSafety:

- Handled 3,112 actionable complaints
- Made 3 formal removal notices
- Made 383 informal requests for removal of harmful material
- Were successful in removing content in 284 cases (74 per cent)
- Interestingly eSafety now receives more complaints under the adult scheme than the child cyberbullying scheme.

In 2022-23 eSafety:

- Handled 2,516 actionable complaints
- Made 601 informal notifications, with material removed in 466 cases (77 per cent)
- Issued 3 removal notices with material removed in all 3 cases.

Image-based abuse scheme

In 2023-24 eSafety:

- Received 7,270 reports of image-based abuse – a 20 per cent decrease
- Requested removal of material from more than 947 locations across 191 platforms and services
- Were successful in having 98 per cent of material removed on request.

In 2022-23 eSafety:

- Received 9,060 reports of image-based abuse
- Requested removal of material and URLs from 6,500 locations
- Were successful in having 87 per cent of material removed on request.

A quick look at these statistics shows that while the image-based abuse scheme appears to be working well, complaints about child cyberbullying and adult cyber abuse are increasing, with a smaller proportion of people who complain (particularly for adult cyber abuse) receiving help through these schemes. While the child cyberbullying scheme numbers do not reflect the additional work eSafety does with schools and parents of affected students, these statistics raise questions about whether the bar for these schemes is set too high.

It is important to have the data to assess the efficacy of existing takedown schemes. Going forward it would be useful for eSafety to have a consistent set of data and language on the use of complaints schemes and for them to include it in their annual report. The data should cover:

- The number of takedown requests received for each of the four schemes
- The number that satisfied the threshold for takedown in each scheme where a threshold is set and the number for those who didn't
- The number of formal removal notices given to the service provider for material to be taken down and the number and percentage taken down by each scheme as a result
- The number of informal takedown requests made under each scheme and the number and percentage taken down in this way; and
- The number of informal requests where the complainant didn't meet the threshold for a formal takedown request and how many were successful.

14

Recommendation 14:

For the avoidance of doubt, the legislation should make it clear that informal requests for takedown are legal and legitimate as they lead to quicker results for individuals who are often in severe distress.

7.3 Complaint and content-based removal schemes can be streamlined and strengthened

It is clear that inconsistencies across the four complaint and content-based removal schemes add complexity for eSafety and those seeking to make a complaint. The tables at Appendix D outline these variations, including who can report, who is protected, the link required to Australia and the available regulatory actions. Changes can be made to strengthen and streamline these existing schemes.

For all four schemes, eSafety has powers to investigate complaints made, but can only take formal action where a complaint meets conditions specified in the Act. Where the conditions are met, eSafety can issue a removal notice as a formal compliance mechanism. These notices can be given to social media services, relevant electronic services, designated internet services, hosting service providers and, in some instances, to the individual who posted the harmful material. For child cyberbullying and adult cyber abuse complaints, eSafety can only issue a removal notice if the complainant has already reported the material to the online platform and the platform has not removed the material within 48 hours of the complaint.

Under the Online Content Scheme, a complaint can be made by a person or government based in Australia if it is suspected that Australians can access Class 1 or Class 2 material (illegal and restricted online content). eSafety can also commence its own investigations. For the other three schemes (cyberbullying, adult cyber abuse and image-based abuse), complaints can only be made by the targeted individual or their representatives. Exceptions do apply where the individual is a child or is mentally or physically incapacitated and a person has been authorised by the individual or by the parent or guardian of the child to make a complaint. This means that harmful content such as deepfake intimate images of a person can still spread online before the targeted individual becomes aware of it and makes a complaint.

7.4 Changes to schemes are needed to better protect people in Australia

While submissions recognise the value of complaint schemes as an important backstop for specific harms, the review identified opportunities to improve and broaden the schemes to provide better protections for members of the Australian community, particularly groups disproportionately experiencing online abuse.

The complaint scheme rules and broader regulatory complexity allow seriously harmful content to remain online for too long

Prerequisite reporting to platforms

The adult cyber abuse and child cyberbullying schemes only allow eSafety to issue a formal removal notice where the online service has failed to act on a complaint within 48 hours. The requirement to first report to online services limits regulatory intervention to circumstances where a person is unable to obtain relief from the service provider. However, this prerequisite adds to the reporting complexity for complainants and provides a 48-hour window in which online harms can amplify. Requiring complainants to report directly to an online service may not always be feasible, particularly where the service benefits from operating a service that is rife with online harm such as websites set up to 'dox' complainants (intentional exposure of the person's identity, private information or personal details without their consent)[61].

[T]he complaints system does not sufficiently consider the permanency of social media. A harmful post can be on social media for under a minute and still cause considerable damage. It may be online forever, for example, if someone reposted it … I strongly suggest amending the Act to lower the time limits that social media have to respond to complaints and require social media to actively search online for every use of the harmful material and remove it.[62]

Removing the platform reporting prerequisite would simplify the complaint process for users seeking help from eSafety. The person would no longer be required to report to the platform first, and would not need to provide evidence of having made a complaint to the platform. This would remove a reporting step, a 48-hour window in which the online harm can amplify, and mitigate the risk of introducing additional harm if a service is motivated to act against the person with malice. For more serious abuse, including threats for example, the person could also contact police.

However, removing the prerequisite reporting requirement would result in an exponential increase in the volume of complaints eSafety receives and require significant additional resourcing. More importantly, the primary responsibility for addressing harmful content and activity must remain on platforms to provide adequate complaint mechanisms.

To uphold this expectation on platforms and reduce the window in which online harms can amplify, the statutory delay to issue a removal notice should be reduced to 24 hours. The regulator should also have discretion to waive the platform reporting prerequisite in circumstances where there is no clear mechanism for online users to submit a complaint to the online service, or where reporting would lead to a reasonably foreseeable risk of further harm to the user experiencing the abuse.

61 eSafety Commissioner (2024), Doxing, Doxing | What is doxing or doxxing? | eSafety Commissioner.

62 Review submission 20 – Associate Professor Marilyn Bromberg, 2.

Case Study: Demands for intimate images

'Anna was in Year 10 when her boyfriend Jason started hassling her for nudes. It was in the middle of a long COVID lockdown while the two were messaging frequently, and their chat soon became overwhelmed with requests for intimate photos. Anna says she felt a lot of pressure and eventually caved in. Then after receiving the first photo, Jason kept asking for more. The harassment continued and Anna soon broke up with Jason. She blocked him on everything, including social media, and thought the "toxic" period of her life was over. It was only after returning to school that she realised it wasn't. A school counsellor pulled her aside one day for a check in … "They said a parent had rung the school and said I had been 'distributing nude images', as if I had been going around texting them to people unsolicited or something," she said … Meanwhile, she says Jason and the boys who were sharing her images were not held accountable. Jason was in the same form class as Anna, so she was forced to sit in the same room as him every day … the following year, she was blocked from running for school captain.'

ABC News, 11 September 2024[63]

63 Thorne, Leonie (2024),' Judged by friends, shamed by staff: How 'revenge porn' left a teenager paying the price for years', ABC News, Wednesday 11 September 2024, Judged by friends, shamed by staff: How 'revenge porn' left a teenager paying the price for years - ABC News accessed 31 October 2024.

15

Recommendation 15:

Users experiencing adult cyber abuse or child cyberbullying should only need to wait 24 hours (not 48 hours) following a complaint to a service before eSafety is able to issue a removal notice.

16

Recommendation 16:

The regulator should be empowered to waive the statutory delay to issue a removal notice for the child cyberbullying and adult cyber abuse scheme where no clear complaint mechanism exists on the online service, or where reporting would lead to a reasonably foreseeable risk of further harm to the user experiencing the abuse.

Individuals experiencing online harms risk falling through the cracks

People in Australia experiencing online harms may seek assistance through multiple agencies and complaint schemes, particularly where their experience involves several types of harm. There is no single regulator or complaint scheme to address the range of harms people experience through online abuse. Specific harms might include image-based abuse (eSafety), cyberstalking and threats (criminal law), reputational damage (courts), doxing (eSafety, criminal and privacy laws) and online hate (eSafety for more extreme forms, criminal or discrimination laws).

Presently, there are a number of Australian government protections against online harms (in a broad sense). These bodies exist in tandem with a range of other agencies and government departments which, working together, regulate Australia's digital environment. This patchwork approach to digital regulation is fraught with danger.

While each body has a unique and important responsibility in this space, there is potential for these responsibilities to overlap and for multiple bodies to work on the same area. There is an inherent risk that allocating responsibilities to each of these disparate bodies is ineffective and causes unnecessary duplication while making it difficult for the platforms and the general public to understand which agency is responsible for what. [64]

Typically, when a person seeks help for a harm experienced online they have suffered something significant, abuse, threats, reputational damage or financial losses. I am concerned that the disparate approach to online regulation can easily result in a situation where an individual is told they can't be helped and effectively turned away. Where this leads to an individual having to approach different organisations, it exacerbates the harmful experience. Each time an individual contacts a new agency or lodges a complaint, they are required to relive the harmful experience and provide evidence of the harm experienced. It must be completely disheartening to be turned away and have to start the process all over again to find someone who can help. Duplicative processes place a significant burden on the person targeted, their representatives, and on the regulators, who collect reports, assess, and action each complaint arising from a pattern of abusive online behaviour separately.

Agencies should collaborate better to avoid turning individuals away, and enable warm handovers within or between agencies to improve user access to assistance, minimise the burden on targeted individuals, and reduce process duplication for regulators.

17

Recommendation 17:

The Government should develop a whole of government 'no wrong door' approach to support individuals seeking help to address online harms. This will require cooperation and information sharing across portfolios, including law enforcement, to address a range of issues such as online safety, child safety, privacy and scams, among others.

64 Review submission 155 – Human Rights Law Centre, 18.

The adult cyber abuse scheme is not adequately protecting communities disproportionately experiencing online abuse

When introduced, the Act included existing complaint schemes for image-based abuse, online content, and child cyberbullying, expanding their application and adding a new scheme to address adult cyber abuse. In a 2018 statutory review of those original complaint schemes, Ms Lynelle Briggs AO acknowledged public submission support for extending the cyberbullying regime to adults and a need to address the online harassment, vitriol and trolling of adults online:

> *I found in this review that the tight limitation on the eSafety Commissioner's role with respect to adults flies in the face of the experience of many people (especially women with high profiles, like journalists and politicians, Aboriginal and Torres Strait Islander women, Islamic spokespeople, and the families of murder and rape victims) with online harassment, vitriol, and predator trolling. A number of these women have approached the eSafety Commissioner for assistance.*
>
> *"[In the words of Dunja Mitjatovic] 'Female journalists and bloggers throughout the globe are being inundated with threats of murder, rape, physical violence and graphic imagery via email, commenting sections and across all social media … Male journalists are also targeted with online abuse, however, the severity, in terms of the sheer amount and content of abuse … is much more for female journalists' …*
>
> *These dangers do not stay online. Following extreme online harassment campaigns, we have had Women in Media members punched in the street and followed home. A couple of our members have had rape and death threats against them and their daughters."*
>
> *Such behaviour is totally unacceptable, and action needs to be taken to prevent it.*[65]

Community groups generally favour harm prevention over reporting once a harm has occurred. This is due to the volume of abuse and the overwhelming burden placed on individuals and community groups to report harms when they occur. However, where the safety net is required and groups have chosen to report, these groups have found that the thresholds for regulatory action (particularly for adult cyber abuse) were too high.

The adult cyber abuse threshold is higher than for child cyberbullying, reflecting that adults generally, or at least in theory, have a higher level of resilience than children and to ensure freedom of expression is not unduly restricted.

> *… we recommend a universal threshold for the current complaints scheme rather than the current two-tiered system of child vs adult, that has led to lower rate of successful complaints for adults vs children. It has also led to a system whereby a child who was bullied two-days before their 18th birthday, would meet the bar, but that same person two days later would face a bigger hurdle in having their complaint upheld. While we completely agree that children do need protecting, the current settings make an assumption that harms are somehow lessened due to age, which is simply untrue. Harm to adults is often severe and can be complicated by a range of different issues such as underlying mental health concerns, socio-economic status, and relationship and family breakdowns and previous history, including childhood history of having experienced online abuse.*[66]

Public consultation highlighted the persistent and compounding impact of online abuse, silencing already marginalised voices. While children were widely recognised as being among the most at-risk in relation to online harms, other sections of the Australian community disproportionately experience online abuse. The risk is higher for First Nations people, women, women in public or prominent positions, people who identify as LGBTQIA+, people from culturally and linguistically diverse backgrounds, people living with disability or medical conditions, people with particular religious beliefs and older Australians.[67]

65 Lynelle Briggs AO (2018), 'Report of the Statutory Review of the Enhancing Online Safety Act 2015 and the Review of Schedules 5 and 7 to the Broadcasting Services Act 1992 (Online Content Scheme)', 32.

66 Review submission 106 - RMIT Digital Ethnography Research Centre, 10.

67 Parliament of the Commonwealth of Australia, House of Representatives Select Committee on Social Media and Online Safety, 'Social Media and Online Safety' March 2022, 36-45, Social Media and Online Safety – Parliament of Australia (aph.gov.au).

The complaint schemes were not seen as adequately addressing the scope and volume of online abuse experienced by people who identify with one or more of these groups.[68] Key issues include the high threshold for removing harmful material (particularly for adult cyber abuse), the cumulative harm arising from the volume, scope and persistence of online abuse, and failure to address abuse targeting a group or individual based on protected characteristics (online hate).

Despite the existence of the adult cyber abuse scheme, the Act itself does not provide for corrective action in respect of online material that amounts to hateful content targeting a particular individual or group, on account of a specific shared characteristic (e.g., religion, ethnic background, culture, disability, age, or gender identity) or those with intersectional characteristics (e.g., gender and race).[69]

While the Act includes a public-facing complaint mechanism that allows users to report harmful content under certain conditions, this mechanism could be expanded to enable more users and communities to seek redress directly.[70]

In addition to increased information sharing and collaboration between security and intelligence organisations and other government agencies, a rapid response capability could be achieved by further expanding the remit of, and a concomitant increase in resourcing for, the eSafety Commissioner to support diaspora groups being targeted.[71]

The threshold at which most adult cyber abuse complaints fail is the objective intent to cause serious harm to an Australian adult. There is no similar threshold for child cyberbullying, which only requires an intent to have an effect on a particular Australian child. Adult cyber abuse material must also be deemed by a reasonable person to be menacing, harassing or offensive in all the circumstances. This second threshold mirrors language used in the Criminal Code offence of using a carriage service in a way that reasonable persons would regard as being, in all the circumstances, menacing, harassing or offensive.[72] The criminal offence carries a maximum penalty of 5 years' imprisonment, but unlike the adult cyber abuse scheme does not require intent to cause serious harm. This seems disproportionately high for a complaint-based content removal scheme.

The consequence or penalty is low. Content is taken down. So why is the threshold so high? [73]

The adult cyber abuse scheme threshold of intent to cause serious harm should be removed and replaced with a threshold more aligned to the child cyberbullying scheme. The new threshold should only require that an ordinary reasonable person would conclude that "it is likely the material was intended to have an effect on a particular Australian adult". This proposed threshold would still require an ordinary reasonable person to "regard the material as being, in all the circumstances, menacing, harassing or offensive." The review heard some concerns that a threshold of 'offensive,' while aligning with language used in the related criminal offence, may be too low for a content removal scheme and that a requirement for material to be 'seriously offensive' should be considered. This proposal has merit and better aligns with terms used in the child cyberbullying scheme.

68 Online Safety Act Review community roundtable.

69 Review submission 149 – Law Council of Australia, 14.

70 Review submission 155 – Human Rights Law Centre, 12.

71 Review submission 21 – Asia-pacific Development, Diplomacy and Defence Dialogue referring to malicious foreign actors. The submission recommended establishing a national body for the information environment, noting potential siloing the way of thinking about threats such as cybersecurity, disinformation, social cohesion, foreign interference, data, privacy and criminal exploitation.

72 Criminal Code, section 474.17.

73 Comment from review civil society group consultation.

Case Study: Online abuse of short-statured people

'Three times in the last few weeks, Samantha Lilly has stumbled across pictures of herself online that she didn't know had been taken. The photos had been posted alongside derogatory captions, attracting dozens of comments from people laughing along and mocking her appearance … One group called "M****t spotting Australia" contained a tagline, "See something small, give us a call". The photos of Samantha that were taken without her knowledge and show her going about her day at the supermarket, near her gym, and in a car park … "[It] made me feel completely powerless, completely subhuman, and something that I don't want anyone else to have to experience." … "But I think the worst thing about all of this is there are photos of people I love on there with quite violent, graphic, disgusting comments — and then there are photos of children." … But getting the Facebook groups removed has been difficult – as some have disappeared from public view, new, more localised ones have been created … '

ABC News, Thursday May 30, 2024[74]

74 'Short statured Australians are facing increased online abuse. They're asking for the public's help to stop it' – ABC News, Thursday 30 May 2024.

18

Recommendation 18:

The adult cyber abuse scheme should be amended by lowering the threshold. The new threshold should require that an ordinary reasonable person would conclude that 'it is likely the material was intended to have an effect on a particular Australian adult', and that an ordinary reasonable person would 'regard the material as being, in all the circumstances, menacing, harassing or seriously offensive.'

Increase the schemes' effectiveness by considering harmful patterns of behaviour as well as individual content

Existing schemes primarily aim to remove or limit access to harmful material identified in a complaint. The schemes differ in actions that can be taken against those who post harmful material online and anonymity of accounts may limit the extent to which these powers can be used.

The image-based abuse scheme establishes civil penalties for posting or threatening to post an intimate image without consent[75] and eSafety can issue a remedial notice to an end-user who has contravened that provision (aimed at preventing further contraventions) or issue them with a removal notice.

There is no equivalent civil wrong for posting child cyberbullying or adult cyber abuse material. The child cyberbullying scheme enables eSafety to issue an 'end-user notice' to a particular person who posted child cyberbullying material, which may be enforceable by court injunction. The notice can require the person to remove the material, refrain from posting cyberbullying material targeting the child, or apologise to the targeted child. The adult cyber abuse scheme enables eSafety to issue a removal notice to a particular end-user who posted the material, but there are no powers to require a person to refrain from posting further cyber abuse targeting the adult.

Account pseudonymity or perceived anonymity can contribute to freedom of speech, and are important privacy concepts that enable individuals to exercise greater control over their personal information and decide how much personal information will be shared or revealed to others.

However, these concepts can also limit accountability for abusive posts and was consistently identified as a contributor to online abuse. Evaluating and removing items of harmful content without addressing the behaviour of online users becomes regulatory 'whack-a-mole' in the face of increasing volumes of online abuse.

One of the significant reasons why so much harm occurs on social media is due to its anonymity. A reason why people write hateful, defamatory and/or harmful material on social media is because they can do so anonymously … I strongly believe that if the Act is modified to remove some of the anonymity associated with social media, it could result in less hate/defamatory/ harmful material being posted and it could be positive for the public's mental health.[76]

Where an online service provider fails to act against accounts persistently generating harmful content, this could be a factor considered in determining whether the statutory duty of care to online users has been met.

A focus on specific items of content also limits the regulatory response to harmful material being reposted after removal. The targeted person must make a new complaint to eSafety before reposted material can be removed, even though the content has previously met regulatory thresholds for action.

75 *Online Safety Act 2021*, section 75.

76 Review submission 20 – Associate Professor Marilyn Bromberg, 3.

Case Study: child cyberbullying

'Jess Tolhurst didn't stand a chance. Teachers tried to keep her away from the bullies at school, her parents kept her safe at home, but no one could keep her tormentors from bombarding her online, from infiltrating her thoughts, from breaking her spirit. Jess took her own life only weeks before the Christmas of 2015, the day before her parents were taking her to the nearest police station to secure an apprehended violence order against her abusers. To this day her mum Melinda Graham just can't comprehend why more won't be done to stop online bullying on social media, why governments won't take strong action, make a stand and stop other children going through the hell that destroyed her daughter. "It was face-to-face, online, every which way, phone calls, all her social media accounts," Ms Graham said. "Snapchat was the biggest one. I used to say to Jess why don't you screenshot and she would say: 'No, I can't, they will be able to see I've done that' … "Messages like: 'Go kill yourself' and: 'If you come back to school I will get you' or calls to: 'Stomp on Jessy's head'."

Ms Graham said the bullying even continued after Jess passed, with her closest friends also becoming targets … "Our daughter was bullied to death, that's the truth of it," she said. "And there are no consequences for the bully. You send screenshots to social media platforms and they do nothing about it. It doesn't go against our community standards, are you kidding me?" '

The Daily Telegraph, Sunday 26 May, 2024[77]

77 'Let Them Be Kids: Bullies who killed Jess were never punished' – The Daily Telegraph, Sunday 26 May 2024.

19

Recommendation 19:

The Act should enable the regulator to issue a removal notice for material that has met the regulatory threshold for removal under a prior complaint, where the regulator becomes aware that the material has been reposted.

20

Recommendation 20:

The Act should include additional powers to require an end-user to stop posting cyber abuse about an Australian adult in an end-user notice, subject to a civil penalty for non-compliance.

This power could align with powers currently available to eSafety through the child cyberbullying scheme end-user notice.

Blocking material that depicts abhorrent violent conduct remains an important power

Under the Act, eSafety can request or require an internet carriage service, such as Optus or Telstra to temporarily block material that depicts abhorrent violent conduct.[78] The powers can be exercised if **material that promotes, incites, instructs in or depicts abhorrent violent conduct** is likely to cause significant harm to the Australian community, for a duration of up to three months.

These blocking powers have not yet been used by eSafety. However, I am not recommending any changes to this scheme and do not consider their lack of use to be a reason to revoke an important crisis response power.

I acknowledge that the Act sets a very high threshold for exercising the blocking powers without procedural fairness requirements, but these powers are intended to operate as a time-limited response in a crisis situation to prevent the rapid distribution of material online:

It would be used under circumstances where such material is being disseminated online in a manner likely to cause significant harm to the Australian community and that warrants a rapid, coordinated and decisive response by the online industry.[79]

Approaches to addressing online hate

Hate speech is not new, but its prevalence online and its ability to spread at a magnitude and order not seen before, is worrying. Online hate has the potential to cause significant harm to individuals and impact community safety more broadly. After hearing the experiences of individuals and community groups, it is clear that further regulatory intervention is needed to address the harms arising from online hate.

[H]ate speech … can be disseminated like never before, worldwide, in a matter of seconds, and sometimes remain persistently available online.[80]

Around the world, we are seeing a disturbing groundswell of xenophobia, racism and intolerance – including rising anti-Semitism, anti-Muslim hatred and persecution of Christians. Social media and other forms of communication are being exploited as platforms for bigotry.[81]

78 *Online Safety Act 2021*, Part 8.

79 Online Safety Bill 2021, Explanatory Memorandum.

80 Delfi AS v Estonia App. no. 64569/09 ECHR, 16 June 2015, 110.

81 United Nations (2019). United Nations Strategy and Plan of Action on Hate Speech, 1.

Experiences of online hate

Throughout the review, I consistently heard of the high volumes of online abuse and hate that the Australian community is experiencing. Abuse targeting individuals or groups is often based on one or more protected characteristics, in particular age, sex, sexuality, sexual identity, race, religion, or disability. Community groups and regulators have described growing amounts of abuse, which is often triggered by current events such as the COVID 19 pandemic, Australia's recent referendum, and conflict in the Middle East.

While there will never be a legislative solution that addresses all varieties of online abuse, what is of particular concern is the racist, sexist, and homophobic abuse that occurs, particularly when it is directed anonymously… An example of where the Act currently falls short, is where online abuse is directed at an individual, but uses collective group language[82]

Online trolls target perceived vulnerabilities, sometimes masking their attacks by using commentary, tropes or images that have a coded or specific meaning or by hiding behind anonymous accounts.

Community consultation also highlighted the context specific nature of online hate, causing some to raise concerns about the risk of religious commentary being censored, or complaint schemes being weaponised to stifle political debate.

Regulatory approaches to online hate

All Australian jurisdictions have frameworks to deal with forms of online and offline hate speech through anti-discrimination, anti-vilification and incitement laws. Vilification definitions and protected characteristics vary between federal and state laws.[83] These variations were helpfully summarised by Purpose[84] (reproduced at Appendix E). Although hate speech is not specifically criminalised under federal law, other criminal laws may apply. These include offences for urging violence, using a postal or carriage service to menace, harass or cause offence, and advocating terrorism. Some relevant existing and proposed legislation includes:

- The *Racial Discrimination Act 1975* is the only federal anti-discrimination law with a hate speech provision. Section 18C makes it unlawful to do an act, otherwise than in private, which is 'reasonably likely' to offend, insult, humiliate, or intimidate another person or group on the basis of their race, colour or national or ethnic origin. This 'racial hatred' is treated as a civil wrong, but does not address anonymous posts because it requires the person posting to be identifiable.
- The proposed Criminal Code Amendment (Hate Crimes) Bill 2024 would introduce crimes for threatening violence against groups, or members of groups.[85] These amendments would broaden the coverage of existing offences for urging violence against groups or members of groups distinguished by race, religion, nationality, national or ethnic origin or political opinion to include groups distinguished by sex, sexual orientation, gender identity, intersex status and disability as protected characteristics.
- The proposed Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024[86] would address harms arising from disseminating material online that is reasonably verifiable as false, misleading or deceptive, and is reasonably likely to cause or contribute to serious harm of a specified type (misinformation and disinformation). 'Serious harm' in this context would include "vilification of a group in Australian society distinguished by race, religion, sex, sexual orientation, gender identity, intersex status, disability, nationality or national or ethnic origin, or vilification of an individual because of a belief that the individual is a member of such a group." The measures proposed focus on systems and processes for digital communications industry participants, rather than on specific items of content.

82 Review submission 153 – Australian Football League, 2.

83 Purpose (2023). Online Hate Speech in Australia: The Role of News Media and Pathways for Change. Part Two: Curbing Dehumanising Hate Speech Online, Online Hate Speech: Role of Media and Pathways for Change (purpose.com), 11, 20, 23.

84 Ibid, 23.

85 Criminal Code Amendments (Hate Crimes) Bill 2024.

86 Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024.

While online platforms can foster positive and inclusive spaces, they are often spaces where racism and dehumanisation occur and misinformation is spread.[87] Platform design, including recommender systems, can also influence the nature of online communications by favouring incendiary or extreme content.[88] Some online service providers have policies around online hate and allow users to report content they believe might be in violation of these policies. Policies outlined in the terms of service or community guidelines of major online platforms often limit or prohibit hateful speech which targets people based on a range of protected characteristics. These could include age, race, religion, ethnicity, caste, national origin, disability, sex, gender, gender identity, sexual orientation, immigration status, veteran status or serious disease. However, online hate is highly contested and context dependent – the policies vary across platforms and are not always enforced.

Platforms' responses to address online hate and other harmful content also vary. Online service providers are increasingly focused on implementing proactive detection technologies to remove harmful content before users see it, and implementing a range of reporting tools and content moderation systems to support removal of harmful content if made public. There was significant variation described across platform reporting tools, content moderation systems (human and automated) and trust and safety resourcing. Some platforms explained that minimising the reach and effects of harmful content may be preferable to content removal.

The focus is on minimising where we may be adding to or exacerbating the effect of harmful content.[89]

The Act does not directly address online hate but provides some protections through its existing complaint schemes. The Basic Online Safety Expectations also set out the Government's expectations that industry ensure services are safe for Australians and require greater transparency around services' safety measures, including measures to enforce their terms of use which usually prohibit the posting of online hate.

Proposed amendments to the Act

There are challenges when it comes to regulating online hate. These include the difficulty of defining online hate (including for global platforms whose policies need to reflect local contexts) and potential impacts on freedom of speech. There are also concerns about overloading the regulator if the volume of complaints significantly increased through new or amended schemes. Despite these challenges, the Act should be amended to complement broader Government measures addressing online hate. This should include defining online hate material, making improvements to the complaint schemes and enhancing online service providers' obligations in relation to systems or processes through an overarching duty of care and due diligence.

There are different views about what constitutes online hate. Defining online hate in the Act would provide greater certainty about when an online post exceeds a threshold deemed acceptable by Australia's Parliament. It ensures the definition is adapted to the online environment and members of the Australian community are protected from online hate irrespective of where they live in Australia. The definition of online hate could also be considered in interpreting whether the duty of care has been met, and the threshold should not be different for public figures.

The proposed definition encompasses community groups who were identified through the review as disproportionately experiencing online abuse but who are not currently protected by Commonwealth vilification laws. The included protected characteristics also align largely with characteristics proposed through the Criminal Code Amendment (Hate Crimes) Bill 2024.

A possible definition:

Online hate material is material which an ordinary reasonable person in the circumstances would conclude contains an online attack against a person or people – rather than ideas, concepts or institutions – on the basis of race, religion, age, sex, sexual orientation, gender identity, intersex status, disability, nationality, national or ethnic origin (a 'protected characteristic').[90] An 'attack' includes violent or dehumanising material, harmful stereotypes, statements of inferiority, expressions

87 Australian Human Rights Commission (2022), National Anti-Racism Framework Scoping Report, National Anti-Racism Framework Scoping Report 2022, 131.

88 Munn,L (2020), Digital Cultures Institute, New Zealand, 'Angry by design: toxic communication and technical architectures' https://doi.org/10.1057/s41599-020-00550-7. Angry by design: toxic communication and technical architectures | Humanities and Social Sciences Communications (nature.com).

89 Online Safety Act Review Stakeholder Engagement Meeting.

90 Under the proposed definition, it is expected that people of short stature would be captured under the disability limb.

of contempt, disgust or dismissal, cursing and calls for exclusion or segregation.

Dehumanising material is material produced or published which an ordinary person would conclude portrays the class of persons identified on the basis of a protected characteristic as not deserving to be treated equally to other humans because they lack qualities intrinsic to humans.

Online hate directed at individuals or at groups should be proactively addressed through an overarching duty of care and due diligence. Most community groups emphasised the importance of prevention rather than acting after the harm has occurred, particularly given the volume of online abuse experienced. However, a safety net should also be available to address online hate directed at an individual in Australia.[91] Individuals and community representatives described harms they experienced arising from persistent, targeted and volumetric online hate attacks that seem to far exceed 'mere ordinary emotional reactions.'[92] They expressed frustration that reported attacks had not been found to meet thresholds for regulatory or law enforcement intervention.

While many of these concerns may be addressed by lowering the adult cyber abuse threshold, the regulator should also be explicitly enabled to consider the cumulative harm arising from online hate in determining whether material meets the threshold for complaint-based removal schemes. Online hate material should be defined in the Act so that the cumulative harm can be appropriately considered under the adult cyber abuse scheme. Regulation through online service providers means the regulator is not required to identify the person who posted the online hate material, one step in ensuring attacks from anonymous accounts are addressed.

While measures to address online hate may raise concerns about excessive content moderation or vexatious complaints being used to stifle public debate, other international jurisdictions have shown how concerns can be addressed through enhanced content moderation, transparency reporting, providing appeal mechanisms for the people who post moderated content, and using external dispute resolution frameworks (see also 7.6).

91 The threshold should not be higher for public figures. This is discussed further below.

92 'Mere ordinary emotional reactions' are currently excluded from the definition of serious harm to a person's mental health. Online Safety Act, section 5.

Case Study: Online abuse of high profile women

'From trolling and harassment to threats of rape and even death, Tara Rae Moss has seen the very worst of social media. Her ghastly brush with online platforms led the best-selling crime author, model and human rights advocate to write, produce and present the TV series Cyber Hate in 2017. Across six episodes, she exposed the toll the online trolling and aggressive social behaviour had taken on her. It continues to this day. "I had many death threats over the years. That's never OK, no matter who you are, but is particularly not OK if aimed at kids," Moss, 51, said. One of the worst messages she received on Twitter read: "Have you no shame, whore? Lying about being raped to sell your garbage book? I hope you do get raped for your lies." Moss joins other high-profile Australians in backing Unplug24, a campaign to boycott online platforms for 24 hours on October 24, the first anniversary of Mac Holdsworth's death. Mac, 17, took his life in 2023 after being "tortured and terrorised" on social media. "Everyone gets negative comments in life, but some comments cross the line, and orchestrated hate campaigns, pile-ons, and death threats can be particularly dangerous," Moss said. "It's important to highlight the importance of taking time out from screen time and social media, particularly for kids who may not have known a world without the kind of tech that can now fill our lives 24/7.

Courier Mail, Brisbane, Wednesday 23 October, 2024[93]

93 'Death-threat survivor Moss backs phone switch-off' – Courier Mail, Brisbane, Wednesday 23 October 2024.

21

Recommendation 21:

The Act should include a definition of online hate material. The definition should acknowledge that online hate involves an attack against a person or people that is based on a protected characteristic and can include dehumanisation. Notably, the definition of online hate material should not include views regarding ideas, concepts or institutions. The definition should also consider potential exclusions (for example where material is posted for artistic, scientific, or journalistic purposes), and potential impacts on the constitutional implied freedom of political communication.

22

Recommendation 22:

The Act should be amended to ensure that, in interpreting the threshold of harm for adult cyber abuse, the reasonably proximate cumulative harm caused by online hate material is taken into account.

Throughout the review, I carefully considered a complaint scheme to enable removal of online hate material targeting groups. This prospect of 'widening the aperture' would be expected to significantly increase the volume of complaints received which could delay complaint handling and draw resources away from eSafety's other regulatory functions. One suggestion considered was to limit complaints to 'trusted flaggers' (approved government or civil society entities). However, this would also place a greater resource burden on those trusted groups. The review also heard concerns about the burden that might be placed on community groups to report online hate.

A cultural shift is needed to address the scale of abusive communications, including online hate material, and this is likely to be more effective by strengthening systems regulation through a statutory duty of care. Prevention of online harms through a duty of care would reduce reliance on complaint-based removal schemes, minimising their impact on the regulator and reducing the reporting burden on targeted communities. Where a service repeatedly fails to take down hateful content, whether aimed at individuals or groups, there would be grounds for eSafety to take legal action for a breach of the duty of care and due diligence requirements.

The recommendations above that lower the regulatory thresholds and expand the scope of the complaint-based removal schemes will also provide additional protections to those individuals who are on the receiving end of online hate. There are also opportunities to streamline the regulatory investigation and response processes to better reflect the cumulative nature of online abuse and address volumetric (or 'pile-on') attacks as a whole (see below), rather than evaluating and acting on individual items of abusive material.

Volumetric ('pile-on') attacks

Volumetric (or 'pile-on') attacks often involve abusive posts connected with the target, which others like, share or repost with additional commentary, and they sometimes involve coordinated and/or disingenuous behaviour. Often the content is shared with an accelerating level of outrage and toxicity, and ultimately a high volume of abuse. These attacks can be among the most serious forms of online abuse.[94] The harm of individual comments can be damaging to the targeted user's wellbeing, and when done on an extensive scale through volumetric attacks, the impacts can magnify and compound.

Through the review I heard many individual experiences of online abuse which included volumetric or 'pile-on' attacks. While these stories were shared in confidence, this account from the House of Representatives Select Committee on Social Media and Online Safety demonstrates the shared experience and long-term impact.

The Act does not define a 'volumetric attack' or 'pile-on' attack. Most experiences of online abuse described involved 'pile-on' attacks, where posts from a large number of people target an individual or smaller group. However, coordinated attacks that occur from a small group of people (such as online trolls) or a single source (such as a bot-generated attack) can also have a similar cumulative impact, inundating the target with an overwhelming volume of attacks in a short period of time. A definition of volumetric attacks must capture the breadth of high-volume attacks while providing certainty of meaning to industry, online users and the regulator.

The distribution of harmful content by various individual users and across different platforms means there is no single point for regulatory action. Under the Basic Online Safety Expectations, service providers are expected to consult and collaborate to promote safe use, including working with other service providers and between services to detect high volume, cross-platform attacks. I found no examples of this collaboration during the review and consider that stronger measures are needed to address the breadth and persistence of attacks people in Australia are experiencing.

Complaints through the child cyberbullying and adult cyber abuse schemes are evaluated by assessing each individual post against thresholds for regulatory action. The intention to cause a volumetric attack, or the fact that a volumetric attack has occurred, may be relevant considerations in an investigator's evaluation, but they may not be able to consider the full scope of an attack across platforms and from multiple accounts.

For the adult cyber abuse scheme, the threshold of intent to cause serious harm may limit an investigator's ability to consider the full extent of a volumetric attack. For example, an end-user who posts abusive material might not be aware of similar attacks occurring on other platforms. I have recommended removing this threshold for adult cyber abuse, instead an investigator should focus on whether the post is menacing, harassing or seriously offensive in all the circumstances.

94 Parliament of the Commonwealth of Australia, House of Representatives Select Committee on Social Media and Online Safety, 'Social Media and Online Safety' March 2022, 17, Social Media and Online Safety – Parliament of Australia (aph.gov.au).

Case Study: Volumetric attacks

The first online attack I received came after my first-ever media appearances on national television. The abuse was predominantly racist in nature, and some of the abuse used such violent language, including calling for the culling of people who look like me. I remember taking screenshots of the pictures of some of the individuals who directed the worst abuse, hoping that, at the very least, I might avoid them in public.

The second attack was more sustained and reached every presence I had online. In what the eSafety Commissioner described at a Senate hearing as 'volumetric attack', I was tracked across all social media platforms and trolled predominantly with racist abuse … This time, though, the abuse and many things that were happening made me take three months off from work. The online abuse was not the only reason, but it played a substantial role in me taking the time to literally try to heal and reconnect again with a sense of safety. Because of that, I no longer share pictures of my children online, I prefer that my family members do not follow me online so they do not receive abuse, and I am constantly on watch to remove abuse that pops up on almost a daily basis.[95]

95 Parliament of the Commonwealth of Australia, House of Representatives Select Committee on Social Media and Online Safety, 'Social Media and Online Safety' March 2022, quoting Nyadol Nyuon, 41, Social Media and Online Safety – Parliament of Australia (aph.gov.au).

23

Recommendation 23:

The Act should define a 'volumetric attack' and the regulator should be empowered to issue a notice or notices to multiple platforms based on a single complaint to address volumetric attacks.

24

Recommendation 24:

The Act should be amended to provide the regulator with the ability to issue a notice to services in relation to a suspected 'volumetric attack', which may require information related to the attack, specify remedial actions to be taken and require the service to report back on steps taken.

Strengthen the Act to better support public figures who experience online abuse

Public figures and people with a public profile are subject to high rates of online abuse and harassment and are often at greater risk of online abuse than everyday private individuals.[96] Women, minority public figures and civil society advocates and activists are among the most targeted.[97]

Surges in online abuse are frequently linked to current news, events and other external drivers such as sports betting. Attacks often target one or more personal characteristics (online hate) rather than political issues and extend to the families of those targeted, including children.

Online abuse targeting public figures, including trolling, stalking, impersonation accounts, image-based abuse and sexual harassment, can have serious professional and personal impacts.[98] While suicide is a complex phenomenon that often cannot be reduced to a single cause or underlying factors, in several cases, online abuse of public figures has preceded suicide.[99] The abuse may also force public figures to withdraw from public life, and stifle the quality of public debate by making it more difficult for public figures to participate safely in online discourse.[100]

The impact of online gendered harm extends to elite sportswomen's online behaviour, participation and feelings of safety. Athletes closed social media accounts (permanently or temporarily), avoided certain social media platforms, stopped posting about certain topics, spent less time online and edited posts to avoid backlash.[101]

Targeted journalists and politicians, particularly women and minority groups, are withdrawing from their roles because of the volume of abuse. Women in local politics, where less support structures are available, describe the toll as being too much. For women journalists, this phenomenon has been coined 'the chilling effect' – the 'chilling' of women's active participation in public debate is described as a threat to the public's right to information and an attack on media freedom and democracy.

As noted in the Issues Paper, online abuse has been described as having a "chilling" effect on women journalists' active participation in public discourse, which is detrimental to media freedom and a threat to democracy. International surveys have reported that 48% of women journalists self-censor, 22% close media accounts and nearly a third consider leaving the profession as a result of online abuse.[102]

Public figures, such as journalists, sports people or politicians often have a professional requirement to be active online and engage with a range of social media platforms. Given this dependence, they may not have the option to remove themselves from abusive online environments. High-profile exposure combined with potential attention on the content they post, increases a public figure's risk of exposure to online abuse. As raised by one public figure:

There is nothing I can say that is safe.[103]

Under the existing Act, high volumes of online abuse may compound into volumetric attacks but not individually meet thresholds for adult cyber abuse. This leaves the targeted person reliant on assistance from their employers, social media screening from friends, family or staff and online services. Platform policies are unclear about how they define public figures, and definitions across platforms are inconsistent. Where defined, platforms often provide fewer protections to public figures on the basis of freedom of expression or public interest. Most policies of larger platforms[104] reflect a higher

96 Cover, R, Henry N, Gleave J, Greenfield S, Grechyn V (2024), 'Protecting Public Figures Online: How Do Platforms and Regulators Define Public Figures?', Media International Australia, 0(0):1-15.

97 Ghaffari, S. (2022), 'Discourses of celebrities on Instagram: digital femininity, self-representation and hate speech', Critical Discourse Studies, 19(2):161-178.

98 eSafety Commissioner (2023), 'What is online abuse?'.

99 Cover et al (2024), 'Protecting Public Figures Online: How do Platforms and Regulators Define Public Figures?', Media International Australia, 2.

100 Ibid, 3.

101 Toffoletti, K, McGrane, C, Reddan, S (2024). Addressing Online Harm in Australian Women's Sport, Deakin University, Report.

102 Review submission 138 – Australian Broadcasting Corporation, 3.

103 Online Safety Act Review Roundtable.

104 Meta reports that it has recently updated its policy in this regard.

threshold for addressing online harms directed at public figures than everyday users. Often, platforms do not differentiate between different types of public figures and fail to acknowledge the varying levels of resources and support each has available.[105]

A lot of people say that that [abuse] comes with being an AFL player. But being bullied or discriminated against is not in the job description.[106]

Through the review I heard that public figures are often reluctant to report online abuse and that the harms they experience are not adequately addressed through existing law enforcement and regulatory frameworks. Sports people were reluctant to report because they wanted to focus on their sport, didn't want to distract the team by raising individual issues, or didn't think reporting would help. Distrust in authorities or past experiences reporting can also influence reporting decisions. Both individuals and employers found that if the abuse could not be controlled through personal or organisational online screening mechanisms, there was limited assistance available.

The online abuse of public figures is a significant and ongoing issue. While I encourage online platforms to assess their policies around online hate and public figures, broader changes proposed in this review will also help to reduce harms and allow public figures to seek support. These include a duty of care, changes to the adult cyber abuse scheme (where public figures are treated the same as other people) and new considerations for the impact of volumetric attacks. The eSafety Commissioner may also use transparency powers under a duty of care to investigate platforms' treatment of public figures and could ultimately issue a code if the harms go unchecked and there isn't a sufficient response.

Case Study: Online abuse of sports professionals

My first experience with online abuse was on Instagram. I was in my second or third year in the AFL and I was new to the platform. I had only been on it for less than a year before I started seeing a few comments on two or three of my recent photos I'd posted, with someone commenting 'monkey' and using the monkey emoji.

Seeing those comments took me back to my childhood, when being abused on the field was fairly commonplace for me. The thing most people don't realise is how inhuman it makes you feel. From those experiences when I was a kid, to that first comment on Instagram, the underlying thing through it all was how it made me feel like I was less than human. Othered.

But it is different when it's online. When I was growing up and copping abuse face- to-face, you always try to brush it off, but at least it's direct — you know who it's coming from and you can deal with it. You can see their faces. You can even try your best to educate them, put them through a cultural awareness session, help them understand why it's offensive and the trauma that's involved.

But when it's an online troll, or a burner Instagram account, you feel like there's no accountability and you can't do anything about it. You feel uneasy and helpless. These people just do the damage and then go about their day without any consequences.

[Abuse] happens to a lot of players in the league, whether it's people of colour, or about their sexuality, anyone who is in a minority is getting slandered for simply being who they are.[107]

107 Chad Wingard (2022), 'As an Indigenous AFL player, I've faced abuse my entire career', GQ Magazine.

105 In circumstances where public figures are supported by employers or others, it is then the supporting individuals who are experiencing the harmful content in place of, or in addition to the public figure.

106 Review Submission 153 – Australian Football League, 1 (quoting Chad Wingard).

© Getty Images. Credit: Mats Silvan.

Case Study: Online abuse of women in government

' ... [one female politician, who has now left politics because of the abuse she received] has been subject to persistent online violence. In an interview with Vice in 2018, [the politician] expressed how the online abuse was overwhelming and questioned how long she would continue in Parliament ... In an Australian study, women MPs were found to be disproportionately targeted by public threats, particularly facing higher rates of online threats involving sexual violence and racist remarks ... Male politicians are also subject to online violence. But when directed at women the violence frequently exhibits a misogynistic character, encompassing derogatory gender-specific language and menacing sexualised threats, constituting gender-based violence ... Without legal recourse, women MPs have two options – tolerate the torrent of abuse, or resign. Both of these options endanger representative democracy ... '

The Conversation, Friday 19 January 2024[108]

108 '[Female politician's] exit from politics shows the toll of online bullying on female MPs' - The Conversation, Friday 19 January 2024

7.5 Striking a balance between protections and freedoms

Adverse impacts on freedom of speech and Australia's democracy

As noted by the Law Council of Australia:

Freedom of expression is associated with other human rights, such as the right to freedom of thought, conscience and religion, and the right to freedom of association — it is the cornerstone of a free and democratic society.[109]

Protecting freedom of speech or expression (and the implied freedom of political communication in Australia's Constitution) was a key concern raised through the review. A number of submissions raised concerns that content moderation limits freedom of speech, with some calling for reduced regulation, abolition of the Act or appropriate governance and oversight of regulatory removal powers.[110] Others described the silencing effects of online abuse, and adverse impacts on their work, health, relationships and personal security. Increased reliance on the internet in Australia to access services means that withdrawing from the internet risks marginalising and disadvantaging targeted groups.

The Australian Human Rights Commission recommended the review consider "the human rights impacts of proposed reform, including specifically the impact on freedom of expression."[111] The Commission noted that all human rights are indivisible and afforded equal status, but that freedom of expression requires specific consideration in online spaces because of the opportunities digital platforms provide for realising the benefits of free speech.[112] The Commission also referenced a United Nations Human Rights Council resolution of 2018[113] which called on member states to protect access and dissemination of information online, while also stressing the importance of combatting advocacy of hatred on the internet.[114]

It is acknowledged that for freedom of speech to flourish online, the 'digital town square' in which discourse occurs should be a safe space for expression. If not, the voices of marginalised groups may be silenced out of fear in engaging in hostile online spaces.[115]

In 2016, the Australian Law Reform Commission explored proportionality of laws that limit freedom of speech, noting:

Free speech and free expression are understood to be integral aspects of a person's right of self-development and fulfilment … At the same time, it is widely recognised that freedom of speech is not absolute.[116]

Several submissions raised the importance of online safety laws being proportionate to the harms regulated.

109 Review Submission 149 – Law Council of Australia, 113.

110 Review Submission 107 – Institute of Public Affairs; Review Submission 98 – Australian Christian Lobby; Review Submission 63 – Free Speech Union of Australia; Review Submission 122 – Affiliation of Australian Women's Action Alliances; Review Submission 34 – LGB Alliance Australia.

111 Review submission 135 – Australian Human Rights Commission, 19.

112 Ibid, 18.

113 United Nations Human Rights Council, 'The Promotion, Protection and Enjoyment of Human Rights on the Internet, 38th sess, UN Doc. A/HRC/38/L.10/Rev.1 (7 April 2018).

114 Review submission 135 – Australian Human Rights Commission, 18.

115 Ibid, 19.

116 Australian Law Reform Commission (2016) 'Traditional Rights and Freedoms – Encroachments by Commonwealth Laws' Report 129, 12 January 2016, Common law foundations | ALRC, 4.3-4.4.

To optimise regulatory efficiency and minimise burdens on platforms, we strongly advocate for coherence and alignment across regulatory frameworks. In this respect, it is critical that online safety regulation should be reasonable and proportionate to the harms it seeks to address and must be balanced against users' rights to privacy, free expression and access to information.[117]

Proportionality considers whether a law is necessary and suitable to achieve a legitimate objective, and whether the public interest of a law outweighs limitation to an individual right.[118] For example, the International Covenant on Civil and Political Rights recognises limitations to freedom of expression where laws are necessary to respect the rights or reputations or others, or for the protection of national security, public order, public health or morals.[119] The review also heard concerns that the right to free speech could be misrepresented to avoid accountability for online harms:

A misguided interpretation of the right to free speech (including by free speech absolutists) has been weaponised to avoid accountability for the harms caused by abuses of the right to free speech. The Australian Government must not be dissuaded from pursuing a comprehensive regulatory regime by such arguments.

The right to free speech ought to be understood in relation to other fundamental rights, including the right to freedom of thought and conscience, right to information, right to participate in public affairs and the right to vote, among others.[120]

A broad range of human rights interact with freedom of speech. Some examples include: freedom of opinion and expression, freedom of thought, conscience and religion or belief, right to take part in public affairs and elections, right of privacy and reputation, right to health, rights of equality and non-discrimination and the rights of the child.[121]

More essentially, the disproportionate harms experienced by First Nations peoples, and other minority groups of the Australian population, undermine attainment of the concepts of human dignity, the right to equality and other fundamental freedoms …

Article 12 of the [United Nations' Universal Declaration of Human Rights] and Article 17 of the [International Covenant on Civil and Political Rights] provide that no one shall be subjected to arbitrary or unlawful interference with his or her privacy, family, home or correspondence, nor to unlawful attacks on his or her honour and reputation. In this context, the current limitations of the framework for online safety require assessment and reform. For example: the current framework is such that the Commissioner may not be able to intervene in situations where a person may be affected by abusive posts targeted at a group of people, such as dehumanising commentary on a particular race or belief …[122]

Community experiences shared through the review indicate that additional regulatory intervention is required to ensure the 'digital town square' in which discourse occurs is a safe space for expression. As noted in an *Australian* editorial, **"There should be no room for hate speech, vilification, bullying or abuse online or in public debate."**[123]

117 Review submission 52 – Reddit, 6.

118 Australian Law Reform Commission (2016) 'Traditional Rights and Freedoms – Encroachments by Commonwealth Laws' Report 129, 12 January 2016, 2.63.

119 International Covenant on Civil and Political Rights, Article 19.

120 Review Submission 155 – Human Rights Law Centre, 7.

121 Ibid, 7-8.

122 Review submission 149 – Law Council of Australia, 27.

123 The Australian, Editorials, 'Silencing Free Speech is a Bad Idea', 27 September 2024.

Many of those experiencing online harm do not meet the threshold for regulatory action, as noted by Youth Law Australia:

It is very common for children and young people … to have experienced cyberbullying over a lengthy period of time, involving peers, often in groups, doing things like sending hurtful, abusive or threatening messages, spreading lies or embarrassing stories or images, or creating fake accounts and impersonating them. In the majority of these cases the statutory threshold in the Act of 'would be likely to have the effect of … seriously threatening, seriously intimidating, seriously harassing or seriously humiliating the Australian child' is not met and the child or young person cannot access relief under the Act like removal notices … there are also rarely grounds for a police response … Typically, the material will also be found to not breach the social media platforms' community guidelines or codes of conduct, so the child or young person may then have to exercise self-help like changing schools to get away from bullies or withdrawing from certain online spaces. In many cases our clients are self-harming or dealing with suicidal feelings on a regular basis …

[W]e also observe that many clients experiencing adult cyber abuse do not meet the statutory threshold for relief under the Act being that an ordinary reasonable person in the position of the targeted Australian adult would regard the material as being, in all the circumstances, menacing, harassing or offensive.[124]

Content and activity moderation, whether through online services own processes or complaint-based regulatory removal schemes, enables rapid removal of harmful material from the online environment. With appropriate oversight and guardrails, content moderation can facilitate rather than prevent participation in political debate. Online safety regulations are balanced with protections against other online harms, such as privacy and cybersecurity laws. Additional regulatory oversight is provided through independent review pathways such as administrative or judicial review of decisions to remove online material. I consider these guardrails proportionate to a regulatory scheme intended to remove exposure to harmful online content.

© Getty Images. Credit: SolStock.

124 Review submission 161 – Youth Law Australia, 9.

7.6 Dispute resolution

Access to good dispute resolution mechanisms is an important part of how we protect people in Australian society.

The eSafety takedown schemes don't catch all types of bad conduct and even world class systems for platforms are not 100 per cent foolproof. In these circumstances, there needs to be somewhere people can go to resolve disputes. This includes people whose posts have been removed who believe they have been taken down unfairly as well as people who have failed to have posts that harms them or their group taken down.

People who are unhappy with a decision made by eSafety can request an internal review, complain to the Commonwealth Ombuds or the Administrative Review Tribunal. People who are unhappy with a decision by a platform or search and app distribution service also need somewhere to take their complaints.

Our laws require financial institutions, energy companies, telcos and others to be members of Ombuds schemes. I believe it should also be a requirement for all online platforms and search and app distribution services to have:

- A simple, user friendly way to make a complaint to the service
- An internal dispute resolution scheme to resolve complaints and respond to the complaint within a reasonable time; and
- An Ombuds scheme where people can go to get a decision if they aren't happy with the service's response or the service fails to respond within a set time.

This recommendation builds on a number of recommendations made by the Australian Competition and Consumer Commission (ACCC) about the need for Internal Dispute Resolution schemes and a Digital Ombuds Scheme in their work on Digital Platforms. In their 5th report, Digital Platforms Services Inquiry, September 2022 they recommended that there be:

- Mandatory internal dispute resolution standards that ensure accessibility, timeliness, accountability, the ability to escalate to a human representative and transparency; and
- Ensuring consumers and small business have access to an independent external Ombuds scheme.[125]

The ACCC reiterated its support for this recommendation in report 6 on social media and report 7 on ecosystems.[126]

While the ACCC was focused on small business and consumer complaints, any Internal Dispute Resolution scheme or Ombuds scheme developed in the digital space, should focus on the full range of harms, including those dealt with in this report.

That more is needed was recognised by Meta in their February 2020 white paper where they noted that:

> *Regulation could also incentivise – or where appropriate, require – additional measures such as … a channel for users to appeal a company's removal (or non-removal) decision on a specific piece of content to some higher authority with the company or some source of authority outside the company.*[127]

Since then they have established an Oversight Board to review Meta's enforcement decisions, though it can only look at a very small number of complaints.

In the Government's 2024 response to the ACCC's Digital Platform Services Inquiry, the Government supported in principle the ACCC's findings that digital platforms "do not have adequate processes for consumers to raise issues and concerns experienced online. A lack of effective dispute resolution processes can reduce trust and confidence in digital platform services and prevent Australians from taking full advantage of the benefits provided by digital platforms. The Government will undertake further work to develop internal and external dispute resolution requirements by calling on industry to develop voluntary internal dispute resolution standards by July 2024." [128]

125 ACCC (2022), Digital platform services inquiry - September 2022 interim report - Regulatory reform, 16.

126 Ibid, 88, 177.

127 Facebook (2020), Charting a way forward: Online Content Regulation, 10. Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.

128 Australian Government (2023), Government's response to the ACCC Digital Platform Service's Inquiry, 2-3.

Work is being done by industry and the Department to develop standards on internal dispute resolution for digital services. This work is being led by DIGI, the industry body, and the Department.

Prior to the Government's response, the Department had been doing a considerable amount of work looking at whether or not there is a need for an Internal Dispute Resolution code and for an Ombuds scheme to be established.

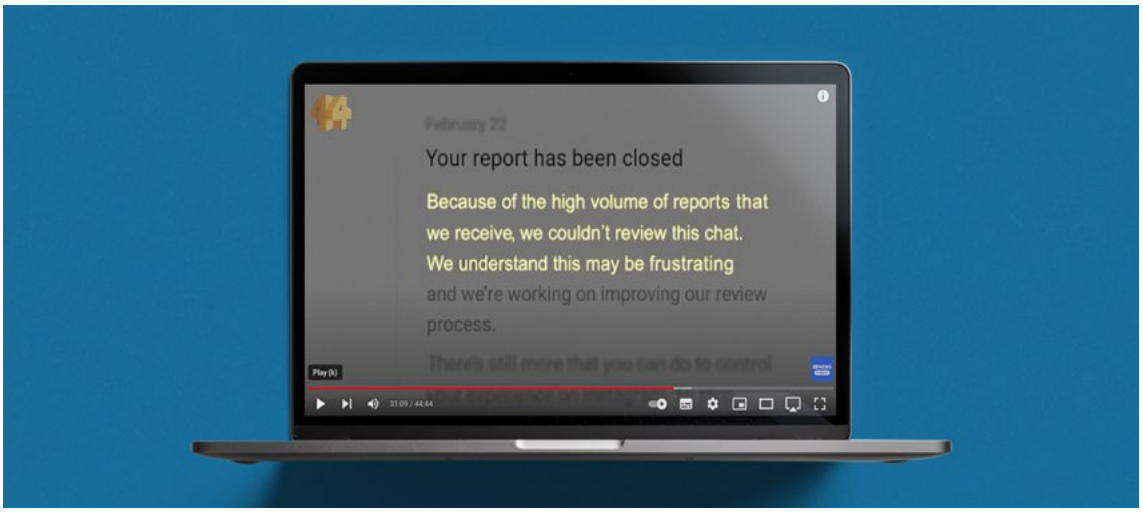
This work included a feasibility study on establishing an Ombuds scheme and work on the cost to the economy of not having effective dispute resolution processes for Australians to access. Back in 2020 they commissioned Accenture to survey Australians to better understand the issues around dispute resolution. The report has not yet been publicly released. They surveyed 8,334 consumers over 18 years old and 1,471 small businesses about their complaints experience. While the focus was not on online safety complaints, the results are still telling.

They found that:

- 2,988 consumers (36 per cent) and 500 small businesses (33.9 per cent) had experienced an in scope issue
- 2 in 3 issues are reported to platforms to resolve
- Adjusted for population size, Australian consumers and small businesses had experienced an estimated 4.9 million in scope issues in 2020. Of these:
  - 1.6 million were resolved by users through user led resolution tools
  - 2.4 million were reported to digital platforms, which resolved 1.6 million or 66 per cent of issues reported to them
  - Of the unresolved issues, 304,000 were escalated to an external body, which resolved 217,000 or 71 per cent of issues reported to them in this way (or, a further 12 per cent of issues initially reported to platforms); and
  - 25 per cent of reports to platforms were unresolved by either the platform or the external body.[129]

The Department's work shows that current digital dispute resolution landscape cost the economy an estimated $4.2 billion in lost time in 2020, of which $3.7 billion was borne by Australian consumers and small businesses. This is in addition to $188 million in direct financial losses incurred by small businesses ($101 million) and consumers ($87 million) – arising, for instance, from scams or lost advertisement spend – and an unknown number of sales and opportunity losses.[130] With the recent increase in online scams, that number is now likely to be considerably higher by a number of multiples.

While these are numbers, an example is always helpful. Perhaps the worst I have seen in this area is when a mother who ran a 'kidfluencer' blog received the following response from Instagram when she complained about comments made by a paedophile in relation to the blog.



Source: Four Corners Series 2024 Kidfluencers

129 Australian Government (2021), Digital platforms industry external dispute resolution scheme: Feasibility study and design project final report, 5-19. [Released under FOI].

130 Ibid.

Having both internal and external dispute resolution schemes is important, including because industry is the one who normally pays the costs of external dispute resolution and this creates the right incentives for services to put in place high quality, fair and effective internal dispute resolution to avoid having to pay the per complaint costs of an external Ombuds scheme.

None of the existing external regulatory bodies, be they the ACCC, eSafety or state and territory consumer protection agencies are set up to deal with the number of disputes that occur in the digital space.

Ombuds schemes, on the other hand, are experienced in handling large numbers of complaints with, for example, the Australian Financial Complaints Authority handling over 104,861 financial complaints in 2023–24.[131] Technology is also helping them to deal with complaints more expeditiously while also helping to identify systemic issues that can be raised with members with a view to dealing with and resolving the issue.

Australia is not the only country looking at this issue. In October of this year the establishment of the Appeals Centre Europe, was announced. It is an out of court dispute resolution settlement body. It has been set up under the EU Digital Services Act. It is backed by Meta and is due to decide cases relating, at least initially, to Facebook, TikTok and YouTube.

If Australia is to set up a Digital Platforms Ombuds scheme it should cover all the types of consumer issues dealt with in the ACCC's work on digital platforms (noting though that the Government is proposing that scams be dealt with by the Australian Financial Complaints Authority so that the respective roles of telcos, platforms and financial services can be considered together)[132] as well as those covered in this report. At a minimum it should require membership of platforms with the highest reach in Australia. Given that these platforms are almost all based overseas, a licensing requirement would better ensure that services complied with this requirement (see Chapter 10).

Useful guidance on establishing the governance of a good Internal Dispute Resolutions scheme can be found in the Treasury's *Benchmarks for Industry-Based Customer Dispute Resolution*[133] along with ASIC's Regulatory Guide 271.[134] Similarly, ASIC's guide on external dispute resolution, Regulatory Guidance 267[135] is a good starting point for designing an Ombuds scheme.

131 Australian Financial Complaints Authority Annual Review 2023-24, 4.

132 It is intended that the Australian Financial Complaints Authority would be the designated external dispute resolution scheme for harms relation to the three initial sectors under the Scams Prevention Framework. This include banks, telecommunication providers and certain digital platforms (social media, paid search engine advertising and direct messaging services).

133 Commonwealth of Australia, 2015, Benchmarks for Industry-based Customer Dispute Resolution.

134 Australian Securities and Investments Commission September 2021 Regulatory Guide RG 271 Internal dispute resolution.

135 Australian Securities and Investments Commission September 2021 Regulatory Guide RG 267 Oversight of the Australian Financial Complaints Authority.

25

Recommendation 25:

All services should be required to have an easily accessible, simple and user-friendly way to make a complaint and internal complaint handling processes that are in line with a code on internal dispute resolution. In particular, this should include a way for non-users to report issues such as when intimate images have been posted without consent on a service. Services should also be required to respond to reports within a reasonable time and for some issues within 24 hours.

26

Recommendation 26:

In line with the Australian Competition and Consumer Commission's Digital Platform Services Inquiry, the Government should develop and implement an Ombuds scheme that covers digital platforms and online search and app distribution services.

WICKED PROBLEMS

08

The introduction of a Duty of Care obligation on services and strengthening the complaints schemes should go a long way to addressing many online harms and result in a significant uplift in safety for all Australians.

That said, there are some serious harms that will require considerably more work to move the dial. Here I'm particularly thinking about the complex issue of targeted technology facilitated abuse; the increasing use of end-to-end encryption and its implications for being able to deal with child sexual exploitation and abuse material and other illegal material; and sextortion, which is often perpetrated offshore along with many other online scams. These are wicked problems that will require a multi-dimensional approach, and multi-stakeholder approach if we are to make a real difference.

8.1 Technology-facilitated abuse

Technology-facilitated abuse is more than online abuse

'Technology-facilitated abuse' is "using technology to enable, assist or amplify abuse or coercive control of a person or group of people."[136] It can include any form of abuse that is enabled through digital technologies, including, but not limited to, online. This includes where technology is used as part of stalking or monitoring, psychological and emotional abuse (including threats), sexual violence or harassment, defamation, bullying or online hate. Specific forms of technology-facilitated abuse include cyber abuse, cyberbullying and image-based abuse or exploitation.

The review heard community experiences of a broad range of technology-facilitated abuse, including the use of locating software, online harassment, bullying and threats, cyberstalking and use of smart technologies to track or intimidate individuals.

A 2022 survey of Australian adults found that one in two Australians had experienced technology-facilitated abuse behaviours in their lifetime.[137] The likelihood was higher for LGBTQIA+ Australians, First Nations people, Australians aged 18-44 years, and Australians with a disability.[138]

Some types of technology-facilitated abuse, such as bullying and harassment and image-based abuse, are dealt with under the Act, and I have proposed that relevant provisions are enhanced to improve coverage of complaints schemes. However, other types, particularly those relating to the use of technology to stalk people or to create abusive material such as deepfakes, are not adequately addressed.

Gendered abuse and violence

Technology-facilitated abuse can also be gendered in nature, particularly in the context of family, domestic and sexual violence. Australian women are significantly more likely to experience abuse perpetrated by a man than by another woman in their most recent experience. Women are also more likely to experience technology-facilitated abuse from an intimate partner or former partner.[139]

This type of abuse when targeted at women, also known as technology-facilitated gender-based violence, manifests differently to abuse targeted at men. It tends to be violent, sexualised, and can include threats towards a woman's children. It will often target a women's physical appearance, fertility and virtue.

These behaviours can result in physical, sexual, psychological, social, political or economic harms or other infringements of rights and freedoms on the basis of gender characteristics.[140] The potential for digital technologies to cause harm and the nexus with 'real life violence' is also highlighted in the *National Plan to End Violence against Women and Children 2022-2032*.[141]

A key message from public consultation is that people intent on causing harm will use emerging technologies and navigate around system and legal obstacles to achieve their objective. Broader community conversations are required to drive cultural change around gendered abuse and around hate against protected groups. However, online content, commentary and tools can facilitate, normalise and amplify the abuse and the hate. These insights have been considered in recommending an overarching duty of care for online service providers and changes to regulated services.

136 World Economic Forum (2023) Typology of Online Harms, 8 WEF_Typology_of_Online_Harms_2023.pdf (weforum.org).

137 Australian National Research Organisation for Women's Safety (ANROWS), July 2022, Technology-facilitated abuse: National survey of Australian adults' experiences, 8, Technology-facilitated abuse : National survey of Australian adults' experiences.

138 ANROWS, July 2022, 8-9.

139 ANROWS, July 2022, 9.

140 See 130.

141 The National Plan to End Violence against Women and Children 2022-2032 | Department of Social Services, Australian Government (dss.gov.au).

As I have said earlier, expanding the complaint-based removal schemes is also necessary so that a broader range of harms can be addressed under the schemes when needed.

Case Study: Technology-facilitated abuse

'A NSW woman tracked and bugged by her spurned ex-lover has described feeling like she was in a film during the ordeal. The woman, who can't be named for legal reasons, told the ABC the stalking infiltrated every part of her life. "It felt like I was in a Netflix movie," she said. "I thought that I was one of the characters and we all know how it ends — the character either dies or the character prevails … thankfully, in my case, my character prevailed. "There were endless sleepless nights, my mind churning with eternal possibilities, the what-ifs. "What if I didn't continue to push against the system, what if I continued to turn a blind eye to my stalker's clandestine and heinous actions?" … NSW Bureau of Crime Statistics and Research (BOCSAR) data shows there has been a big spike in the use of bugs and trackers in the state over the past two years. BOCSAR executive director Jackie Fitzgerald said the rise could not be underestimated. "When we delve into the figures, we're really seeing that phones are very commonly used and computers and electronic devices are much more commonly associated with stalking and intimidation incidents these days," she said … Victims of Crime Assistance League (VOCAL) practice lead Sophie Wheeler said the spike in technology-facilitated stalking was a big concern. "This is something that is a really serious area in terms of … the far-reaching tools of intimidation and harassment and how perpetrators can use technology to stalk and intimidate," she said … '

ABC News, Thursday 19 September, 2023[142]

142 'NSW stalking victim felt like she was 'in a Netflix movie' amid spike in technology-assisted tracking' - ABC News, Tuesday 19 September 2023.

Abusive technology

However, a different approach is needed in relation to technologies that enable stalking and deepfakes, in particular those that are sexually explicit. Online 'nudify' apps or services allow someone to upload an image of a real person to generate a fake but photorealistic image of what that person might look like undressed. These apps and services can be used to sexualise and humiliate women and girls, and sometimes to create synthetic child sexual exploitation and abuse material.[143] They are often cheap and easy to use. While content generated by deepfake apps and services is captured by removal schemes, given the purpose of this technology, it is difficult to conceive a legitimate purpose and to justify its availability.

Technology, and how people use it, is evolving at a rapid pace. New and emerging technologies pose a particular risk to human rights. eSafety's work in producing Tech trends and challenges position papers plays a pivotal role in educating and raising awareness for people and policy makers across the country, as well as ensuring that eSafety is giving early consideration to emerging technologies and risks[144]

Cyberstalking is particularly insidious and frequently used against women. Apps, services and products available in Australia enable the user to track another person's location and track activities such as texts, calls and internet browsing, and may be undetectable on the owner's device[145] and information about using spyware is readily available online. While there may be legitimate uses for some tracking devices such as those that help you find your phone or luggage or, with permission, see where family and friends are, it is difficult to envisage any legitimate purpose for a stalking app that you can't detect on your phone and which can see all of your communications and online use. These apps should be banned and search engines should no longer show results for them and app stores no longer provide them. Prohibition of apps such as nudify and stalking apps may not sit with the Online Safety Act and will need cross Government consideration. As noted there are an ever-expanding range of tracking technologies which while helpful, can also be put to nefarious use, and it is vital that safety by design is held paramount in their development and release to the public.

143 eSafety Commissioner, Addressing deepfake image-based abuse, 24 July 2024, Addressing deepfake image-based abuse | eSafety Commissioner

144 Review Submission 135 – Australian Human Rights Commission, 14

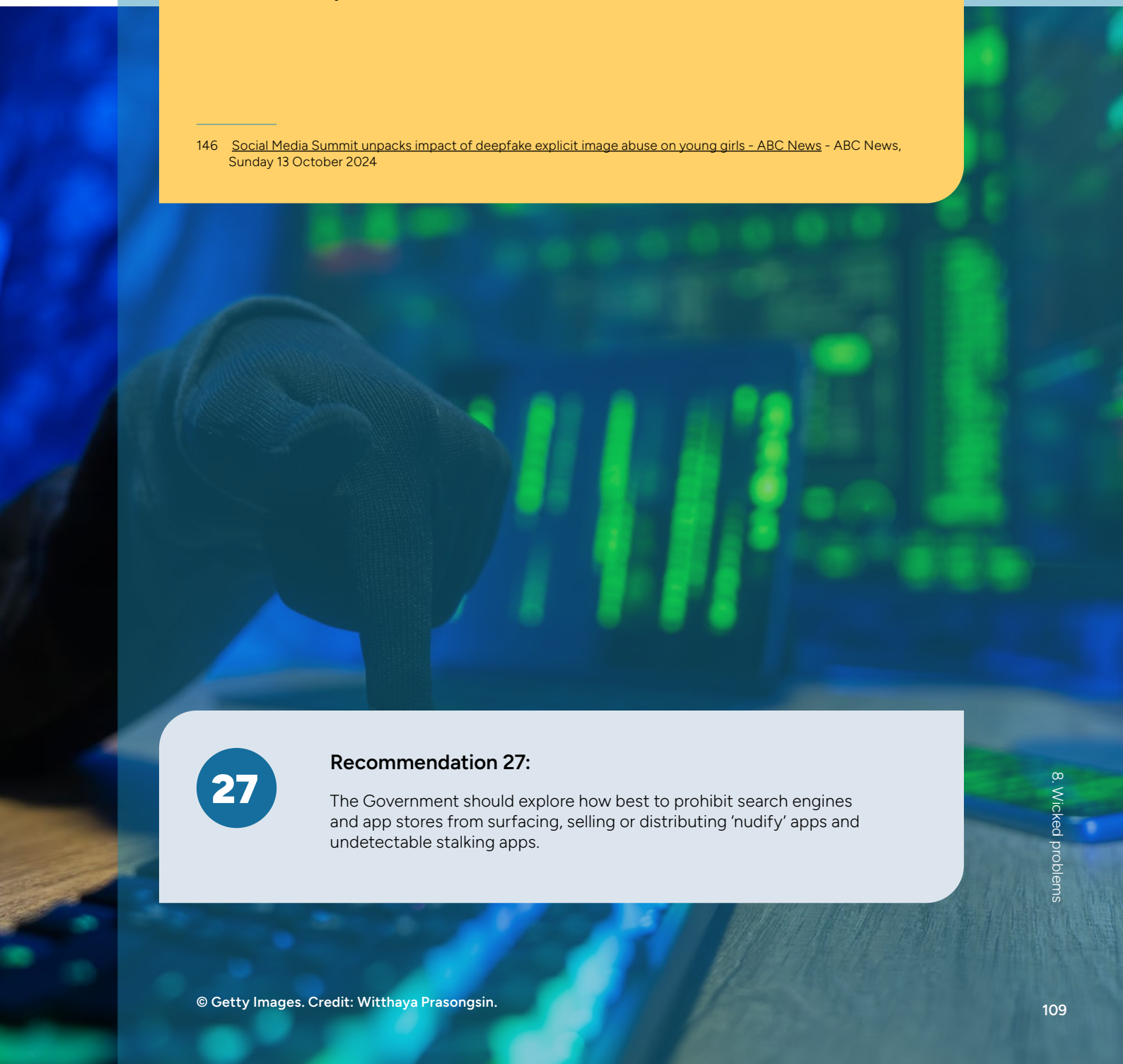
145 Anne Summers, How tech became the next frontier in domestic violence, The Saturday Paper, 16 March 2024, How tech became the next frontier in domestic violence | The Saturday Paper.

Case Study: Deepfakes

It was a typical day for Ms Mason at her work when she received the call from her daughter, Matilda "Tilly" Rosewarne, at lunchtime in November 2020. Hearing 14-year-old Tilly sobbing on the phone, Ms Mason said she knew "instantly something was very wrong". Her daughter explained that a fake nude image — which depicted Tilly's likeness — was circulating at school. Tilly was eight years old when she first started being bullied. Ms Mason said the bullying Tilly experienced continued into high school, occurring around their hometown of Bathurst in NSW. The bullying then morphed onto social media. The impact was swift … "We came to understand how far the image had been spread amongst students in Bathurst," she noted in her submission to the Joint Select Committee on Social Media and Australian Society. For months following, the cyberbullying worsened. Just after 3am on February 16, 2022, Tilly took her own life. She was just 15 years old. Ms Mason said: "Sadly, Tilly died from a thousand cuts that occurred over the course of her short life. This was a death from bullying — exacerbated by social media."

ABC News, Sunday 13 October 2024[146]

146 Social Media Summit unpacks impact of deepfake explicit image abuse on young girls - ABC News - ABC News, Sunday 13 October 2024



27

Recommendation 27:

The Government should explore how best to prohibit search engines and app stores from surfacing, selling or distributing 'nudify' apps and undetectable stalking apps.

8.2 End-to-end encryption

The complexity of balancing privacy and safety

End-to-end encryption was a particularly polarising issue among submitters and stakeholders I consulted for the review. It is increasingly being adopted by services which offer messaging functions to consumers and there are many who support this move as a means of protecting users' right to privacy. However, others especially those in the law enforcement and child protection areas, consider the increased adoption of this feature a huge set back in being able to detect and stop child sexual exploitation and abuse. I have great sympathy with this view.

By its very nature, end-to-end encryption renders message content unintelligible to everyone except the sender and receiver. It can therefore conceal harmful conduct or hinder investigation of the distribution of harmful and illegal online content such as child sexual exploitation and abuse material. In its submission the Australian Federal Police note that in 2022-2023, 96.1 per cent of content it lawfully intercepted was unintelligible due to encryption, and suggested that it also prevents communications service providers from identifying illegal content on their own platforms and reporting it to law enforcement.[147]

Proponents of end-to-end encryption argue that it supports a range of rights, in particular privacy, and in so doing benefits online safety. Submitters including Digital Rights Watch suggest that it guards against a range of other online harms "such as harmful targeted marketing and targeted extreme content and disinformation, data breaches and identity theft."[148] Further, "it protects the privacy of victims of domestic violence, confidential sources of journalists, safety of political dissidents and all activists, lawyers, and reporters."[149]

Privacy and Digital Rights Organisations in their joint submission advocated against service providers being required to:

Remove or circumvent … end-to-end encryption in order to meet a duty of care if one was introduced as it 'would pave the way for pervasive surveillance and damage online safety as well as the human rights to privacy and free expression."[150]

Meta in its submission asserted that:

An independent Human Rights Impact Assessment of Meta's expansion of end-to-end encryption - conducted by NGO Business for Social Responsibility in line with UN Guiding Principles on Business and Human Rights - found, among other areas, that encryption increased the realisation of privacy, freedom of expression, protection against cybercrime threats, physical safety, freedom of belief and religious practices and freedom from state-sponsored surveillance and espionage.[151]

There is a legitimate place for encryption

A number of submitters acknowledged the complexity of balancing privacy and safety posed by end-to-end encryption and called for a proportionate, considered and risk-based approach.

For example, the UTS Centre for Media Transition, while generally supportive of a duty of care proposal, recommended that "the difficult issue of access to encrypted content should be addressed separately from the proposed duty of care" and recommended "a prohibition on generalised monitoring of users, such as that included in the EU Digital Services Act."[152]

147 Review submission 141 - Australian Federal Police, 27.

148 Review submission 112 - Digital Rights Watch, 4.

149 Ibid, 15.

150 Review submission 109 - Privacy and Digital Rights Organisations joint submission, 1.

151 Review submission 166 – Meta, 46.

152 Review submission 134 - UTS Centre for Media Transition, 10.

The Office of the Australian Information Commissioner (OAIC), while not making specific recommendations regarding end-to-end encryption, recommended that,

In balancing privacy, security and safety, the Online Safety Act Review should consider whether any impact on privacy is reasonable, necessary, and proportionate to pursuing a legitimate objective.[153]

Even the Australian Federal Police, while being largely critical of the broad application of end-to-end encryption and its disruptive impact on the fight against child sexual exploitation and abuse, acknowledge that encryption has a legitimate role in protecting certain information, including banking and identity data.[154] The stakes are high on both sides of the argument.

I am interested to observe the pragmatic approach taken by eSafety in developing Phase 1 standards in relation to illegal material, including child sexual exploitation and abuse material, for Designated Internet Services and Relevant Electronic Services (whose industry codes were rejected due to their insufficient protections). The standards that have since been developed by eSafety, which are intended to commence in December 2024, work around end-to-end encryption and take an outcomes-based approach. They require services to:

implement appropriate systems, processes and technologies to detect and remove known child sexual abuse and pro-terror material where it is technically feasible and reasonably practicable to do so.[155]

Importantly, they include the qualification that,

Providers will not be required to implement systems or technology to detect and remove material where doing so would require the provider to implement or build a systemic weakness, or a systemic vulnerability, into the service or where it would require an end-to-end encrypted service to implement or build a new decryption capability or render methods of encryption used in the service less effective.[156]

Similar language is used in the *Online Safety (Basic Online Safety Expectations) Determination 2022*, which specifies services are not required to introduce 'systemic weaknesses' in their encryption to satisfy relevant expectations but are expected to take,

reasonable steps to develop and implement processes to detect and address material or activity on the service that is unlawful or harmful.[157]

A similar approach has been taken in development of online safety legislation overseas. The UK Online Safety Act gives Ofcom the power to require that a company use 'accredited technology', or "make best efforts to develop technology", to tackle child sexual exploitation and abuse on any part of its service including public and private channels.[158] However, it is yet to be determined, through consultation and code-making, what this technology will be.

153 Review submission 146 – Office of the Australian Information Commissioner, 9.

154 Review submission 141 - Australian Federal Police, 28.

155 Explanatory Statement to the Online Safety (Designated Internet Services—Class 1A and 1B Material) Industry Standard 2024, 3. Identical provision is in Explanatory Statement for the Online Safety (Relevant Electronic Services—Class 1A and 1B Material) Industry Standard 2024.

156 Ibid.

157 Online Safety (Basic Online Safety Expectations) Determination 2022, section 8(1), 8(2)(a).

158 Home Office (UK)(2023), Guidance: End-to-end encryption and child safety, End-to-end encryption and child safety - GOV.UK.

Other strategies must be employed for fighting child abuse online

Whatever one may think, it is clear that services have no appetite to wind back end-to-end encryption and that governments, weighing up the reduced ability to detect illegal content against other safety and privacy concerns, have chosen not to legislate against use of encryption. Separately, I understand that governments in many countries, including Australia, are working with industry to identify feasible technical solutions to lawful access which maintain privacy.[159] In relation to child sexual exploitation and abuse material, the approach has been to require services to employ other means to detect child sexual exploitation and abuse material.

Meta in its submission notes that they continue to invest in behavioural analysis and metadata as effective harm prevention rather than undermine encryption.[160] Meta has also said in media statements that it uses machine learning to detect patterns of behaviour, including posting child sexualised content, coded language in bios or joining certain groups, and is able to stop suspicious accounts before they can contact children or share content. The Attorney General's Department in its submission note the success of methods used by service providers to monitor engagement on their platforms to proactively identify and prevent child abuse. Successful prevention efforts have targeted preparatory grooming behaviours, such as those observed in chat logs between a child and perpetrator, and behavioural signals that are linked to online child sexual exploitation and abuse and are visible in encrypted spaces, including suspicious user activity (such as mass contacting of unknown or underage accounts), and the use of specific emojis and vernacular.[161] I believe that while services must be required to cooperate fully with law enforcement in conducting their investigations, adoption of proactive and preventative strategies should be done as part of everyday business by platforms themselves.

I consider that requirements to develop and deploy effective detection methods for child sexual exploitation and abuse material must be mandatory and absolute, rather than aspirational or a case of 'best attempts.' Services must also take every measure available, and continue to develop measures, to prevent the use of technology to make child abuse material, as well as detect its dissemination online. International Justice Mission, in its submission notes that end-to-end encryption is one of many features of major platforms that can be used to facilitate sexual abuse of children.[162] It advocates a greater emphasis on device manufacturers and operating systems in safety by design requirements, including requiring 'camera-enabled devices' to have safety features designed to prevent the capture and rendering of child sexual exploitation and abuse material.[163] During consultation for the review with the child protection sector, I heard that technology already exists that can scan images for age and nudity, but it has not been adopted by device or operating system developers.

While appropriate prevention and detection measures should be addressed through a statutory duty of care, given the seriousness of this issue additional focus may need to be given to ensure the appropriate technology is developed and that it continues to be fit for purpose.

Marshalling the experts through dedicated fusion cells

For most problems that need a solution, competition is frequently the driving force for change. However, some issues are so important and difficult that cross sectoral collaboration is needed. I consider that given the multifaceted nature of technology-facilitated abuse, including the combination of social and technological aspects that drive it, a concentrated and collaborative effort is required. It is clear that a single organisation is unlikely to find a perfect solution and where the harm is complex and causing significant harm, we should bring our best people together for a time-limited period to come up with solutions that can be shared across industry, government and civil society.

There has been some success in establishing 'fusion cells' – a multi-stakeholder approach where you bring the best minds together to solve an issue. Industries including telecommunications providers, financial services, and energy providers

159 Department of Home Affairs Five Country Ministerial 2023.

160 Review submission 166 – Meta, 46.

161 Review submission 132 – Attorney General's Department, 6.

162 Review submission 95 – International Justice Mission, 18.

163 Ibid, 14.

are developing strategies to prevent technology-facilitated harms from occurring on their services using fusion cells, and such an approach is already bearing fruit in the scams area, where the ACCC's National Anti-Scam Centre has set up an investment scam fusion cell.[164] It is also occurring in response to family violence issues. An existing example of a multi-stakeholder approach is the Australian Digital Platform Regulators Forum (established in 2022 and consisting of the ACCC, Australian Communications and Media Authorty [the ACMA], eSafety and OAIC) which has collectively considered key safety-related matters like examining the function of multimodal foundation models, and considering the harms and risks of algorithms. However, there is benefit in bringing a broader range of expertise and establishing a time limited, tailored group.

In the online safety context, developing a fusion cell might better support online services to address technology-facilitated gender-based violence on their service, as well as a range of other problems.

Fusion cells are particularly effective when they are time limited and can harness those with the greatest expertise and have particular access to intelligence and data to understand the core drivers of an issue and potential solutions. In the case of technology-facilitated abuse, this could involve, in addition to regulators and law enforcement, representatives of the technology sector, women's safety groups, and industry.

Noting the vital importance of collaboration to date in fighting child sexual exploitation and abuse, a dedicated fusion cell may also be a useful approach to the specific issue of detecting child sexual exploitation and abuse material on encrypted services. A fusion cell for this purpose would work across industry, academia, government, regulatory and child protection sectors.

28

Recommendation 28:

The Government and the regulator should both be able to convene multi-stakeholder 'fusion cells' to analyse 'wicked problems' (such as the implications of end-to-end encryption for combatting child sexual exploitation and abuse, and technology-facilitated abuse and gender-based violence) and develop coordinated multi-stakeholder solutions.

164 National Anti-Scam Centre's first fusion cell to disrupt investment scams | ACCC.

8.3 Sextortion

The crime of sextortion and its impacts

Sextortion involves the blackmailing of victims, often adolescent boys and young men, using sexualised images that the victims have been pressured and/or tricked into sharing. Data from the Australian Federal Police showed more than 90 per cent of victims were male and predominantly 15-17 years of age, however police had seen victims as young as 10 years old.[165]

eSafety notes there are other forms of sextortion, such as escalating demands for increasingly explicit sexual content, or for direct sexual engagement. Here the focus is on financial sextortion.

Scammers generally contact victims through social media posing as attractive women, they then often move the chat to another service and the chat becomes sexual. Victims are manipulated into sharing nudes or other sexual images of themselves and then scammers threaten to share the images with the victim's family or social network if their demands are not met. Demands may include money, gift cards or online gaming credits; sometimes repeated demands are made over time, as the victim is held to ransom by their fear of being found out and of their images being seen by everyone they know. The offenders will often continue to harass victims until there is no longer a viable avenue for communication – this can occur if the victim stops communicating with the offender including blocking accounts, user reporting of offender behaviour, law enforcement involvement or the service provider suspends communication due to suspected or reported behaviour that violates their terms of service.

The impacts of this sexualised and extremely violating form of scam can be devastating, as noted by the Australian Federal Police in their submission: "offenders exploit young victims" feelings that they have done something wrong and will be reprimanded by parents or carers and even prosecuted by the law if their actions are discovered.'[166] This can cause victims to react in panic, fear and shame to the extent that they do not seek help and do not think there is a way out – tragically, there have been several cases which have led to victims dying by suicide or attempting suicide.

Incidence has increased in recent years

It is very troubling that the incidence of sextortion seems to be increasing over time. The Australian Centre to Counter Child Exploitation recorded a 100-fold increase in reports of this form of sextortion from 2021 to 2022.[167] Similarly, eSafety found a 1,332 per cent increase in reports of sexual extortion- from 432 reports in 2018–19 to 6,187 reports in 2022–23.[168] eSafety noted that the dramatic increase could be due partly to increased awareness of its image-based abuse scheme, but also noted that due to stigma among victims sextortion was likely to be underreported.[169]

165 AFP and AUSTRAC target offshore sextortion syndicates preying on Australian youth | Australian Federal Police.

166 Review submission 141 – Australian Federal Police, 23.

167 Australian Federal Police (2022), AFP and AUSTRAC target offshore sextortion syndicates preying on Australian youth | Australian Federal Police..

168 eSafety Commissioner (2024), Lifting the veil on sextortion, Lifting the veil on sextortion | eSafety Commissioner.

169 Ibid.

Current and future responses to an evolving problem

In relation to eSafety's ability to respond to sextortion and support victims, reports can be made under the image-based abuse scheme for those over 18. eSafety directs minors to report to the Australian Centre to Counter Child Exploitation. However, a major challenge, not only for eSafety but for law enforcement, is that most perpetrators operate out of offshore crime organisations.

There have been some gains in the fight against this crime, for example law enforcement operations such as AFP-led Operation Huntsman shut down over 500 Australian accounts linked to offshore crime organisations linked to sextortion in 2022.[170]

However, the main response to date in stopping the spread of sextortion has been education. The Australian Federal Police in their submission note that its ThinkUKnow program has been delivered across Australia since 2009 educating students, parents, carers and teachers about online child sexual exploitation and how to keep children and young people safe online.[171] eSafety has developed targeted resources aimed at young men, with succinct advice on how to handle sextortion if an image has been shared, and tips on identifying sextortion to stay safe.

Clearly, while education will always be an important ingredient in the prevention of online harms, the onus must be on online services to make their services safe. It is my hope that a duty of care will compel services to develop technological solutions to prevent bad actors using their products for criminal activity.

Some online platforms are beginning to respond more comprehensively to the problem of sextortion, and time will tell how effective these measures are and whether other services follow suit. I am heartened to see that Snap, Meta and Apple have this year announced strategies to help address sextortion.

In June this year, Snap introduced features including an in-app warning to alert young users if they receive a chat from someone who has been blocked or reported by others, or from a region outside their current network. There are features that also prevent delivery of friend requests from users without mutual friends in known suspicious locations, and the ability to block unwanted requests from multiple accounts created on the same device.[172]

In October, Meta announced a range of new features on Instagram, including automatic blocking of follow requests to teenage users from suspicious accounts, a 'nudity protection feature' in Instagram direct messages, automatic prompts for child users that appear before sending detected images, and removing the ability to take screenshots or screen recordings of disappearing photos and videos sent through Instagram direct messages or Facebook Messenger. Instagram will also display an in-feed sextortion education message to users in Australia, the United States, Canada and the United Kingdom.[173] They have not, however, done this for all of their services.

Apple has announced enhancements to its safety features this October, launching first in Australia. Currently, an iPhone automatically detects images and videos containing nudity that children receive, or attempt to send, in iMessage, AirDrop, FaceTime and Photos. The child is shown two intervention screens before they can proceed, including prompts to contact a parent or guardian. Users will now have the option to report the images and videos to Apple.[174]

I note that some commentators have suggested that these safety features do not go far enough. A recently unredacted complaint against Snap from the New Mexico Attorney General noted that in 2022 Snap staff were fielding around 10,000 reports of sextortion each month on the platform[175] – illustrating the point that action by Snap – and across industry – on sextortion is overdue. There is skepticism around safety features that put the onus on users, especially children, to protect themselves,[176] and those such as nude image detectors that are activated after such an image has been produced or received by a child.[177] I am sure that more can be done across

170 Australian Federal Police (2022), AFP and AUSTRAC target offshore sextortion syndicates preying on Australian youth | Australian Federal Police.

171 Review submission 141 - Australian Federal Police, 7.

172 Snapchat focuses on user safety with new features to combat sextortion and bullying | Mi3 26 June 2024.

173 Meta unveils features to combat teen 'sextortion' - NBC News 17 Oct 2024, accessed 30 October 2024.

174 New iMessage feature allows children to report nudity to Apple | Apple | The Guardian - 24 October 2024.

175 Snapchat ignored sextortion, child grooming, New Mexico lawsuit alleges | Mashable 1 October 2024.

176 Meta unveils features to combat teen 'sextortion' - NBC News 17 Oct 2024, accessed 30 October 2024.

177 Instagram to block some screenshots to help prevent sextortion – BBC, 18 October 2024.

the industry and that responses to sextortion will need to continue to adapt as criminal enterprises rapidly adopt new technology and find ways to circumvent these safety features.

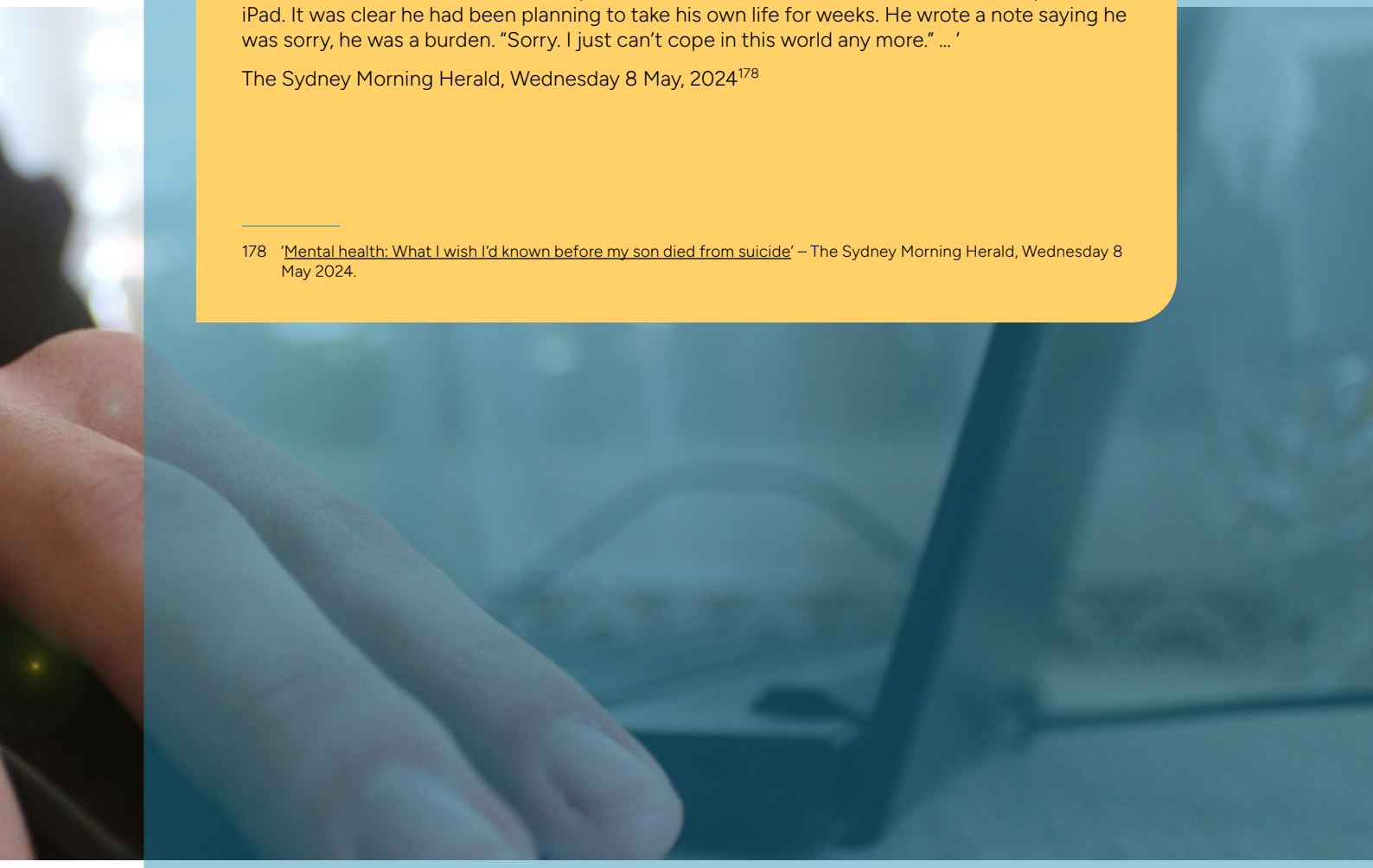
Given the importance of a cooperative approach in addressing it, transparency is also an essential tool in the fight against sextortion. The regulator and law enforcement bodies should also have access to data from services on reports of harms such as sextortion, what they are doing to address such harms and how effective these efforts are. If competition doesn't spur on other services it could be that a fusion cell on sextortion would also be useful.

© Getty Images. Credit: Goran Babic.

Case Study: Sextortion

'I wish I'd known that in the hours before someone decides to take their own life, they act as if they don't have a care in the world. I wish I'd known that. My son might still be alive. I didn't know anything about suicide before last October. I didn't know anything until my son took his own life after sextortion demands made on him. Yes, we found the criminal, but it was too late for my son to make a victim's impact statement. And I was not allowed to give a victim statement on his behalf. That person received a six-month jail sentence. He'd already been in jail for three months and now, he's out on the streets again. I can't get distracted by that, though … My son was a victim of sextortion. I knew about it. He'd come to me and said: "Hey Dad, I've made a mistake." We talked about what he could do. He'd paid them $500 and then another $500 but still the threats came. He made a statement to police and when the scammer called again, he'd put me on and I'd pretended to be a police officer to warn them off. But the thought of his friends seeing the images he'd sent really rocked him. Nothing I said seemed to help, but then he appeared to be OK. I discovered later that the scammers had tried again two or three weeks after their first attempt. A few weeks after he died, I looked at his computer and iPad. It was clear he had been planning to take his own life for weeks. He wrote a note saying he was sorry, he was a burden. "Sorry. I just can't cope in this world any more." … '

The Sydney Morning Herald, Wednesday 8 May, 2024[178]



178 'Mental health: What I wish I'd known before my son died from suicide' – The Sydney Morning Herald, Wednesday 8 May 2024.

LINKS TO THE NATIONAL CLASSIFICATION SCHEME

09

Part 9 of the Online Safety Act, the Online Content Scheme, relates to restricted and illegal content and references the National Classification Scheme (Classification Scheme) to determine whether certain online content is restricted or illegal. Class 1, put simply, is material that is or would be Refused Classification under the Classification Scheme, and Class 2 is material that is or would be restricted to adults.

One of the issues that both eSafety and industry raised with me in consultations and submissions was the need to decouple the regulation of Class 1 and Class 2 material from the Classification Scheme. I agree with them.

Using these borrowed thresholds, which entail applying a range of considerations under the Classification Scheme, is not fit for purpose as a framework for efficient decision making of dynamic and potentially high-volume online content and impedes rapid responses to illegal and harmful content when speed is important.

That said, if you are going to remove content, it is important that frameworks are adopted that enable transparent and consistent decisions by the regulator, are simple for industry to comply with and do not lead to gaps, allowing illegal and harmful content to slip through the cracks, or amount to unnecessary censorship.

Decoupling the Act from the Classification Scheme raises a number of issues, not the least of which is content that is currently considered to be 'classifiable content' under the Classification Scheme but which is accessed online. This includes sexually explicit content, commonly known as pornography, as well as a range of online content. This is partly due to definitional issues under the *Classification (Films, Publications and Computer Games) Act 1995* (Classifcation Act) which are being considered by the Government concurrently to this review as part of a program of reform to the Classification Scheme. However, it is appropriate for this review to consider, from the perspective of the online safety regulator, where responsibilities should lie.

9.1 The connection between the two frameworks is a legacy of older legislation

Linkages between the Act and the Classification Scheme are a legacy of previous online safety legislation under Schedules 5 and 7 of the *Broadcasting Services Act 1992*. The 'Online Content Scheme' (Schedules 5 and 7 of the Broadcasting Services Act) was enacted in two stages in 1999 and 2007. A core function of the Online Content Scheme, like its current iteration, was removal of illegal and potentially harmful content (in particular content harmful for children). Under Schedule 7, the then Australian Broadcasting Authority, later the Australian Communications and Media Authority (ACMA) had powers to address prohibited content on various Australian hosted online services. Content classified MA 15+ (mature accompanied), R 18+, X 18+ or Refused Classification (RC) was deemed prohibited for particular or all audiences and could be removed or addressed under Schedule 7. The Classification Board was required to classify the material before it could be removed.

The *Online Safety Act 2021* removed the need for the Classification Board to classify online content before it could be removed, instead retaining an option for the Commissioner to consult the Board or "stand in the shoes of the Board" in deciding whether to take something down or issue a notice. However, Part 9 retained references to the Classification Scheme in defining Class 1 and Class 2 material. Class 1 material is solely defined with reference to the Refused Classification category of the Classification Scheme, which includes child sexual exploitation and abuse material, pro-terror material and content inciting, promoting or instructing in crime and violence. The Refused Classification category, and therefore Class 1, also includes a range of content that is considered to fall outside community standards. Class 2 includes material that is or would be classified X 18+ (a classification category for depictions of actual sexual activity between consenting adults) and R 18+ (high impact material that is restricted to adults).

The Commissioner has removal powers in relation to Class 1 material even if it is hosted outside Australia. For Class 2 material, the Commissioner must establish that the material is provided from Australia before issuing a removal notice to the service provider, or a remedial direction requiring certain steps be taken to establish the age of end-users.

Part 9 also provides for enforceable industry codes and standards in relation to Class 1 and Class 2 material. Codes and Standards in relation to Class 1 material have been made and the Class 2 codes are currently being developed.

9.2 The classification framework is not suited to responding to illegal and harmful online content

The Classification Scheme was enacted in 1995, prior to the ascent of online services to provide for the classification of films, publications and computer games for commercial release in Australia (G, PG, M, MA 15+ or R 18+), for regulation of sexually explicit films and publications (X 18+ films and Category 1 and 2 Restricted for sexually explicit magazines) and for refusal of classification (banning) content that is illegal or would be considered unacceptable to a reasonable adult. It was established to provide for individual, professionally produced items, with predetermined and fixed content, to be assigned a classification to determine a suitable audience for commercial release. The Classification Scheme also provides a mechanism to refuse classification to (ban) material which instructs or incites in matters of crime or violence or offends against the standards of morality, decency and propriety generally accepted by reasonable adults.

The process of classification under the Classification Scheme is intricate. It involves applying the 'matters to be taken into account' specified in s11 of the Classification Act, the principles and categories in the National Classification Code and the relevant thresholds for categories as described in the separate Guidelines for the Classification of Films 2012, Guidelines for the Classification of Computer Games 2023 and Guidelines for the Classification of Publications 2005. Central to the Classification Scheme is balancing the right of adults to see, play and read what they want and the need to protect children from harmful and disturbing content.

This nuanced framework is not workable as the basis for a regulatory regime designed to apply to vast volumes of online content, that as eSafety suggest in their submission, is dynamic, fluid and even ephemeral.[179] What is needed is clear rules to determine whether certain material is illegal or harmful in order to trigger rapid removal, appropriate regulatory action and efficient compliance by online services.

eSafety also suggest that the underpinning concept for classification standards under the Classification Scheme is offence rather than harm.[180] This distinction is overplayed to an extent, however, it is true to say the focus of the Classification Scheme is far broader than harms, necessitating a more complex set of considerations, whereas harm is the sole focus of the Act, and as such a less complex decision-making process should apply.

There is broad support from industry for decoupling Class 1 and 2 from the Classification Scheme. Telstra, in its submission, notes that industry should not need to know how to classify content in order to comply with industry codes and instead should be subject to risk-based criteria.[181] DIGI suggested that harmful but lawful material should be subject to more streamlined and objective standards of harmfulness,[182] and Meta noted that AI/human hybrid moderation of user generated content required 'clear targets' rather than detailed assessment.[183] The International Social Games Association noted that Australia's application of media classification frameworks to online safety regulation is inconsistent with approaches in other countries.[184]

179 Review submission 73 – eSafety Commissioner, 5.

180 Ibid.

181 Review submission 168 – Telstra, 4.

182 Review submission 144 – DIGI, 5.

183 Review submission 166 – Meta, 19.

184 Review submission 140 – The International Social Games Association, 3.

9.3 Alternative framework for Class 1 and 2 material

There is general support from both the regulator and regulated entities under the Act that the current definitions for Class 1 and 2 material are not fit for purpose, and compelling reasons for adopting an alternative.

Any new framework must be transparent, objective and easily scalable. There must be strong confidence that the framework is acceptable to the community, particularly as it will result in the removal of material. While the framework borrowed from the Classification Scheme is unwieldly, it has been developed through a rigorous process, as provided for in the *Intergovernmental Agreement on Censorship 1995*, changes to the National Classification Code and guidelines require public consultation and agreement of all participating Ministers.[185]

There is no intergovernmental agreement needed in relation to the Act, however as noted by the Communications Alliance,[186] providers should not be required to restrict categories of material where a decision has not been taken by Parliament that the category of material should be restricted.

When it comes to removal or restriction of content, even while moving away from the exact criteria that applies under the Classification Scheme, there should be a high degree of consistency between the standards under both schemes.

29

Recommendation 29:

The Act should be decoupled from the National Classification Scheme with new Class 1 and Class 2 definitions and thresholds specified in the Act and, as far as possible, be based on equivalent standards in the National Classification Scheme.

185 Intergovernmental Agreement on Censorship 1995, Part 5.

186 Review submission 157 – Communications Alliance, 18.

9.4 A new way to categorise material

Having regard to the input I have received on categorisation approaches, I propose a significantly streamlined typology for Class 1 and Class 2 content. It should be clearer and simpler for application of removal and remedial powers, and for industry compliance under the relevant codes and standards. Tables 9.1 and 9.2 show the proposed frameworks for Class 1 and 2 alongside a summary of the current framework based on the Classification Scheme (in their entirety, existing classification criteria are several pages long).

The proposed framework is far more succinct than the current classification-based framework, and differs in that it lists the kinds of material that would sit in either category, or makes reference to definitions in relevant legislation, rather than describing qualities of material (e.g. gratuitous, high impact) as is done in classification instruments. These changes are designed to make this alternative framework significantly easier to follow for both the regulator and regulated entities.

As noted earlier, it is important for standards to align between schemes. A new categorisation of material could largely align with current thresholds under the Classification Scheme in terms of how they are applied to online material, and Class 1 could retain reference to Part 9A of the Classification Act dealing with terrorism content. However, certain differences to the current classification framework would be introduced.

The first difference is greater alignment with the Criminal Code. As Part 9 includes powers with respect to illegal material, the link to the Criminal Code should not be through references under the Classification Scheme, but instead be direct, and should be readily updated to reflect any relevant amendments to the Criminal Code. Recent amendments include criminalising the use of a carriage service to deal with violent extremist material under amendments which came into force in January 2024. In addition, amendments to strengthen offences targeting the creation and non-consensual sharing of sexually explicit material online, including 'deepfakes' received Royal Assent on 2 September.

Another proposed change is the transfer of fetish material from Class 1 to Class 2. This change aligns with a recommendation of the most recent review of the National Classification Scheme (the 2020 Stevens Review).[187] Stevens noted, as has eSafety, that non-injurious fetish practices are legal to do between consenting adults, so the blanket prohibition on representations of these practices does not seem justified. More recently, in developing industry codes under Part 9, eSafety has separated fetish material from other Class 1 content, as it was considered more appropriate to address this material alongside Class 2 pornography rather than grouping it with content such as child sex abuse and terrorism material.

The next proposed change is that the material that would be Class 2 in the alternative framework covers a narrower range of material than is currently the case, with a focus on content which is legal but harmful. In the proposed framework, Class 2 includes actual sex as well as high impact violence. It would also include material promoting harmful practices including disordered eating, self-harm and substance use. This differs from the Classification Scheme that covers *any* aspect of real *or depicted* content themes (violence, sex, nudity, drug use or coarse language) that are assessed to have a high impact, as this would lead to an R 18+ classification. This difference at first may appear quite substantial, however it speaks to the differences in the material that eSafety chiefly deals with online and the desire to focus on harms, rather than offence or generalised age-appropriateness.

A further difference is in the specification of harmful practice material under Class 2, which has an implied equivalence in the classification system, where such content may be considered to have high impact themes. This also speaks to the difference between traditional media, such as a film, and the potentially more impactful presentation of such content online, for example by an influencer. I understand that eSafety's proposed approach in developing Phase 2 Codes is both to ensure children are prevented from accessing such material, and asking industry to establish measures to enable adults to avoid this content if they wish. I should note that in the long term I would see disordered eating, self-harm and substance abuse as fitting better under a code dealing with mental and physical wellbeing as these are problems that also impact adults.

187 Neville Stephens AO (2020) Review of Australian Classification Regulation, Recommendation 9.4.

30

Recommendation 30:

New Class 1 definitions and thresholds should clearly focus on illegal and seriously harmful material and directly correspond to the Criminal Code where appropriate. Sexually explicit material that includes violent and seriously injurious practices, such as choking, should sit under Class 1.

31

Recommendation 31:

New Class 2 definitions and thresholds should include material that is legal but may be harmful, particularly for minors, and consensual sexually explicit material including non-injurious fetish material.

32

Recommendation 32:

Class 2 definitions and thresholds should also capture material dealing with harmful practices such as disordered eating, self-harm and substance use to address their heightened impact, especially on young people, in the context of social media. In the longer term, industry should be obliged to prevent dissemination of such content through a broader code dealing with mental and physical wellbeing under duty of care provisions.

Treatment of online pornography in the proposed standards

The treatment of online pornography is a highly topical and controversial matter. There is general agreement sexually explicit material should not be available to children, and the Government is currently progressing an age assurance trial to support restriction of online pornography to adults, which would align with the categorisation of pornography as Class 2 material.

There is also concern about the impact on adults of repeated exposure to violent and degrading sexual practices, such as choking. While pornography can have positive effects, it can have negative effects on understanding consent, expectations about sex, ideas about intimate relationships and gender stereotypes.[188] While the priority must be to prevent children's access, including through continuing to designate pornographic material Class 2, the normalisation of these practices and this content should also be addressed, by stronger means than age restriction. To my mind such practices should be classified as Class 1. Much of this is gendered violence and our society needs to do more to prevent such practices becoming normalised and generally send a message that gendered violence won't be tolerated.

The proposed categorisation includes sexual activity between consenting adults and any fetishes that are non-injurious. The intention of the reference to non-injurious fetishes is to clarify not only the practices that should *not* be subject to removal (but should be restricted to adults), but also to rule 'injurious' practices out of Class 2.

Attempts to sub-categorise pornography beyond this distinction would require eSafety to make exactly the kind of detailed assessment, similar to classification, and consideration of 'good taste' that introducing a new framework aims to prevent.

If new Class 1 and Class 2 thresholds are introduced in the Online Safety Act, the regulator should have the ability to make legislative instruments (which would be subject to Parliamentary scrutiny) to add material to both Class 1 and Class 2. Provision should also be made for the regulator to give further guidance if needed.

188 eSafety Commissioner (2023), Accidental, unsolicited and in your face.

Table 9.1: Comparison of current and proposed framework for Class 1 material

Class 1 – Current approach referencing the National Classification Scheme

s106: Material, including films, computer games and publications (or parts thereof), that is, or would be, classified as Refused Classification (RC) under the Classification Act.

Section 9A of the Classification Act

- Advocates the doing of a terrorist act.

National Classification Code

- Depicts, expresses (and, for publications only, describes), or otherwise deals with matters of sex, drug misuse or addiction, crime, cruelty, violence or revolting or abhorrent phenomena in such a way that they offend against the standards of morality, decency and propriety generally accepted by reasonable adults to the extent that they should not be classified
- Describes or depicts in a way that is likely to cause offence to a reasonable adult, a person who is, or appears to be, a child under 18 (whether the person is engaged in sexual activity or not)
- Promotes, incites or instructs in matters of crime or violence.

Guidelines for the Classification of Films

- Impact test (for themes of violence, sex, nudity, drug use, coarse language) exceeds the **R 18+** classification
- Films outside the bounds of **X 18+** because they contain: violence, sexual violence, sexualised violence or coercion, sexually assaultive language, consensual depictions which purposefully demean anyone involved in that activity for the enjoyment of viewers, fetishes, depictions of non-adult persons, including those aged 16 or 17, adult persons who look like they are under 18 years, persons 18 years of age or over portrayed as minors
- Other criteria including: Descriptions or depictions of child sexual abuse; gratuitous cruelty or real violence which are very detailed; depictions of bestiality; gratuitous depictions of incest or other fantasies which are offensive or abhorrent; instruction in the use of proscribed drugs or promotion of proscribed drug use.

Guidelines for the Classification of Computer Games

- Impact test (for themes, violence, sex, nudity, drug use, coarse language) exceeds the **R 18+** classification
- Other criteria including: excessively frequent, prolonged, detailed or repetitive violence; actual sexual violence; implied sexual violence that is visually depicted, interactive, not justified by context or related to incentives or rewards; depictions of actual sexual activity; depictions of simulated sexual activity that are explicit and realistic; drug use related to incentives and rewards; interactive illicit or proscribed drug use that is detailed and realistic.

Guidelines for the classification of publications

- Publications that exceed the bounds of Category 1 or 2 restricted because they contain (examples): descriptions and depictions of violence that are excessive or are in a sexual context; descriptions and depictions of fetishes in which non-consent or physical harm are apparent; sexualised nudity or sexual activity involving minors; exploitative descriptions or depictions of: violence in a sexual context; sexual activity accompanied by fetishes or practices which are revolting or abhorrent; incest fantasies or other fantasies which are offensive or revolting or abhorrent.

Table 9.1: Comparison of current and proposed framework for Class 1 material *cont.*

Class 1 – Proposed approach
Material and activity that: ○ Promotes, incites, instructs in, or depicts abhorrent violent conduct (within the meaning of s 474.32 of the Criminal Code) ○ Depicts or describes[189] child abuse ('child abuse material' being within the meaning of section 473.1 of the Criminal Code) ○ Depicts child sexual exploitation ('child sexual exploitation material' being adapted from a definition contained within the Online Safety (Designated Internet Services—Class 1A and Class 1B Material) Industry Standard 2024 and Online Safety (Relevant Electronic Services—Class 1A and Class 1B Material) Industry Standard 2024) ○ Depicts bestiality ○ Depicts real sexual violence including where AI generated material is indistinguishable from actual footage of an assault ○ Depicts injurious sexual practices such as choking. ○ Advocates the doing of a terrorist act (within the meaning of section 9A of the Classification Act) ○ Promotes, incites, or instructs in serious crime or violence ○ Counsels or incites, or promotes or instructs in particular methods of suicide (per ss 474.29A and 474.29B of the Criminal Code) ○ Is non-consensually shared sexually explicit material including that which is AI generated per ss474.17A and 474.17AA of the Criminal Code) ○ Is violent extremist material as defined in s474.45A of the Criminal Code.

189 Includes material such as child abuse fantasies engaged in with chat bots.

Table 9.2: Comparison of current and proposed framework for Class 2 material

Class 2 – Current approach referencing the National Classification Scheme
S107: Material, including films and computer games (or parts thereof), that are, or would be, classified R 18+ or X 18+ under the Classification Act. Publications or parts thereof that are or would be classified Category 1 restricted or Category 2 restricted under the Classification Act. National Classification Code **X 18+ Films** - (except RC films) that: contain real depictions of actual sexual activity between consenting adults in which there is no violence, sexual violence, sexualised violence, coercion, sexually assaultive language, or fetishes or depictions which purposefully demean anyone involved in that activity for the enjoyment of viewers, in a way that is likely to cause offence to a reasonable adult; and are unsuitable for a minor to see. **R 18+ Films** - (except RC films and X 18+ films) that are unsuitable for a minor to see. **R 18+ computer games** - (except RC computer games) that are unsuitable for viewing or playing by a minor. **Category 1 restricted publications** (except RC publications and Category 2 restricted publications) that: explicitly depict nudity, or describe or impliedly depict sexual or sexually related activity between consenting adults, in a way that is likely to cause offence to a reasonable adult; or describe or express in detail violence or sexual activity between consenting adults in a way that is likely to cause offence to a reasonable adult; or are unsuitable for a minor to see or read. **Category 2 restricted publications** (except RC publications) that: explicitly depict sexual or related activity between consenting adults in a way that is likely to cause offence to a reasonable adult; or depict, describe or express revolting or abhorrent phenomena in a way that is likely to cause offence to a reasonable adult and are unsuitable for a minor to see or read. **Additional criteria** in the Guidelines for the Classification of Films, Guidelines for the Classification of Computer Games and Guidelines for the Classification of Publications, mainly relating to the allowance of high impact (but not very high impact) sex, violence, nudity, drug use, coarse language and themes, and that certain content is not permitted (see RC in previous table). From September 2024 games which simulate gambling are to be classified R 18+.

Table 9.2: Comparison of current and proposed framework for Class 2 material *cont.*

Class 2 – proposed approach
Online material that: ○ Is not a computer game or publication subject to the National Classification Scheme ○ Depicts real sexual activity between consenting adults ○ Depicts realistically simulated sexual activity between consenting adults[190] ○ Depicts high impact adult nudity ○ Depicts non-injurious fetish practices in real sexual activity between consenting adults ○ Promotes serious ill health through eating disorder material ○ Promotes serious ill health through self-harm material ○ Promotes serious ill health, including dependency and addiction, through alcohol and drug material ○ Depicts real violence (other than Class 1 violence) with a high impact in a way that is gratuitous or excessive ○ Depicts high impact fictional violence in a way that is gratuitous or excessive.

190 Includes AI generated material and sexually explicit material accessed through immersive technology.

9.5 The regulatory overlap between the Act and the Classification Scheme needs to be resolved

Aside from the borrowing of standards developed under the Classification Scheme for the operation of the Online Content Scheme, there is also an issue of regulatory overlap between eSafety and the Classification Board. This is due partly to the broad definitions of 'film' and 'publication' in the older Classification Act, which covers a range of online content,[191] and to the migration of traditional classifiable content, especially films and computer games, to online platforms. The Stevens Review noted the significant overlap between the schemes and that the practical implications are likely to gain significance over time.[192] Again, I acknowledge the important work being done to reform the Classification Scheme and the need for reforms to both frameworks to be coordinated.

Illegal material needs to be captured under both schemes and a common threshold is needed

As noted previously, it is important that standards under both schemes remain as consistent as is practicable. One of the advantages of the simplified framework for Class 1 and Class 2 regulation is that it will in fact make consistency in practice easier to achieve. In particular, both schemes must have a consistent and community endorsed approach to removing content, to avoid unnecessary censorship. One of the advantages of the proposed simplified framework for Class 1 and Class 2 material is that it will in fact make consistency in practice easier to achieve.

© Getty Images. Credit: Miguel Sotomayor.

191 *Classification (Publications, Films and Computer Games) Act 1995*, section 5.

192 Neville Stephens AO (2020) Review of Australian Classification Regulation, 29.

eSafety should be responsible for harmful online material, but not commercial films, games or publications that are classified

The Stevens Review proposed that there should be greater clarity around content requiring classification, and in particular that user generated content should not need to be classified under the Classification Scheme.[193] The stage 2 classification reform consultation paper included a proposal that content to be covered by the Classification Scheme should be limited to films, computer games and publications meeting the following criteria:

- Professionally produced – content with higher quality production values
- Distributed on a commercial basis – to capture organisations or individuals that distribute media content as part of their business, as opposed to individuals or community groups whose main purpose is not to distribute media content for commercial gain and
- Directed at an Australian audience – a selection of content is specifically made available for Australia or marketing is specifically directed at Australians.
- Classification is the responsibility of the service provider who makes the content available in Australia, regardless of who originally makes the content; and
- Online content is only classifiable where it is uploaded by the service provider itself to clarify that user-generated content that is professionally produced and distributed on a commercial basis does not require classification.[194]

These criteria would capture online streaming providers and online games stores directed at Australian consumers, but would not capture user-generated material that has been posted online which has historically been captured due to the broad definition of 'film' in the Classification Act.

Conversely, a number of submitters to this review, representing online games and streaming video on demand services- which consist largely of professionally produced content- have also expressed concern about being subject to overlapping regulatory regimes because they are online. These services suggest that compliance with the Classification Scheme should be sufficient and that the Online Safety Act, in particular the restricted access systems provisions, should not apply to them.[195] Submitters such as the Tech Council support clarifying distinctions between the types of content that are captured by each scheme, with the Classification Scheme covering services that primarily distribute professionally produced content and the Act covering intermediary services that contain user-generated content.[196]

However, the distinction between user generated and professional content is increasingly blurred, with the emergence of influencer channels and services like OnlyFans that have high production values and significant audiences. Therefore, commerciality alone would not be enough to rule material out of the remit of the Act. One advantage of introducing a statutory duty of care and a risk-based tiering of obligations is that services such as OnlyFans would be required to assess the risks inherent to their service and make suitable mitigations. These may be over and above requirements under codes for Class 1 and Class 2 material.

I would support a risk-based approach to delineating responsibilities between the two regulatory frameworks, which would allow eSafety to focus primarily on content, contact and conduct online which is most likely to present serious harms to users and others. At the very least, it should exclude online films (such as those on subscription video on demand), computer games and any e-books which would be subject to the Classification Scheme. The exception to this would be any social media enabled interactivity, for example the inclusion of user-generated content in a game, or chat functions tied to online gaming.

The other potential exception to this proposed delineation of responsibilities is pornography, as outlined below.

193 Ibid, 41

194 Public Consultation Paper: Modernising Australia's Classification (infrastructure.gov.au), April 2024 9.

195 Review Submission 117 – Australia New Zealand Screen Association, 4; Review Submission 140 – International Social Games Association, 3; Review Submission 142 – Interactive Games and Entertainment Association, 13.

196 Review Submission 147 – Tech Council, 5.

Clarifying responsibility for responding to pornography

Sexually explicit content, commonly known as pornography, is currently regulated under both the Classification Scheme and the Act, depending on whether it is distributed online or via traditional media.

As sexually explicit content online has become more prevalent, the number of X 18+ films submitted for classification has steeply declined. Since 2016, not a single X 18+ commercial film has been submitted for classification.[197] The growth of online pornography has largely involved sites which are hosted overseas, leaving limited options for enforcing compliance with classification requirements. The growth of online pornography has also seen a blurring of boundaries between professionally produced content and user generated content.

In proposing a decoupling of standards from those used in the Classification Scheme, I have suggested slightly different standards in relation to pornographic material than those that currently apply under the Classification Scheme, with greater emphasis on harms.

There are growing concerns about potential links between the rise of pornography use, due to its proliferation and easy access by children online, and the development of misogynistic attitudes or otherwise unhealthy understandings or behaviours with regard to sexuality and relationships. This has prompted the Government to, among other things, bring forward work on online age verification. Preventing children's access and empowering other users to control their exposure to pornography is a key aspect of Phase 2 industry codes development. The Government has also stated that reforms to the Classification Scheme could have a role in addressing harmful online pornography.[198]

These factors highlight a need to clarify regulatory responsibility for online pornographic content. I would support giving eSafety the lead in determining appropriate standards for this content when it is delivered online in Australia, which would be achieved through decoupling Class 1 and 2 thresholds from the Classification Scheme. I also consider that it should be made clear that eSafety has regulatory responsibility for online pornography. Making eSafety the sole regulator for online pornographic content would also be consistent with the approach in the UK's Online Safety Act.

If the standards for online pornography proposed earlier as part of decoupling from the Classification Scheme were adopted, it could open up a process for review of thresholds for all pornographic content, noting that consistency is desirable. However, subject to development of stage 2 classification reform, the Government may wish for the Classification Scheme to retain responsibility for both pornographic films released in cinemas/offline formats and for sexually explicit publications and to look separately at these standards.

Age verification and age assurance

Online age verification and age assurance is a relatively under-developed area of regulation, with Australia and other jurisdictions yet to legislate that services must employ specific enabling technology.

The eSafety Commissioner has the power to declare that an access-control system is a restricted access system for the purposes of the Act. The current Online Safety (Restricted Access Systems) Declaration 2021 seeks to ensure that the methods for limiting access to relevant Class 2 material meet a minimum standard. In certain circumstances, the Commissioner can investigate access to Class 2 material by end-users in Australia, investigate any such material hosted in Australia or determine whether material needs to be placed under a Restricted Access System. Draft Phase 2 industry codes, which are currently being consulted on and which will be enforceable once registered, also contain some age assurance measures for Class 2 material.

While the age assurance trial has been framed in relation to preventing children's access to pornography and in the context of implementing an age limit on social media, the potential for application of such requirements to other content online, including as an adjunct to classification for online films and games that are age restricted, is obvious.

This poses the question as to whether eSafety should be responsible for enforcing age verification or age assurance on *all* online platforms, including for online films, computer games and publications that are subject to the Classification Scheme. For traditional films, games and relevant publications regulated under the Classification Scheme, state and territory

197 Classification Board and Classification Review Board Annual Report 2015-16 through 2022-23.

198 Tackling online harms | Our ministers – Attorney-General's portfolio, Media Release, The Hon Mark Dreyfus KC MP, 1 May 2024.

legislation enforces age restrictions that apply as a result of classification (e.g. shops must not sell R 18+ games to minors). For online films, games and publications, state and territory classification legislation does not provide for enforcement of age restrictions.

On balance I think that eSafety should continue to be able to determine the adequacy of age verification technology employed by social media platforms, pornography sites and any online services where age verification or age assurance may be warranted, other than those currently covered under other legislation. For example, I note this is currently one of the considerations of the ongoing Phase 2 codes process. However, this should not preclude other agencies retaining responsibility for ensuring that age assurance laws are complied with. This would mean, for example, that while eSafety was responsible for enforcing requirements around implementing effective age verification as it relates to Class 2 material, responsibility for enforcing age verification for online films, computer games and publications under the Classification Scheme would be a matter to be resolved by the Government in consultation with states and territories.

33

Recommendation 33:

In reforming the Act and the National Classification Scheme, the regulatory remit of eSafety should be clarified. Content that is subject to the National Classification Scheme should fall outside eSafety's remit (except features that are uniquely social media enabled).

PENALTIES AND ENFORCEMENT

10

The Act seeks to address content, conduct and contact online which too often results in significant harm to Australians and the broader community, and yet penalties are relatively mild. Stronger maximum penalties are needed to create a persuasive deterrent, especially for those online services which are among the richest global corporations in the world. Should new obligations be placed on services under a duty of care, appropriate and persuasive penalties must be in place. Coupled with stronger penalties, there needs to be a range of enforcement options available to the regulator including those with a remedial focus. A major challenge to enforcement in Australia also lies in questions of jurisdiction and the extra-territorial enforceability of penalties, and this must be overcome for regulation to be effective against all international services.

10.1 New penalty and enforcement options are needed to enforce the duty of care

Effective compliance with an overarching duty of care and its corresponding obligations requires new enforcement mechanisms and penalties. There are a number of systemic obligations under the duty of care which could potentially trigger enforcement action, including obligations to:

- Conduct risk assessments, mitigation, measurement and reporting
- Produce regular transparency reports
- Conduct audits when required to by the regulator
- Comply with lawful requests for information
- Cooperate with investigations
- Comply with codes made by the regulator
- Submit to obligations when they have been duly designated (e.g. 'reach or risk' designations); or
- Otherwise honestly provide all information necessary, and perform all actions necessary, for the regulator to be able to determine services' obligations and monitor their compliance.

In the first instance, failure to comply with these or other necessary obligations will generally require action by the service to remedy the non-compliance. The regulator may first seek to secure this through informal effort to bring a service into compliance voluntarily. However, where this fails – or where it is inappropriate or likely to fail – the regulator needs powers to enforce compliance. These powers should at least include:

- Enforceable undertakings, where a service enters into a legally binding agreement to undertake remedial actions agreed to by the service and the regulator; and
- Remedial directions, where the regulator may independently direct a service to undertake actions deemed necessary to bring it into compliance.

But when remedial action fails, enforcement of the duty will require enforcement through:

- Infringement notices
- Court injunctions; and/or
- Civil penalty orders made by the Court.

The maximum civil penalty that a Court can impose will require sufficient range to capture the full potential severity of failures by services to address systemic risks. These failures are potentially of a different order of magnitude to failures to address specific harmful content or activity under the removal schemes, as they may increase the risk of there being a greater prevalence or persistence of harms on a service more generally.

The Act currently sets 500 penalty units as a maximum penalty for most of its offences. At current levels, this amounts to a maximum penalty of $782,500 for companies, as at 31 October 2024.[199] While this penalty amount may be compounded on daily basis where required actions are not taken, it is still very low when considering:

- The size of major platforms, and the resources available to them
- Maximum penalties available to regulators under comparable laws internationally; and
- Maximum penalties which can be imposed on companies under other Australian laws.

Technology companies like Apple, Microsoft, Amazon, Alphabet (Google) and Meta all have market capitalisations in excess of $1 trillion USD – and over $3 trillion in Apple and Microsoft's case[200]. By way of comparison, Australia's reported GDP in April 2024 was $1.7 trillion USD[201]. It is easy to see that for these and many other companies regulated by the Act, a penalty of $782,500 would barely be considered a "parking ticket"[202] – and potentially even far less of a deterrent than one.

199 Due to the passage and Royal Assent of the Crimes and Other Legislation Amendment (Omnibus No. 1) Act 2024, the amount of the Commonwealth penalty unit will change from $313 to $330. This will take effect from 7 November 2024.

200 Companies Market Cap (2024) Largest Companies by Market Cap, https://companiesmarketcap.com/

201 IMF World Economic Outlook Database, https://www.imf.org/en/Publications/WEO/weo-database/2024/April/

202 Review submission 64 – Uniting Church in Australia, Synod of Victoria and Tasmania and Synod of Queensland, 16.

As noted by the issues paper, and by many submitters, more recently enacted laws internationally (see Table 10.1) have set maximum penalties at a much higher level and as the higher of either a percentage of global annual turnover or a high pecuniary amount in the millions.

Table 10.1: International comparison of online safety maximum financial penalties

	Australia Online Safety Act 2021	Ireland Online Safety and Media Regulation Act 2022	EU Digital Services Act	UK Online Safety Act 2023
Maximum Fines / Penalties	**$156,500** for individuals; or **$782,500** for corporations	**€20 million;** or **10 per cent** of annual turnover in the prior financial year	**6 per cent** annual worldwide turnover in the preceding financial year; or for periodic penalties, **5 per cent** of the average daily worldwide turnover	**£18 million;** or **10 per cent** of annual global turnover

Maximum penalties should be strong *in proportion* to the size of a service in order to be credible and effective. This is why the approach taken in the United Kingdom and Europe, and proposed in Canada, where maximum penalties for online services are primarily framed as a percentage of that service's *global* annual turnover, is one that should be employed in the Act in relation to duty of care breaches. Furthermore, a variation of this approach is already taken in Australia where the *Competition and Consumer Act 2010* and the Privacy Act have penalty regimes that consider turnover. This is also under consideration in the Communications Legislation Amendment (Combatting Misinformation and Disinformation Bill) 2024, where the maximum penalty is the greater of 5 per cent of global turnover or 25,000 penalty units ($7.825 million).

34

Recommendation 34:

The maximum civil penalty that a court can impose should be increased to the greater of 5 per cent of global annual turnover or $50 million.

10.2 Stronger civil penalties for complaint schemes

In addition to a new maximum penalty to enforce a duty of care, and for sustained and systemic code breaches, change is needed to penalties linked to the Act's removal schemes. While the maximum civil penalty for individuals (500 penalty units, or $156,500) should be a serious enough disincentive in most cases, a maximum of $782,500 under the corporate multiplier is not necessarily significant for companies, even though the penalty will accrue on a daily basis for every day the content is not taken down. The potential for harm when services are resistant or unresponsive in removing abusive or illegal material from their service is serious enough to warrant far higher penalties.

These penalties should not be the same as those for non-compliance with systemic requirements, as the failure to comply in relation to individual pieces of material is a different order of offence to a systemic failure. The duty of care model recognises that services have greater responsibility for the design of their services and their systems and processes, than for particular pieces of content. However, such failures, especially repeat failures, may be indicative of systemic deficiencies and could be considered as factors for such by the regulator.

Some submitters have recommended that the penalties for different removal schemes should be differentiated, to account for relative harmfulness, legality or severity of the material (e.g. child sexual exploitation material vs adult cyber abuse). It is true that this is currently an issue given that the maximum penalty is low, but applying a higher maximum to all schemes will address the problem by allowing the Courts to decide on the severity of an offence. While many see child sexual exploitation and abuse material as inherently worse than the other harms covered in the schemes, there needs to be flexibility for courts to decide their relative severity in particular cases – e.g. where failure to remove bullying or cyber abuse material contributed to a person's self-harm or suicide.

35

Recommendation 35:

The civil penalties for non-compliance with removal notices should be increased to a maximum of $10 million for companies.

10.3 The regulator and courts should be able to more broadly order and enforce remedial actions by services

In cases of non-compliance, it is often good regulatory practice to first seek remedial action from the regulated entity, to bring them into compliance without immediate resort to penalties. This is particularly good practice where the non-compliance is less serious or could have been made out of ignorance or a misunderstanding. Working first to seek remedial action reduces the potential burden on regulators, regulated entities, and the courts – and focuses on fixing harms and mitigating risks, which are ideally the objectives of regulation. Penalties are just one tool for achieving these objectives, and not necessarily the best one to employ initially.

It is also likely that there will be circumstances of non-compliance where the payment of a pecuniary penalty is insufficient for addressing the underlying causes of the non-compliance. Offences arising from the systems, processes, policies or practices of a service should lead to corresponding changes in these factors, to remedy these underlying causes and prevent future offences (as noted above).

Currently, remedial direction powers only exist for the image-based abuse scheme and (in certain circumstances) the Online Content Scheme. The Act should therefore be amended to provide the regulator with powers to issue remedial directions or enforceable undertakings for all instances of non-compliance where they may be useful, including in relation to:

- Compliance with the duty of care (see above)
- Complaints schemes; and
- For both end-users and online service providers.

36

Recommendation 36:

The Act should be amended to empower the regulator to use enforceable undertakings or issue remedial directions to services in relation to all relevant penalty provisions, to seek to bring them back into compliance.

10.4 Enforcement action related to content reporting and removal should be streamlined and consistent across schemes

Removing illegal, harmful or abusive material is only one side of reducing the harm of such material. Where initial efforts to have material removed may take time or be unsuccessful, the regulator should have the option to immediately order search engines to limit the discoverability of this material.

Link-deletion notices can currently be issued to search engines services by eSafety to require them to 'de-index' Class 1 material (rendering the material unsearchable on that service), but only where efforts to have the material removed by the provider of the service on which it is posted, or the host of that service, have failed, and the material could be accessed by Australian end-users at least twice during the previous 12 months. This means that over that period, in which eSafety has been aware of the existence of illegal content, such content may still be discoverable to users on search engine services. The Act should be amended to allow for the issuing of link-deletion notices immediately, to immediately minimise the discoverability of such material while it is still available.

Immediate link-deletion powers should also be extended to all of the content removal schemes, not just the Online Content Scheme. This would rapidly limit the discoverability of abusive material relating to a person which might appear in searches. For example, if image-based abuse material of a person is available online it may appear in search results relating to that person's name. The regulator should have the power to immediately address this.

37

Recommendation 37:

The Act should allow removal and link-deletion notices to be issued simultaneously under the Online Content Scheme.

38

Recommendation 38:

The Act should empower the regulator to simultaneously issue link removal notices for all harmful content under removal schemes.

10.5 A pattern of repeated non-compliance with takedown notices should be treated as a breach of the duty of care

Where a service demonstrates multiple or repeated disregard for enforcement actions such as removal notices, they are demonstrating that the standard penalties are not a sufficient deterrent and that a more severe penalty ought to be applicable. The regulator should be able to identify and penalise demonstrably and habitually bad actors with heightened enforcement action. This should also include consideration of such actors' repeat non-compliance as evidence of a systemic failure in meeting the duty of care.

39

Recommendation 39:

The finalised duty of care model should include scope to consider repeated non-compliance by services in removing content as evidence of non-compliance with the duty of care.

© Getty Images. Credit: Boonchai Wedmakawand.

10.6 Additional powers should be considered to hold individuals accountable

The regulator should be empowered to pursue all possible avenues for requiring the removal of harmful content and the deterrence of harmful conduct under its schemes. This should include the end-user of a service responsible for the content or conduct. End-user notices, requiring the person who posted material to remove it, should be available for all removal schemes, including (where appropriate)[203] for the Online Content Scheme. The availability of end-user notices should not shift the onus away from service providers in exercising responsibility for safety on their service. In most cases, the service will be in the best position to respond to removal requests and notices, as overseas end-users may be unresponsive or identifying information may be difficult to obtain. However, there may be circumstances where the end-user is best placed to remove content, such as where they have posted content to multiple services or where a service is being unresponsive to removal action.

The Act lacks escalated powers to deal with repeated abusive conduct by end-users, and to prevent such conduct from continuing. It is common for people engaging in online abuse to be habitual abusers, or 'trolls'.

Proposed amendments to address repeat abuse include:

- Removing the need to issue separate notices for every instance of conduct by an end-user
- Powers to issue remedial directions to end-users under all schemes, to require actions such as removing all instances of abusive material they have posted against a person on services, and to refrain from abuse in future
- In serious cases, powers to require the suspension or removal of an end-user's account or account privileges. It may not be feasible to fully prevent determined or resourceful end-users from creating new accounts, however services should take reasonable steps to prevent it, and at any rate it would create a disruption or friction which would act as a deterrent for many users. Such a power would also need to be balanced against consideration of the effects of removing the account (e.g. what range of services a user requires the account in order to access); and
- In very serious cases, power to request Court orders restraining end-users from engaging in certain online activity.

40

Recommendation 40:

The Act should include consistent powers to require end-users to remove content and refrain from posting abuse in the future.

203 The Online Content Scheme deals with the worst of the worst kinds of online material, and an end-user notice would not always be appropriate. For example, in the case of child sexual exploitation and abuse material, issuing a notice to an end-user could also jeopardise law enforcement investigations.

10.7 Australia should work to align and cooperate with international partners on enforcement

Australia's enforcement of online safety laws will be most effective if it is 'interoperable' and coordinated with like action by our international partners, such as the UK and EU. The Act should provide for and enable maximal cooperation with international partners through mechanisms such as information-sharing, cross-border enforcement and mutual assistance agreements.

10.8 Business disruption and access restriction powers should be considered for severe or repeated violations

My observation from my time at the Australian Competition and Consumer Commission (ACCC) is that in most cases overseas entities will submit to the jurisdiction of Australian Courts when a regulator brings an action, even when they don't have a presence here and they will pay any penalties or orders imposed. That said, when they have no presence in Australia they could choose to not submit to our courts or ignore any rulings, including in regard to penalties.

We are not alone with this problem given that the major services are located only in a very small number of countries and primarily in the United States of America.

In cases of severe or consistent non-compliance, the Courts should be able to take steps beyond financial penalties. Especially important in enforcing compliance against overseas services are measures which leverage onshore services which provide access to or do business with these services. These powers can be categorised as:

- Access restriction powers – blocking or limiting access to a service from Australia; or
- Business disruption powers – disrupting or hampering a service's ability to do business or receive revenue in Australia.

The Act currently only provides for internet service provider (ISP) blocking powers in relation to material depicting abhorrent violent conduct. While the Act provides for court orders to services to cease providing a service in the case of serious non-compliance under the Online Content Scheme, there is no provision for when a service refuses to comply. In this case it may be simpler to be able to order ISPs to block access. Workarounds to such orders may exist, but they will at least have the effect of seriously disrupting a service's operation.

Business disruption powers are exercised indirectly against a service in order to disrupt its ability to conduct business or make revenue. For example, the UK's Online Safety Act provides for such orders to be made to an "ancillary service", which is a service that "facilitates the provision of a regulated service" or "displays or promotes content relating to the regulated service"."[204]

Such powers might include court orders that such ancillary services withdraw assistance or service to non-compliant companies, for example by:

- Not processing payments to a service
- Not advertising on a service or providing advertising services to a service

204 UK Online Safety Act 2023, section 144(11).

- Not displaying a service in search results; or
- Otherwise contracting or conducting business with a service.

Such powers would need to be carefully drafted, proportionate to the harm, and subject to strict due process. Consideration would need to be given to how the exercise of such powers may affect the interests, rights and obligations of blameless third parties (including small businesses who rely on the service to attract customers) and should not affect contracts and arrangements entered into prior to their commencement. Further advice and consideration, including regarding the constitutionality of such provisions will also be necessary.

Consideration could also be given to business disruption powers which do not directly affect third parties or completely restrict access. One potential example may be a requirement for ISPs to 'throttle' (slow down) access to a particular service. Facebook whistleblower Frances Haugen noted in my meeting with her that Meta's research found users disengaged when Instagram feeds were slowed. Externally imposing such a slowdown could be used as a business disruption measure. While not blocking the service entirely, this would disincentivise use of the service, affecting its business and potentially nudging it back to compliance.

41

Recommendation 41:

The Government should expand access restriction powers against services for seriously harmful non-compliance.

42

Recommendation 42:

The Government should consider options for business disruption powers for seriously harmful non-compliance.

10.9 Australia should explore options for requiring a domestic presence for major platforms

While most major overseas services have been cooperative with eSafety in the administration of the Act, for example in responding to removal requests/notices and Basic Online Safety Expectations, recent experience has demonstrated an increased willingness to challenge Australian jurisdiction over online activity or material affecting Australians. Even simple administrative or legal actions such as the service of notices are potentially complicated by services' lack of onshore presence, designated contacts or complex company structures.

To address this issue, the Government should investigate the feasibility of requiring major online services to establish a domestic legal presence in Australia as a condition of operating in the country. Such a requirement would better enable the regulator to exercise and enforce its powers, and otherwise work with major services in achieving their compliance with Australian law.

In the meantime, major services should be required to designate and report a point of contact to the regulator for the purposes of complying with the Act. This is currently an expectation in the Basic Online Safety Expectations, but should be made a requirement for services designated under the reach and risk thresholds for duty of care obligations.

43

Recommendation 43:

The Government should consider the feasibility of requiring major platforms to have a local presence for the purpose of facilitating enforcement action.

44

Recommendation 43:

The Act should require major platforms, that is those designated under the reach or risk criteria under the duty of care requirements, to have a contact point for service in Australia.

10.10 Consideration should be given to introducing a licensing scheme

Licensing is a common requirement for many activities or forms of enterprise where it is considered necessary to behave or operate to an enforceable standard for the safety and wellbeing of others and for upholding the law. Licensing is required in many areas of everyday life, such as driving, but also for many sectors of industry. For example:

- Premises and entertainment venues which serve alcohol must be licensed
- Companies that operate infrastructure and carry communications in Australia must be licensed
- Persons who engage in construction and related trades such as electrical work must be licensed in order to operate
- Commercial Television and Radio broadcasting services must be licensed; and
- Banking and financial services must also be licensed.

In addition to supporting enforcement of the Act, a licensing scheme for online platforms could support:

- The news media bargaining code
- Requiring services to be members of the recommended digital services Ombuds scheme; and
- Cost-recovery from industry for the costs of maintaining the online safety regulatory regime. Licence holders could be required to pay a fee for the licence, or holders earning above a certain revenue threshold could be required to pay an industry levy.

Introducing a licensing scheme for online platforms would not be without challenges though. First, in seeking to overcome the problem of extra-territoriality, licensing may only provide leverage where a service has significant enough interest in maintaining its availability to Australians. While it is likely that most services would comply, a compliance regime cannot simply depend upon voluntary compliance. This problem is not unique to the licensing of online services, but extends to all obligations placed on services that are domiciled overseas, but unfortunately the online world is without borders and therefore much easier to circumvent. However, this is not a reason not to do anything.

A licensing regime would also likely raise complex legal issues, including determining the criteria for designating who falls under the regime and other legal risks. That said, I think there are good reasons to continue to explore licensing. Promising work is currently being undertaken by Dr Rob Nicholls of the University of Sydney and others are also looking at this issue. Licensing regimes are central to how we regulate sectors that carry high risks in Australia, and it seems incongruous that one of our highest risk sectors is not covered.

45

Recommendation 45:

The Government should consider options for introducing a licensing scheme for major services as a condition for operation.

INVESTIGATIONS AND INFORMATION GATHERING POWERS

11

The powers for investigation and information gathering in the Act should be broadened and strengthened, to underpin the new duty of care regime and to better support the investigative work eSafety is already tasked with. These powers include a broader ability to initiate investigations, and to collect the information and evidence required to monitor compliance, investigate potential non-compliance, and take enforcement action. Investigators should be empowered to use technology and methods that are the most effective and appropriate for online investigations concerning online harms. In addition to ***information gathering***, the Act should also provide a broader scope for ***sharing*** information with relevant agencies and stakeholders domestically and internationally – recognising that the effort to tackle online harms crosses portfolio boundaries, jurisdictional boundaries, and often requires cooperation of non-governmental organisations.

11.1 New investigative, information gathering and monitoring powers would support increased powers and scope of the Act

Proactive investigative powers

The majority of the Commissioner's current investigative work under the Act is focused on complaints made under the Act's various removal schemes, although investigations powers also apply to industry codes and standards. With the exception of investigations relating to Class 1 and 2 material, this investigative work is necessarily reactive, with reports made to eSafety often serving as the trigger for investigations (although broader triggers may apply to codes and standards). Therefore, outside of the Online Content Scheme, eSafety is currently constrained in their ability to proactively conduct investigations on their own initiative. With the adoption of a duty of care with proactive and systemic obligations however, the Act's will need to allow for more general and robust investigations powers. In particular, it is necessary to provide the regulator with the power to conduct investigations on its own initiative in relation to:

- Suspected non-compliance by online services with their duty of care, and more specifically with related obligations such as to conduct proper risk assessments, produce accurate transparency reports, comply with codes and submit to audits where the regulator requires them to be undertaken; and
- The subsequent posting, reposting or spread of material previously reported and removed under the Act's removal schemes. For example, where reported image-based abuse of a person has continued to be posted online the targeted person should not be required to re-report new instances of the same material – a requirement which has the potential to retraumatise the person. This capability can be assisted through technological means, with the appropriate authority (see below).

Proactive, own-initiative investigations powers for regulators are not uncommon. Section 21 of the *Interactive Gambling Act 2001*, for example, explicitly gives the Australian Communications and Media Authority (ACMA) the authority to investigate matters "on its own initiative". Internationally, the EU Digital Services Act gives Digital Services Coordinators and the European Commission powers to initiate investigations and proceedings against online services, and the UK Online Safety Act provides powers to Ofcom, the regulator, to initiate investigations into compliance.

Monitoring and investigative powers

To support its investigations authority under an expanded Act, the regulator will require sufficient powers to conduct investigations, monitor compliance, and to inspect, audit and validate information provided by the service. While the Act currently provides specific information gathering and investigative powers, these were designed to support an investigations function primarily intended to support specific content reporting and removal schemes rather than more broad and systemic compliance obligations. Indeed, eSafety report that these powers are not always sufficient for the purposes of enforcing the Act in its current form. For example, they report that when service providers, hosts or end-users deny having the ability to remove the material or do not respond to removal notices, they have insufficient power to gather evidence to prove violation of a civil penalty provision.

Therefore, an essential step in building up an investigative capability to underpin the new powers and service obligations in the Act will be to incorporate the broad monitoring and investigations provisions, such as those available in Parts 2 and 3 of the *Regulatory Powers (Standard Provisions) Act 2014* (Regulatory Powers Act). These provisions empower regulators to:

- Monitor compliance with legislative requirements, and whether information given in compliance with legislative requirements is correct; and
- Gather evidence relating to the contravention of offences or civil penalty provisions.

While the eSafety Commissioner has some of these powers in Part 14 of the Act, especially section 199, they are specific to certain sections of the Act. A broadening could include such powers as:

- With consent or under a warrant, accessing, copying or seizing documents, records and electronic systems. They would enable investigators to access communications, digital logs, content, emails and records in gathering evidence of a suspected non-compliance or breaches. This information could be used to monitor platform compliance with the Act or to identify end-users relevant to the investigation of harmful material; and
- With consent or under a warrant, gathering evidence from all relevant sources. Part 13 of the Act is restricted to gathering end-user information from social media, relevant electronic and designated internet services. However, in many cases these services do not hold sufficient information about the identity of end-users, which can hamper the progress of investigations. Applying the monitoring provisions would enable the regulator to obtain the needed information from any entity that may hold it, such as financial institutions, government agencies or individuals.

By adopting these provisions, the Act would gain a standardised and existing set of monitoring, investigation and enforcement procedures that are already widely used by Commonwealth regulators. This would reduce legal complexity, make enforcement action timelier and more efficient, and allow the regulator to act promptly to obtain the necessary information required to enforce and monitor compliance with the Act, remove harmful content, and identify end-users.

These powers are similar to those provided in the EU Digital Services Act and UK Online Safety Act.

46

Recommendation 46:

The Act should be amended to empower the regulator with stronger powers in relation to investigations, including to:

- Incorporate the monitoring and investigations provisions of the Regulatory Powers Act into the Act
- Initiate investigations of a services' compliance with the duty of care; and
- Initiate investigations into reposted material that was previously reported and taken down.

Investigations concerning online and technologically mediated harms, are out of necessity themselves mediated technologically. The technologies and processes, which enable the harms under investigation, also have the potential to limit or enable the conduct of those investigations. It is essential that regulators responsible for operating on this digital terrain are appropriately enabled and equipped to effectively do so using tools and practices adapted to that terrain. The Act should be amended to provide the regulator with the ability to employ the most appropriate and effective methods and tools for conducting investigations and gathering intelligence. Consideration could be given to a provision similar to that at section 22 of the *Interactive Gambling Act 2001*, which empowers the ACMA to conduct investigations as it "thinks fit".

Technological tools

With the appropriate authority, the regulator could be empowered to use technological and methodological tools needed to make their investigations and enforcement more effective and versatile. For example, sophisticated tools are available which could enable the regulator to automatically identify and flag harmful material that has been reposted after having been reported and removed. This would be particularly helpful in regard to image-based abuse and child sexual exploitation and abuse material. The Act should be amended to establish and clarify the regulator's authority to use tools such as these, especially where their status might otherwise be uncertain under the law.

'Sock puppet' accounts

Investigating compliance and gathering regulatory intelligence on online services often necessitates that investigators actively use those services, making use of accounts created on those services. In many cases, investigators and officers will only be able to effectively do this if they are capable of doing so pseudonymously – without identifying themselves as regulatory officials. The use of fake, anonymous or pseudonymous 'sock puppet' accounts that are used to observe or interact with online users can be an important tool to monitor social media platforms, such as in identifying harmful content, tackling online abuse or detecting non-compliant behaviour. Such accounts can be also be used to evaluate harms arising from the use of recommender systems and algorithms by social media platforms. These algorithms are designed to treat users differently according to their activity and characteristics. Therefore sock-puppet accounts can be used to test content recommendations, analyse the potential exposure to harmful or illegal material, and to monitor how algorithms interact with user behaviour. They can also be used to monitor changes with recommender systems specifically with the promotion or suppression of harmful content, as well as to test the accuracy of information provided under transparency powers.

The use of sock-puppet accounts is a necessary tool for investigations and intelligence gathering, and the use of them should be clarified in the Act. The Act will also need to clarify the regulator's authority to use sock-puppet accounts under relevant Commonwealth laws. More generally, consideration could be given to explicitly authorising investigators to breach a service's terms of use where necessary for the exercise of their official duties.

47

Recommendation 47:

Amend the Act to provide the regulator with appropriate flexibility to conduct investigations as it thinks fit, including the use of technological tools to assist with investigations and content removal, and the use of sock-puppet accounts.

11.2 Existing information and evidence gathering powers should be strengthened and supported

Obtaining end-user information (basic subscriber information)

Section 194 (s194) in Part 13 of the Act empowers the Commissioner to require from online services contact details or information about the identity of particular end-users if the Commissioner believes on reasonable grounds that the online service has information about the identity or contact details of the end-user and that information is relevant to the operation of the Act. This power may be necessary, for example, in issuing end-user removal notices for child cyberbullying, image-based abuse or adult cyber abuse material. In these or other circumstances where an investigation or the exercise of regulatory powers requires, this section empowers the Commissioner to unmask the anonymity of users.

Most services, however, do not collect verified information about the identity of their users. Contact information collected by services might also be unreliable, as accounts for most services can easily be set up using 'burner' email accounts. The regulator's ability to collect this information could be strengthened by such measures as:

- Expanding the scope and definition of the type of end-user information services may be required to provide, such as to include as much necessary and relevant data on the user gathered by the service as they can provide
- Expanding the scope of persons the regulator can require identifying information from, beyond providers of social media, relevant electronic and designated internet services. As noted above, this could be achieved through the incorporation of general monitoring and investigations powers; and
- Requiring services to preserve accounts and account information for the purposes of an investigation.

Another issue with section 194 is the lack of a confidentiality requirement to prevent services from informing end-users when they have received a s194 notice in relation to them. If services are able to tip off end-users whose activity is under investigation, complainants may be put at risk of further harm. The regulator should be able to require information to be kept confidential in appropriate circumstances – both information given to a service for the purpose of a s194 notice, and information given by a service in response. This is also an issue with section 199 discussed below, and confidentiality requirements should apply to both sections.

Obtaining identifying information of an end-user from services could be achieved by placing requirements on services to obtain reliable identifying information as a condition for opening an account, such as a valid phone number. This is already something services can do, such as through requirements for multi-factor authorisation. The Act should enable eSafety to obtain end-user information under Part 13, including a requirement that prevents services from informing end-users when they have received a notice under Part 13, a requirement for services to collect a user's phone number as a condition for opening an account, and provide a new power to compel the preservation of accounts for investigative purposes.

It must be noted that improving the regulator's ability to obtain end-user information under Part 13 raises complex privacy and security risks. Reforms must be carefully designed to ensure that limitations on privacy are aimed at achieving a legitimate objective and are reasonable, necessary and proportionate.

Evidence gathering

Evidence gathering is necessary for the Commissioner to exercise its investigations. Section 199 of the Act provides eSafety with the authority to issue a written notice to a person to produce information and/or documents and to answer questions for the purposes of an investigation.

Section 199 does not specify a time period for a written notice to provide documents or other information. The Act should continue to refrain from specifying a time period as the scope and urgency of information requests will vary. eSafety should be provided with flexibility to manage time periods to consider various factors, including the time it might take to generate information and the urgency of the information requests.

Section 205 of the Act provides authority to issue penalties where an individual has not complied with a request to provide evidence or produce documents under Part 14 of the Act. However, section 205 does not provide authority to issue penalties where an individual has not complied with a request to provide other information under Part 14. The Act should be amended to allow for penalties to be issued where a person is required to produce information under Part 14.

eSafety reports that services are sometimes reluctant to provide detail concerning the actions they have or have not taken in response to their requests and actions. The Act should be amended to allow the regulator to require such information. This should include information on action taken in response to both formal and informal requests – including informal requests that a service take action to enforce their own terms of use.

48

Recommendation 48:

Provide additional powers to the regulator to improve its ability to obtain end-user information under Part 13, including a requirement that prevents services from informing end-users when they have received a notice under Part 13, a requirement for services to collect a user's phone number as a condition for opening an account, and provide a new power to compel the preservation of accounts for investigative purposes.

49

Recommendation 49:

The Act should be amended to empower the regulator with stronger information gathering powers, including to:

- Improve its ability to obtain end-user information under Part 13 of the Act; and
- Set the time period for a written notice to provide evidence under Part 14 of the Act.

50

Recommendation 50:

Section 205 of the Act should be amended to confirm that non-compliance with a requirement to give evidence includes information as requested under section 199 (and other sections in Part 14 of the Act).

51

Recommendation 51:

The Act should be amended to require services to inform the regulator of all actions the service has taken in response to the regulator's actions and requests (including informal requests).

11.3 Services should be required to retain certain records relevant to investigations under the Act

To assist with the Commissioner's investigations, services should be required to maintain records of certain documentation, such as complaints relevant to schemes, measures taken to comply with obligations under the Act and risk assessments including details on how the risk assessment was informed. These records should be retained for five years.

52

Recommendation 52:

The Act should be amended to require services to maintain certain records, such as measures taken to comply with obligations under the Act and any actions taken in response to eSafety requests and risk assessments, for the purposes of the regulator's investigations.

11.4 The Act should provide for broader disclosure of information to relevant persons and agencies

Key to improving investigations of online services and online harms, both domestically and internationally, is the ability to share information with relevant agencies or persons, where necessary and relevant to the conduct of investigations and mitigation of harms. Part 15 of the Act currently authorises disclosures of information to:

- The Minister
- The Secretary or APS employees for the purpose of advising the Minister
- The ACMA
- Royal Commissions
- National Children's Commissioner
- Office of the Australian Information Commissioner (OAIC)
- The Australian Federal Police
- The Director of Public Prosecutions
- State and Territory law enforcement authorities
- Foreign regulators responsible for regulating social media, relevant electronic and designated internet services
- Foreign law enforcement authorities in relation to online safety on social media, relevant electronic and designated internet services; and
- Teachers, school principals, parents or guardians in relation to the resolution of child cyberbullying complaints.

While this list is quite extensive, it does not cover all the agencies or bodies it might be necessary or desirable to disclose information to. eSafety reports that the current disclosure provisions do not include all the organisations that they would want to share information with in order to better perform their functions. Expanding the scope and powers of the Act will only reinforce this issue and increase the need for broader disclosure powers.

The regulator should have authority to disclose to a broader range of Commonwealth agencies and departments. For example, the Australian Competition and Consumer Commission (ACCC) should be included – to cover all members of the Digital Platform Regulators Forum – as well as Home Affairs, the Attorney-General's Department and the Australian Security Intelligence Organisation. However, considering the pervasive influence and reach of online services across all domains of policy and regulation, it might not be advisable to specify in advance a limited range of agencies or departments that might have an interest or stake in information held by the regulator. Rather than a closed list, section 212 of the Act could be amended to permit disclosure to any head of a Commonwealth agency or department.

In addition to foreign regulators and law enforcement authorities, the Act should be amended to authorise disclosure to international authorities and appropriate non-governmental organisations (NGOs). Expanded authority in this area is essential for the regulator's functions, particularly in ensuring the safety of children and young people and responding to online crisis events. Relevant NGOs include INHOPE (Association of Internet Hotline Providers), C3P (Canadian Centre for Child Protection), and the Internet Watch Foundation, organisations that perform an important function in combatting child sexual exploitation and abuse online internationally. This expanded authority should of course come with appropriate limitations on disclosure and criteria for determining whether an NGO is appropriate for disclosure, including whether the receiving body has satisfactory arrangements in place for protecting the information or documents.

Sections 213 and 214 permit disclosure of information to teachers, school principals, parents and guardians relating to child cyberbullying complaints, to assist in their resolution. This provision should be expanded to include image-based abuse, which is unfortunately a form of online harm that occurs among children and in schools.

The Government should also consider clarifying and making explicit procedural fairness requirements relating disclosures.

© Getty Images. Credit: Davidf.

53

Recommendation 53:

The Act should be amended to allow the regulator to disclose information to:

- Any head of a Commonwealth agency or department
- International authorities; and
- Teachers, school principals, parents or guardians regarding complaints from a child about image-based abuse (as can be done for child cyberbullying).

54

Recommendation 54:

Allow the regulator to disclose certain information to Non-Government Organisations who have an approved role in assisting the regulator with enforcement activities.

PROMOTION, EDUCATION AND RESEARCH

12

A core function of eSafety, which has been in place from the very start, is that of promotion and education. Teaching the community about online safety, supporting others to deliver online safety education and promoting the supports provided by eSafety to those who are experiencing online harms is crucial to the effectiveness of the online safety regulatory framework.

Under section 27 of the Act, the Commissioner's functions include to:

- Promote online safety for Australians
- Support and encourage the implementation of measures to improve online safety for Australians
- Collect, analyse, interpret and disseminate information relating to online safety for Australians
- Support, encourage, conduct, accredit and evaluate educational, promotional and community awareness programs that are relevant to online safety for Australians
- Make grants of financial assistance in relation to online safety for Australians; and
- Support, encourage, conduct and evaluate research about online safety for Australians.

In practice these functions have translated into four key areas of activity: awareness raising about eSafety, education and capacity building to prevent online harms, strategic partnerships, and research and evaluation.

I have set out to understand the breadth of awareness raising, education and capacity building, research and evaluation and partnership activities undertaken by eSafety, and have given particular focus to what the available data, including research and feedback from submissions and stakeholder engagement, says about awareness of eSafety and the services it provides, as this is crucial to eSafety's effectiveness and has been raised repeatedly during consultations for this review.

12.1 Awareness raising about eSafety and help seeking

As we have seen, eSafety's powers relating to ordering content removal when triggered by a complaint from individual community members is highly valued by the community. However, a fundamental barrier to making a complaint is not knowing how to make a complaint or where to go. To ensure the effectiveness of complaints schemes, community members must know where to get help and be encouraged to seek help, therefore it has been necessary for eSafety to place strong emphasis on self-promotion.

Awareness raising activity for eSafety has been multifaceted, and included media appearances and features about the Commissioner, a Government funded advertising campaign conducted by the Department, and a range of targeted marketing and communications activity focussing on particular groups. In addition to general awareness raising, much of the educational material and activity by eSafety includes information on help seeking, including through eSafety itself. The main source of information about both eSafety and online safety in general is eSafety's main website. I will discuss awareness raising about eSafety and broader online safety education material separately.

Broad based activity to raise awareness of eSafety

The Government has invested $4.5 million over five years for a broad online safety awareness campaign, focussing on directing victims of online harms to eSafety for help. The first overarching awareness raising campaign for eSafety was done in 2023. eSafety's primary campaign activity on building awareness included the 'Your eSafety Kit', and participation in Safer Internet Day. Other activity has included a national awareness campaign aimed at parents and carers. eSafety was also allocated funding under the National Online Safety Awareness Campaign to invest in a youth-based SCROLL campaign on Instagram and has targeted messaging to young people on social media channels with Government funding of $100,000 per year for five years from 2022–23 to 2026–27.

It is worth noting here that eSafety considers that changing its name to the Online Safety Commission regulator would enhance its reach, impact and brand recognition. This is based on online safety being the more commonly used term to describe the domain in which they work. They believe that this change would be beneficial for their recognition and discoverability through search functions. Later in the report I have a recommendation that the name is changed to the Online Safety Commission. Should this change happen it is likely to coincide with the other reforms to the Act and it would be appropriate to have a new awareness raising campaign.

Targeted promotional activity

eSafety has also sought to raise awareness of its reporting schemes and education and prevention tools through a range of targeted activity.

For example, eSafety has directed its promotion activities to those most at risk of harm and to particular programs (like BeConnected, for older Australians), particular audiences (like LGBTQIA+ persons through a 'Learning Lounge' resource package) or particular apps.

More recently, eSafety has made moves to boost its profile in regional communities by attending and conducting awareness raising (and educational) activities in regional centres. eSafety staff distribute hardcopy materials and showbags, and answer questions at stalls at community events. Typically, regional visits also involve engaging with relevant community leaders and services, including schools, councils, police and local community organisations.

eSafety also conducts outreach in metropolitan areas, with a focus on young people, including physical stalls at festivals, posters in bars, universities, community groups, youth services and sporting clubs. Similar activity is also carried out at migrant resource centres, where resources in multiple languages are made available.

eSafety has also worked with the education sector, attending education conferences to encourage teachers, principals and education bodies to refer cyberbullying and online harms to eSafety. As well, eSafety also has targeted engagement with peak parent organisations attached to the education sector. eSafety's National Online Safety Education Council has direct engagement with 27 education authorities responsible for most Australian schools (around 10,000 schools) and eSafety has 986 eSafety Champions in schools across Australia. Direct in-person engagement with the school sector also occurs through eSafety's Trusted eSafety Provider initiative which is discussed further below.

Some of the targeted outreach by eSafety has intentionally focused on hardcopy resources, that direct people to eSafety's broader digital offering.

There is clearly a comprehensive range of broad-based and targeted promotional activity taking place across a number of channels, both online and offline, which is appropriate and necessary.

Awareness of eSafety and the range of resources it has developed is understood to be increasing. In its submission, eSafety noted " … we see increasing numbers of people aware of eSafety, reporting to eSafety, and increased uptake of educational resources and training."[205]

Research findings about awareness are mixed. A survey commissioned by the Department in 2022, found low (2.1 per cent) unprompted awareness among parents of eSafety as a source of help with negative online experiences, but fair prompted awareness (45.14 per cent).[206]

However, results from eSafety's 2019 and 2022 Australian Adults Online survey shows that prompted awareness of eSafety among adults rose from 15 per cent to 37 per cent in this period.[207]

A 2023 Government survey on online safety issues[208] found that 38.44 per cent of children surveyed had not heard of the eSafety Commissioner, and 24.72 per cent knew the name but were not sure what eSafety does. On the other hand, 26.22 per cent knew a little about the eSafety Commissioner, and 10.26 per cent knew 'a lot'. Those who had previously experienced online harm were more likely to know 'a lot' about eSafety.

There is clear demand for more awareness raising about eSafety and the supports it provides. Other unpublished eSafety data from 2022 found that 35 per cent of adults did not know where to go to report a negative online incident.[209] Another unpublished report indicated 75 per cent of teenagers aged 13–17 years were interested in seeing more content on social media about youth online safety issues.[210]

eSafety's annual report indicates significantly increased engagement with their websites, particularly their primary site, in 2023–24. They found a "strong upward trend in unique visitors" to esafety.gov.au, largely due to the continued implementation of a Search Engine Optimisation strategy. The number of users driven to the site more than doubled on the previous reporting period. In addition, 'direct' traffic increased by 80 per cent. eSafety credits a range of activity for the latter increase, including participation in Safer Internet Day, an inhouse brand awareness initiative and social media activations, email campaigns and advertising and promotion around the Act.[211]

205 Review submission 73 – Office of the eSafety Commissioner, 1-2.

206 Department of Infrastructure, Transport, Regional Development, Communications and the Arts (2022) The 2022 National Online Safety Survey – summary report, July 2022 https://www.infrastructure.gov.au/sites/default/files/documents/national-online-safety-survey-2022-wcag-accessible-report-25july2022-final.pdf.

207 eSafety (2022), Australian Adults Online, unpublished.

208 Social Research Centre (2023), 2023 Online Safety Issues Survey - Summary Report, 57.

209 Unpublished data from the Office of the eSafety Commissioner, 2022.

210 Omnipoll (Unpublished). Evaluation of the 2022 SCROLL / eSafety digital campaign among young people aged 13 to 17: a baseline report. eSafety Commissioner.

211 Annual report 2023–24 Australian Communications and Media Authority and eSafety Commissioner, 193.

© Getty Images. Credit: Boobi Lockyer/Refinery29 Australia - We are many image gallery.

Despite the impressive trends outlined in the Annual Report, demand for greater awareness raising about eSafety surfaced during this review, as told to me by representatives of affected communities and front-line support organisations, and through submissions. A common theme was that eSafety's online resources were highly regarded, but those who needed them most may not be aware of them. As noted by Good Things Foundation in its submission:

There are fantastic resources available through the eSafety website, but relying on online resources and access to webinars is not enough, ensuring that there is local support … is the key.[212]

A number of submissions also called for targeted resources and activities that are in fact already occurring, suggesting a need for more such activity, but also that there is a lack of awareness, even from key stakeholders in the online safety environment, about what is already available. For example, Dolly's Dream in its submission advocated greater awareness raising about eSafety's reporting schemes targeting rural, remote and regional Australian families.[213] Other submissions advocated conducting targeted outreach and awareness campaigns to educate vulnerable communities about their rights and the Act's protective measures[214] and working towards:

…A greater public awareness of government complaints mechanisms including particularly the eSafety Commissioner's platform which, according to some of our nascent ethnographic work, is virtually unknown among those adult users who are most vulnerable or most in need of it …[215]

212 Review submission 116 - Good Things Foundation, 2.

213 Review submission 49 - Dolly's Dream, 1.

214 Review submission 57 - Black Ink Legal, 8.

215 Review submission 106 - RMIT Digital Ethnography Research Centre, 7.

Despite work already done to target First Nations people (which I will discuss below), I heard repeatedly that there are ongoing gaps in awareness of eSafety and that more needs to be done to reach First Nations people, who experience an appalling degree of online harm. In its submission, the First Nations Digital Inclusion Advisory Group emphasised the need for culturally appropriate awareness raising about rights to safety online and how to seek help:

We strongly recommend that targeted communication with First Nations communities, using First Nations media and broadcasters. Without an effective communications campaign using First Nations media and broadcasters, there is a significant risk that this message and information will not be effectively received by communities.[216]

When I consulted with First Nations groups I was told more work needed to be done to raise awareness of eSafety and online safety more broadly. It was suggested that this could include distributing a 'digital rights' card, increasing in-person outreach in communities, providing resources to advocacy services to better deal with racial abuse, and supplying good quality and durable merchandise that has the contact details for eSafety.

Similarly, when I met with organisations to discuss technology facilitated abuse, it was suggested that many clients of family violence services are not aware of eSafety or what they can do to help, and that even young people experiencing technology facilitated abuse, despite being highly technologically literate, do not know about where to seek help.

Given the regular coverage of online safety issues in the press, there is an opportunity to use these stories to promote eSafety as a source of help, similar to the inclusion of information about mental health services at the end of news articles relating to suicide and mental health. I am aware that eSafety is quite proactive in the media, but adding these details could go a step further in prompting those who need it to seek eSafety's help and/or resources.

216 Review submission 169 - First Nations Digital Inclusion Advisory Group, 2.

55

Recommendation 55:

The regulator's continued awareness raising activities should include in-person outreach, including in hard to reach communities, and hard copy resources.

56

Recommendation 56:

Educational and promotional material should not only focus on what the regulator does for people experiencing harms, but also include simple messaging about how to make a complaint. Online safety education delivered at schools should focus on awareness of the regulator as a source of help. News media outlets should be encouraged to provide information about the regulator at the end of articles detailing experiences of online harms.

57

Recommendation 57:

If a decision to make structural changes to the regulator includes a change to its name, a major campaign re-launching the regulator should be conducted. The timing of this campaign should be coordinated to align with major changes to the Act.

12.2 Education and capacity building to prevent online harms

eSafety undertakes a range of education activities to help inform Australians about how to stay safe online and use technology safely. Educating people about how to keep themselves safe online has, up to this point, been one of the few means of preventative action available to eSafety under the Act. Resources produced by eSafety encourage behavioural change and aim to reduce the likelihood of online harms occurring.

While the focus of the Act would shift significantly towards prevention through duty of care obligations on platforms, it would continue to be important for community members to be vigilant and to know their rights and where to seek help when things do go wrong.

The range of resources and projects that eSafety has produced to educate the community, both directly and through important sectors such as education and community services, is impressive. Resources are co-designed with key audiences and underpinned by research and ongoing consultation, so they are appropriately tailored to specific groups.

For example, a dedicated section of the eSafety website includes resources for First Nations people, including information available in 9 Indigenous languages. There is also an area of the website, eSafety Kids, created for children, with a simple layout, messages and language and appealing characters representing different safety attributes (safe, curious, kind and secure).

There is a separate area for young people with simple graphics covering a range of relevant topics, ranging from 'my nudes have been shared' to 'how to be an upstander'.

The site also has tabs leading to tailored information and resources for educators, parents, women, seniors, communities (Culturally and Linguistically Diverse, LGBTQIA+, living with disability, sports hub) and industry. Each of these sections contains not only information about relevant issues, but also links to activities and resources such as webinars, videos and relevant research, such as the eSafety's Women in the Spotlight program which is aimed at professional women with an active online presence and teaching them how best to protect themselves from abuse.

In another example, the eSafety Commissioner partners with the Department of Social Services and Good Things Foundation to provide the in-person supports offered under the Be Connected program. Be Connected training courses are aimed at developing the digital skills of older Australians so they can confidently engage with digital devices and use the internet safely. This program is also delivered in person, due to the nature of the target group. In 2023–24, there were 3,902 attendees at Be Connected webinars and 412,811 learners accessed learning resources on the Be Connected website. Some 30,429 learners also attended in-person Be Connected classes.

Schools and frontline services

Training is provided to key audiences with direct influence on children and young people, including educators, parents and carers. Other audiences include frontline workers supporting people experiencing family, domestic and sexual violence, those working with clients in vulnerable situations and communities, senior Australians, and others. The following initiatives are some examples.

eSafety has developed a Best Practice Framework for Online Safety Education. This aims to establish a consistent national approach to online safety education that supports education systems across Australia to deliver high quality programs with clearly defined elements and effective practices.

eSafety engages with online safety education providers through the Trusted eSafety Provider program, where approved online safety education providers help raise awareness of eSafety's role and resources when delivering their online safety education programs. In 2023–24, nearly 1.4 million school students, parents and educators participated in training run by education providers endorsed under this program. eSafety has also set up an eSafety Champions Network in schools to champion the resources of eSafety and deal with online safety issues in their school and community.

Professional development and training opportunities are available to frontline workers, such as disability support workers and those who work with people experiencing technology-facilitated abuse. eSafety Women educates frontline workers and specialists about gender-based online violence against women and provides tools for identifying and responding to technology-facilitated abuse. In 2023–24, over 15,000 individuals participated in frontline training and professional learning sessions.

In 2023-24, the Online Safety Grants Program provided funding of $3 million to recipients under the Preventing Tech-Based Abuse of Women Grants Program, which forms part of the Government's commitment to the aims and objectives of the National Plan to End Violence against Women and Children 2022–32.

Interestingly, although the 2023 Online Safety Issues survey mentioned above found relatively low awareness of eSafety and its role, almost 80 per cent of parents reported that their children had received information about online safety at school, but most either did not know or did not specify who produced the content. I understand that under a new agreement, participants in the Trusted eSafety Provider program are required to raise awareness of eSafety specifically, which is likely to substantially boost the prominence of eSafety and understanding of its role.

The range, variety and quality of educational resources on eSafety's website are excellent and they should be resourced to continue with this work. A key consideration will need to be around how to ensure that all affected communities can find out that these resources exist and to ensure that barriers such as technological and general literacy and language barriers can be overcome. This is likely to require more outreach, including in person in remote communities and at targeted locations, and ongoing funding should be provided for this resource intensive but important work.

Another gap is reaching parents who may be less engaged with online safety concerns but who should be. As noted at one roundtable of parents and carers I met with:

There is real diversity in parental digital literacy and having a suite of resources that reflects that and doesn't talk down to parents about their level of understanding is really important.

The challenge is how do you reach that cohort of parents who are responsible for the children who are behaving badly online but are not motivated, or don't have the time.[217]

This is a perennial challenge for campaigns promoting behaviour change, which I will not attempt to solve here, but will need further consideration by those with appropriate expertise. The solution is likely to involve finding ways for people to accidentally discover messaging and to frame it in a way that is engaging and non-judgemental. One idea is to get it incorporated as a story line in popular entertainment offerings. Another option is to tie in with existing communities of belonging, such as sporting codes. eSafety is already doing work in this area and there are opportunities to do more, as discussed below.

217 Online Safety Act Review Roundtable.

12.3 Strategic partnerships to promote online safety

To date, eSafety has been very active in leveraging free marketing by building relationships with media organisations such as SBS, the ABC and commercial channels, peak bodies, state and territory education departments, police, community and sporting organisations (including the AFL, NRL and Netball Australia) to boost its profile in the community.

For example, eSafety engaged in a cross promotion with the Australian Football League in mid-July 2022, which reached approximately 25,000 at the stadium plus an additional 840,000 thousand viewers on live and streamed television.[218] The AFL's collaboration with eSafety appears to have been mutually beneficial noting the issue of online abuse of public figures also includes high profile AFL players. In its submission the AFL indicated that it would like this partnership to continue:

The AFL is committed to continuing to work with the eSafety Commissioner, our clubs and across our sphere of influence, to continue to protect those in our industry from online harm.[219]

In return, as noted in the submission, the AFL is able to give eSafety a platform to reach a broad range of communities, including those that can be difficult to engage through other means:

We know that some social groups, including Aboriginal and Torres Strait Islander people, people from LGBTQIA+ communities, people with particular religious beliefs and older Australians, may be at higher risk of online harm. Because of the wide cross-section of Australians who love and play our code, ensuring that everyone can be safe online is a high priority.[220]

The collaborative work eSafety does is vital to increasing its reach and promoting awareness of complaints schemes, powers and resources. eSafety should continue to build these relationships, with a focus on forming connections to on the ground services in communities (including rural and remote communities and First Nations), as well as sources of broad engagement such as sporting codes and high-profile events.

218 eSafety and AFL event data (July 2022, unpublished). Source: Game event undertaken as part of the memorandum of understanding between eSafety and the AFL to help improve online safety for AFL players, fans and the broader community and reaise awareness of the steps Australians can take to '#PlayItFairOnline'. See AFL and eSafety commit to #PlayitFairOnline | eSafety Commissioner.

219 Review submission 154 – Australian Football League, 4.

220 Ibid.

12.4 Research, consultation and evaluation to inform eSafety's work

eSafety research

Research is important to ensure that eSafety's work is evidence based and appropriately targeted.

eSafety undertakes research to inform its activities, support the development of programs and resources as well as support its broader online safety advocacy work.

For example, in 2023-24, eSafety published research on a range of issues including:

- Technology-facilitated family, domestic and sexual violence
- Young peoples' attitudes towards online pornography and age assurance
- The digital experiences of young people with disability
- Experiences in the metaverse
- The digital experiences of LGBTQIA+ teens
- The digital experiences of young people
- Requests for child sexual exploitation on online platforms; and
- The risks and benefits of online gaming for children and young people.

Community input to develop online safety resources

In April 2022, based on recommendations from youth engagement and online safety research, eSafety set up the Youth Council, made up of members aged 13-24 from diverse locations, genders and backgrounds.[221] The Youth Council makes sure that that young people's views and experiences are considered when developing resources, determining priority areas, and improving engagement and awareness of eSafety among young people.

eSafety's Advisory Committee and its new National Online Safety Education Council informs briefings on trends in online safety reports and emerging issues, examining recent research, and sharing quality online safety education programs and resources.

They also foster greater cooperation with Government, Catholic and independent school education sectors in each state and territory.

Evaluations of education and awareness raising activity

It is important that eSafety's major campaigns and programs are able to be evaluated independently to ensure their effectiveness, support continuous improvement and innovation and to promote accountability around investment in this activity. It is also valuable to be able to systematically reflect on and identify improvement opportunities in smaller initiatives, without the need for an external evaluation which can be resource intensive.

eSafety regularly evaluates its education programs, awareness-raising efforts, and regulatory activities. I understand that eSafety has both commissioned independent evaluations and engaged evaluation experts to design evaluation frameworks for eSafety to use internally.

eSafety is also planning to conduct a brand awareness survey, which will be an important information source to guide future awareness raising work. Future awareness research should seek to determine not only levels of awareness but understanding about how to seek help. Insights should also be sought on where people are finding out about eSafety so that resourcing can be appropriately targeted.

221 Moody, L, Marsden, L, Nguyen, B & Third, A (2021) Consultations with young people to inform the eSafety Commissioner's Engagement Strategy for Young People, Young and Resilient Research Centre, Western Sydney University: Sydney.

GOVERNANCE – A FUTURE-PROOFED REGULATOR

13

The functions and powers of the eSafety Commissioner have increased substantially since the creation of the role in 2015. This has also been at a time when the operating environment has become more complicated and contested. A new governance structure is needed to meet these challenges and future-proof the regulator. The recommended governance structure is for eSafety to adopt a Commission model comprised initially of a Chair, Deputy Chair and a Commissioner to support collective decision-making. As eSafety functions and responsibilities increase over time, there are also compelling reasons for eSafety to transition into a standalone regulatory agency, and be established as a separate Commission. Importantly, eSafety must be appropriately resourced so that it can deliver on its mandate of protecting Australians online.

13.1 A Commission model will see better decision-making in an increasingly complex environment

A regulator that can best deliver online safety for all Australians is one that:

- Is deeply knowledgeable about the online sector and the regulatory environment
- Can be dynamic and adapt in response to changing community expectations and technological innovations; and
- Brings a diversity of views and experiences to ensure all decisions are thoroughly considered and well-informed.

The Commissioner has done an excellent job administering the Act and her track record in delivering positive online safety outcomes for Australians is undeniable. However, as the Commissioner's legislative mandate continues to increase and as the operating environment becomes more complex and challenging, there is merit in considering whether alternative governance models may be required.

In its 2014 paper, *The Governance of Regulators*, the Organisation for Economic Cooperation and Development, identified three types of governance structures for independent regulators:

- **Governance board model** – the board is generally responsible for governance, risk management and strategy. Regulatory decision-making would be delegated to a chief executive officer (CEO) and staff, with the board being responsible for the appointment of the CEO and monitoring performance and compliance with the law, the body's governing documents and policies.
- **Commission model** – the board is responsible for collectively making most substantive regulatory decisions.
- **Single member regulator** – an individual is appointed and is responsible for making most substantive regulatory decisions, and delegates other decisions to staff (such as the current eSafety Commissioner model).[222]

The Organisation for Economic Cooperation and Development's guidance goes on to provide a range of circumstances where multi-member decision making model may be appropriate, including where:

- There is benefit in having a diversity of experiences and perceptions that can be brought to bear on substantive regulatory decisions, particularly when decisions involve a high level of judgement
- There is importance in ensuring consistency of decision-making over time – where regulatory decisions require a high degree of judgement, a multi-member decision-making body provides for more 'corporate memory' over time and the ongoing development of expertise; and
- Where independence of decision-making is critical – a multi-member decision-making body is less likely to be susceptible to industry or political influence in comparison to a single decision maker.

There are strong arguments for favouring a multi-member decision-making model in the context of the Act. The Commissioner's decisions under the Act can be complex, contested, and subject to media and public scrutiny. Should the main recommendations in this report be accepted, the need to appropriately exercise regulatory judgement will be even more important.

In light of this, the scope for collective decision-making facilitated through the governance board and Commission models are particularly attractive. A distinct advantage of the Commission model is that members have greater involvement in making significant decisions, allowing them to leverage from their substantial experience, while also continuing to develop their expertise on online safety matters. The Commission model has been successfully operationalised in the Australian context, including by the Australian Communications and Media Authority (ACMA) and the Australian Competition and Consumer Commission (ACCC).

222 Organisation for Economic Co-operation and Development (2014) OECD Best Practice Principles for Regulatory Policy – the Governance of Regulators, 69.

This model is appropriate for both eSafety's existing governance arrangements, and if it were to become a standalone regulatory agency, should that recommendation be adopted (see below). This model shares the burden of decision-making and will enable multiple perspectives to be considered to ensure robust decision-making. It will also result in greater independence (both in reality and perception).

58

Recommendation 58:

To support collective decision making, the regulator should move to a Commission model of governance and be known as the 'Online Safety Commission'.

© Getty Images. Credit: Halfpoint Images.

13.2 The Commission must be appropriately skilled and transparent in its conduct

Commission composition and the right mix of experts

The Online Safety Commission (the Commission) should be comprised of a Chair, a Deputy Chair and a Commissioner, with flexibility in legislation to appoint up to a total of nine Commission members. Allowing for a greater number of Commissioners than is currently needed is intended to futureproof the Commission, ensuring that it has the capacity to expand should the demands and mandate of the regulator evolve.

To ensure the Commission can make the best decisions in an increasingly complex operational and regulatory environment, it is vital that the Chair, Deputy Chair and Commissioner (and other future members) are appropriately skilled and have the right mix of expertise and experiences. Currently, the criteria for appointment of the Commissioner focus on ensuring the officeholder has substantial experience or knowledge and significant standing in at least one of the following fields:

- The operation of social media services
- The operation of the internet industry
- Public engagement on issues relating to online safety; or
- Public policy in relation to the communications sector.

Consideration should be given to whether other expertise should be reflected in the appointment criteria for Commission members. For instance, there may be merit in ensuring that those comprising the Commission have expertise and experience in regulation, technology, economics and market dynamics, and relevant areas of law. While promoting online safety will always be the north star guiding the Commission's decisions, ensuring a collection of skills and expertise will drive better decision-making, and help realise the value of a Commission model of governance.

Internal governance and transparency

A range of measures to support strong internal governance and transparency are listed below:

1. **There should be freedom for the Commission to assign roles and responsibilities to Commission members as it sees fit.** Members could be asked to lead particular thematic areas of the Commission's work, chair subject matter committees and lead the Commission both internally and publicly on their respective areas of focus. A division of responsibility among Commission members would allow for the sharing of the growing workload under the Act, the development of expertise in particular subject matters, and the ability to leverage the unique and existing expertise and experience of Commission members.

2. **The Commission's focus should be on setting strategic direction and making significant decisions.** Significant decisions are those that involve an element of judgment and discretion, where the consequences for an industry participant (or section of industry, or industry as a whole) are significant, or where novel matters arise which require the Commission to set a precedent. This could include decisions such as: determining whether an industry participant has breached their duty of care, whether legal proceedings should be instituted for failure to comply with the Act, the development and finalisation of codes and the publication of key documents such as regulatory guidance and major educational initiatives.

3. **There also needs be clarity on what decisions are to be made by the Commission as a whole, and what decisions are to be delegated to Commission members or senior staff of the Commission.** Not all decisions under the Act would require the consideration of the Commission as a whole. Senior staff require the ability to make informal and formal removal requests under the various complaint schemes, which rely on speedy decisions for minimising online harms.

4. **How the Commission exercises its regulatory functions should be clearly documented and made public.** Like its counterparts the ACCC and the ACMA, the Commission could prepare a Code of Conduct for its members.[223] The Code could address how the Commission makes decisions, the roles and responsibilities of members, the duties of members (taking into account the Act, the Public Governance, Performance and Accountability Act 2013 (PGPA Act) and the Public Service Act 1999), the Commission's values, processes for the identification and resolution of real and perceived conflicts of interest and the purpose and scope of any committees that are created to support the Commission's work.

5. **Consistent with current practice, the Minister for Communications should issue an updated Statement of Expectations to the Commission that outlines the Government's expectations for how the Commission will achieve its objectives, carry out its functions and exercises its powers.** The Commission should be required to respond to the Statement of Expectations through a Statement of Intent outlining how they will meet the Australian Government's objectives. Both statements should be made public.

6. **Establishing common ground with industry can be assisted by having the Commission publish its regulatory priorities for each financial year, to ensure resources are focused on those areas of greatest impact and concern and where improved industry compliance is required.**[224] This will be instructive for the public, and provide industry with an opportunity to lift their standards ahead of any investigations or enforcement actions. Regulatory priorities should still retain flexibility so that the regulator can pivot where there are new or significant risks that require an immediate focus. The Commission should consider the utility of publicly identifying its regulatory priorities for each financial year as other regulators, such as the ACCC, do.

7. The public and industry benefit when the regulator **publishes guidance on how it will administer the schemes under the Act**. This provides for a shared understanding of how the Act is administered by the regulator and ensures consistency of decision-making. The Commissioner has already published a range of highly informative regulatory guidance on various schemes within the Act. This practice should continue, and is all the more necessary for any new schemes that are implemented following this review.

8. The Commission should ensure it has **access to external experts and representative voices** to inform its decision-making. It is unrealistic to expect that a small number of Commission members could adequately represent all the interests that must be considered when arriving at sound regulatory decisions. In particular, the Commission should have mechanisms for hearing the perspectives of groups who are disproportionately impacted by online harms, such as First Nations groups, culturally and linguistically diverse groups, disability groups, LGBTQIA+ groups, young people, and women. Other important voices will be industry, academia, technology experts, educators and other agencies with regulatory functions. The Commissioner has already established foundations for this through bodies such as the Youth Council, the Women in the Spotlight program, and the Digital Platform Regulators' Forum.

223 ACCC Code of Conduct: Code of Conduct for Commission Members and Associate Members 2024 (accc.gov.au). ACMA Code of Conduct: Microsoft Word – Authority Code of Conduct June 2024.docx (acma.gov.au).

224 Compliance and enforcement priorities | ACCC, Compliance priorities 2024–25 | ACMA, OAIC regulatory priorities | OAIC.

9. **Annual reporting** will be crucial in explaining to the public the story of the Commission's work. This story should capture the Commission's successes, the new and enduring challenges, important trends, the efficacy of the schemes under the Act, and the nature and scope of its preventative activities. Currently, the Act requires the Commissioner to prepare an annual report as soon as practicable after the end of each financial year. The Act specifies a range of matters that the annual report should cover, including reporting on the operation of the complaint schemes, formal and informal actions taken to address harmful material, breakdown of particular harms by ground or category, and the number of applications for internal review. Annual reporting on these matters should continue and reflect any new statutory obligations under the new Commission. Ideally, data should also be published more frequently than annually and be consistent with the Consumer Policy Research Centre's proposals for best practice data publication.[225]

   The annual report should also include enforcement actions taken such as the institution and status of court proceedings, external merits review processes, education and awareness raising activities that have been undertaken, collaboration with other Australian government departments and agencies and with international counterparts, and other information relating to the Commission's functions under the Act.

   As noted in the Minister's current Statement of Expectations for eSafety, it is important that annual reports are prepared in line with best practice principles and consistent with the PGPA Act. The Statement of Expectations also require eSafety to produce detailed corporate plans – this will be another vital governance practice that the Commission should continue.

225 Consumer Policy Research Centre (2024) Am I the only one, 6. Am I The Only One - CPRC.

59

Recommendation 59:

That the Commission should be comprised of a Chair, Deputy Chair and a Commissioner, with flexibility for the Commission to grow up to nine members as the functions and powers of the regulator increase.

60

Recommendation 60:

That in moving to a Commission, the Act should require Commission members to have an appropriate mix of skills to support informed and robust decision-making.

61

Recommendation 61:

That a newly formed Commission has strong internal governance processes, is transparent in how it does its work and ensures that it reports meaningfully on its performance.

13.3 Transitioning to a standalone Commission to support a growing regulatory remit

The existing governance arrangements are unusual

The establishment of the Commissioner in 2015[226] as a statutory office operating with the support of the ACMA was appropriate when the Commissioner's role was limited to protecting children online and management of the Online Content Scheme. In the following years, the Commissioner's role was expanded to cover the online safety of all Australians (2017), administration of the image-based abuse scheme (2018)[227], powers to deal with abhorrent violent material (2019)[228] and the management of numerous additional programs seeking to promote online safety.

With the implementation of the Act, there came a suite of new functions and powers such as the adult cyber abuse scheme, the Basic Online Safety Expectations framework and industry codes. Importantly, the Act carried over several key governance measures from the former *Enhancing Online Safety Act 2015* to support the Commissioner's independence. This includes a clear statement that the Commissioner is not subject to direction by the ACMA in relation to the performance of her functions or the exercise of her powers[229], and that amounts from the Online Safety Special Account (which funds eSafety) cannot be debited from the Account without the written approval of the Commissioner[230].

Crucially, the Act also made improvements to eSafety's governance arrangements by making the Commissioner an official of the ACMA for the purposes of the finance law (within the meaning of the PGPA Act). A practical effect of this was to enable the ACMA Chair to delegate certain functions to the Commissioner and eSafety staff (who are ACMA employees). Delegations made by the ACMA Chair relating to financial matters and staffing have enhanced eSafety's independence.

However, these arrangements are dependent on strong relationships, goodwill, and trust between the ACMA and eSafety, and are ultimately at the discretion of the ACMA Chair. While the ACMA Chair has ensured that the Commissioner can operate as independently as possible, these arrangements may not provide long-term certainty and whenever there is a new ACMA Chair, new delegations will need to be negotiated and implemented.

Relevantly, the ACMA Chair also ultimately remains the Accountable Authority for eSafety under the PGPA Act and the Agency Head for eSafety under the *Public Service Act 1999*. This means, for instance, that for the purposes of the PGPA Act, it is the ACMA Chair that is responsible for ensuring the proper use and management of eSafety resources. While arrangements can be put in place to ensure that the ACMA Chair can comply with her responsibilities as the Accountable Authority while simultaneously providing eSafety with the functional and financial independence the Online Safety Act envisions, it should be noted that these types of measures, while necessary, are unusual and add a layer of complexity to both ACMA and eSafety's operations.

226 The Office of the eSafety Commissioner was established under the *Enhancing Online Safety Act 2015* (EOSA). Previous to this, the Office was called the Children's e-Safety Commissioner under the *Enhancing Online Safety for Children Act 2015*.

227 Implementation of the *Enhancing Online Safety (Non-Consensual Sharing of Intimate Images) Act 2018*.

228 *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019*.

229 *Online Safety Act 2021*, section 186.

230 *Online Safety Act 2021*, section 190.

© Getty Images. Credit: Fly View Productions.

The Briggs Review (2018)

The question of whether the governance arrangements for the Commissioner remain appropriate is not new. Ms Lynelle Briggs AO who undertook a statutory review (the Briggs Review) of the then *Enhancing Online Safety Act 2015* in 2018, considered this same question.

The Briggs Review made two recommendations about eSafety's governance structure:

- That governance arrangements be improved by moving the eSafety Commissioner and her Office (along with associated ASL [average staffing level], contractors, resources, programs and responsibilities) out of the ACMA and into the Department of Communications and the Arts. Under this arrangement, the eSafety Commissioner would retain the independence of her office in a new departmental online safety stream of business and assume responsibility for a new departmental online safety outcome, all of her staff and resources, and be brought under the *Public Service Act 1999* and the *Public Governance, Performance and Accountability Act 2013*, with the special purpose vehicle being abolished.
- That these new governance arrangements be reviewed after a transition period of 3 years to assess the possibility of setting up a new standalone online safety entity.

At the time, these recommendations were supported in principle but were not ultimately implemented.

A standalone independent regulator is the ideal end state

Making the Commissioner an official of the ACMA, and other developments such as service level agreements for corporate services, have no doubt smoothed the operation of the existing governance arrangements. However, with eSafety's functions and responsibilities growing (and with a commensurate increase in its staffing profile), there is a legitimate question as to whether eSafety will outgrow the current governance arrangements.

The ideal end state is for eSafety to transition into a standalone, independent regulatory agency (which would also be called the **Online Safety Commission**). This will provide eSafety with total independence in how it conducts its work, both internally and externally, and sets them up to succeed in an increasingly complex and rapidly evolving operating environment. Alternative options such as moving eSafety into the Department are not preferable as there may still be limitations on its independence.

Making eSafety a standalone entity, would see the Commissioner become the Accountable Authority under the PGPA Act and captured by the *Public Service Act 1999*, increasing transparency and accountability of the regulator. The creation of an independent eSafety builds on eSafety's existing record of trust with stakeholders.

A standalone entity would also align eSafety's governance arrangements to its closest comparable agencies, including the OAIC, the ACCC, the ACMA, and international counterparts such as Ofcom in the United Kingdom.

Transitioning to a standalone regulatory agency will require considerable effort and planning, and must be managed carefully. It would be a great disservice to eSafety and the people eSafety protect if these reforms impacted on eSafety's ability to exercise its regulatory functions. Should the main recommendations of this report be adopted, it will substantially increase eSafety's responsibilities and remit, and this must be where the most energy and time is initially dedicated. The transition will take time to ensure that critical online safety functions and other reforms can be delivered while the transition is designed and implemented. Care must also be taken to ensure that any new and urgent work that the ACMA must undertake, such as under the proposed mis/disinformation legislation, is not unduly disrupted through the transition process.

Additionally, it will be important to ensure that in transitioning eSafety to a standalone regulator, that there are no adverse resourcing implications for the ACMA that will impact its ability to deliver on its own remit.

Creating a new standalone agency is also likely to involve additional costs in the short-term, including needing to have specific corporate and administration functions, and establishing new systems, processes and internal governance bodies. The benefits of eSafety transitioning to a standalone regulator agency must be carefully balanced against these costs and any benefits the existing governance arrangements provide.

Some of the costs could be minimised through corporate services agreements. Shared services will generate efficiencies and leverage economies of scale. But any such agreements must also recognise that eSafety may have unique needs (such as specialist IT assets to support its investigative and oversight functions) that may exceed what shared services may be able to provide.

While the existing governance arrangements are far from perfect, the challenges they pose are not insurmountable in the short term.

62

Recommendation 62:

That following consideration of the regulator's functions and responsibilities under a new regulatory framework, the regulator should transition to a standalone, independent regulator to support its growing functions and responsibilities, and to future-proof the regulator.

13.4 eSafety must be appropriately resourced and set up to succeed

eSafety must be appropriately resourced to fulfil its mandate. The public expect eSafety to protect them online and carry out its functions to the fullest extent possible. If eSafety cannot deliver on its objectives effectively and promptly, the public's confidence will erode over time. The regulator must be given the tools to succeed and adequate funding is vital in supporting and amplifying eSafety's success.

The recommendations made throughout this report, if implemented, will lead to increased responsibilities and workload for eSafety. With any new functions and powers, there must be accompanying resourcing. Otherwise, the regulator will be left to make unenviable resourcing decisions, including diverting resources from other mission critical and legislated functions – this is neither desirable nor appropriate.

eSafety must be supported by strong regulatory infrastructure. For instance, eSafety should have a well-funded legal team with a General Counsel capable of running multiple enforcement and litigation actions. The complex and contested operating environment means complicated and novel legal challenges that eSafety must be ready to face. Strong regulatory infrastructure also means having corporate oversight and officers who manage the day-to-day operations of the regulator. This could be achieved in many ways, including by having roles such as a chief operating officer or a chief executive officer. Strong corporate leadership is also necessary to support eSafety's transition to becoming a standalone regulatory agency.

In addition to having the right number and mix of personnel, eSafety must also be equipped with the technological resources required to meet the operating environment. Appropriate IT infrastructure is key to ensuring that eSafety can carry out its investigative and oversight functions (such as evidence gathering, technical research, auditing and testing algorithms) effectively and efficiently, and for protecting the often highly sensitive information it holds.

These changes would bring eSafety in line with most other regulators and make sure that eSafety is set up for success.

63

Recommendation 63:

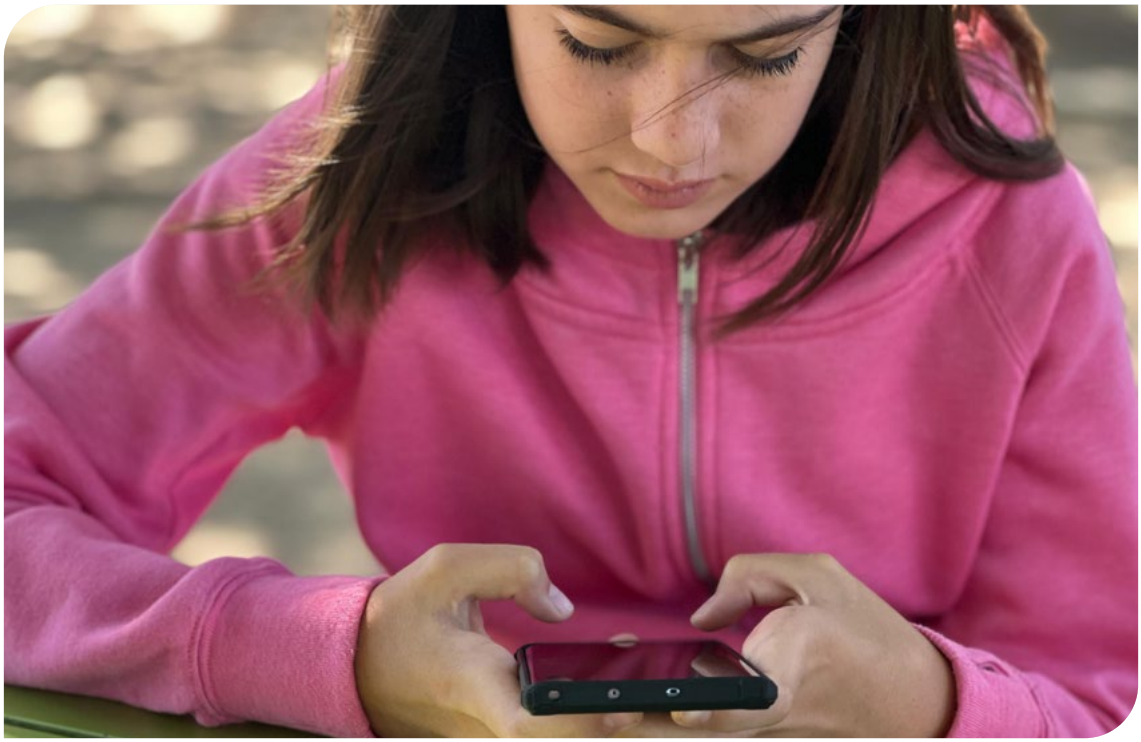
That the regulator should be appropriately resourced to implement the right regulatory infrastructure and carry out its functions. This includes having an ongoing dedicated and appropriately resourced legal team, appropriate corporate management and the information technology it needs to do its job well. Consideration should be given to how other regulators operate to determine what may be appropriate in the regulator's context.

13.5 Industry must bear the cost of regulation

Funding for eSafety could either come from the Government (as it does currently), from industry or a combination of both. Industry funding would typically occur through cost recovery initiatives such as an industry levy. Cost recovery involves charging non-government entities, like private business, for some or all of the efficient costs[231] of a specific government activity.[232] The relevant government activity in this instance is online safety regulation which is carried out by the eSafety Commissioner.

There is a growing international trend towards charging industry for the cost of online safety regulation, with counterparts in the United Kingdom, the European Union, Ireland and Canada all enacting or proposing to enact legislation establishing cost recovery mechanisms.

The rationale for cost recovery is clear. Australians experience online harms because of the online services that are made available. The role of the eSafety Commissioner was established for the purpose of preventing, minimising, and investigating online harms that Australians are or may be subject to, and where appropriate, taking regulatory action against industry where they fail to meet their obligations under the Act. In this instance, it is appropriate that industry bears the cost of eSafety's activities, rather than the general public.



© Getty Images. Credit: Rafael Ben-Ari.

231 'Efficient costs' means the minimum costs necessary to provide a particular government activity while achieving the policy objectives and legislative functions of the Australian Government (Australian Government Cost Recovery Policy | Department of Finance).

232 Australian Government Cost Recovery Policy | Department of Finance.

Details of a cost recovery mechanism should be developed in consultation with industry

While the case for cost recovery may be self-evident, the design and implementation of a cost recovery mechanism requires considerable thought. Given that this issue is also being considered in other contexts, this review of the Act is not the appropriate vehicle to determine the details of a cost recovery mechanism.

Careful consideration will be needed to identify which activities should be cost recovered. The costs associated with the administration of any schemes within the Act and taking regulatory actions should be subject to cost recovery. However, eSafety also has a range of other functions under the Act, such as coordinating Commonwealth departments and agencies on online safety matters, and preparing reports and papers. Consideration should be given to whether it would be appropriate to cost recover for these and other non-regulatory functions of eSafety.

Another critical question, and potentially the most complex, is to identify which industry participants will be subject to cost recovery which will require further consultation with industry. There are various options for this. For instance, cost recovery could be directed at all online service providers who are captured within the scope of the Act – this may be the most equitable approach for cost recovery, but identifying and recovering costs from all online service providers captured under the Act, especially those that have a limited presence in Australia may be complex and resource intensive, if it can be achieved at all.

Alternatively, cost recovery could be directed at all online service providers captured within the scope of the Act who meet particular thresholds. Examples of potential thresholds include those services who reach a particular number of average Australian monthly active users, or those services with over a specific amount in annual revenue in Australian markets, or a combination of these. Consideration could also be given to a threshold that enables cost recovery to be directed at online service providers who are considered high-risk. It is also worth exploring whether there are other metrics by which to determine who is in scope of cost recovery.

The advantage of a more targeted basis for charging industry is that it would focus on those online services that create the most regulatory burden for eSafety and where significant issues are most likely to arise. However, consideration must be given to whether targeting a particular section or subsection of industry could have adverse consequences.

In determining how the cost would be apportioned among industry participants, there will be lessons learned from our international counterparts who are working through these same questions. Consultation with industry will be vital in ensuring that the feasibility and implications of particular cost recovery models are fully understood before any decisions on a cost recovery model are taken.

A fair and reasonable cost recovery mechanism must be supported by rigorous transparency. Consistent with the Australian Government Cost Recovery Policy, eSafety will be required to document its cost recovery activities in a cost recovery implementation statement (CRIS) before commencement.[233]

64

Recommendation 64:

A cost recovery mechanism should be developed to fund the cost of regulating industry, with details to be settled by government in consultation with industry.

233 Ibid.

A REFORM PATHWAY

14

The recommendations contained in this report are substantial and will be a step change in how our online safety laws operate. Should the key recommendations in this report be adopted, their development and implementation will take time to get right. However, this does not detract from the urgency of implementing the recommendations as soon as practicable and, if required, prioritising those changes that provide the most immediate benefits to Australians.

This review has also highlighted one of the enduring challenges of attempting to regulate the online world – that it is continuously evolving and governments all over the world are constantly playing catch-up. While I have attempted to future-proof the Act with the recommendations outlined in this report, the Act should be subject to regular statutory reviews to ensure we have the right policy settings, and that the Act is operating effectively. A future review could consider whether there is benefit in consolidating the regulation of online harms rather than relying on the current patchwork of administrative, legislative and regulatory arrangements which can be complex and create inefficiencies.

14.1 Priority areas for online safety reform

The breadth of reforms contemplated by this report are broad and their implementation will be complex. While this is a matter for Government, it is useful to consider whether it is preferable to attempt to implement these recommendations in one extensive legislative reform project, or instead to separate out the reforms into several tranches of smaller legislative projects. One argument in favour of the latter approach is that while one significant legislative reform project could take years to develop and pass through Parliament, breaking the project down into thematic and manageable parts will likely see swifter development and implementation. The immediate benefit of this approach is that the most important online safety protections could be in place sooner.

Implementing a duty of care and supporting eSafety are the first priority

The headline reform proposed in this report is for Australia to adopt a duty of care approach to prevent online harm. As outlined in Chapter 5, an overarching duty of care places responsibility on service providers to take reasonable steps to address and prevent foreseeable harms on their service. A duty of care is a priority for two important reasons – (a) it will be the most effective and immediate means of improving online safety for Australians, and (b) because once the details are settled, online services must be given a reasonable time to adapt to the new regulatory model.

Implementing a duty of care model also requires a range of associated measures identified in this report to be implemented, including risk assessment, mitigation, measurement and transparency measures, and settling new categories of industry sections. To ensure that the duty of care obligation is enforceable, changes to eSafety's investigatory and enforcement powers must also be implemented to ensure eSafety can carry out its responsibilities under the duty of care framework and that there are appropriate penalties to incentivise online services to comply with their duty of care.

The urgency of progressing these reforms must be balanced with ensuring they are subject to appropriate scrutiny and the model and detailed consideration is given to developing a robust and clever approach. The proposed duty of care model and associated reforms are complex and, in the time available to complete this review, I am sure I have not addressed all the nuances required to implement a duty of care in the Australian context. Consultation will be key to getting these critical reforms right, including reflecting on international experiences so that we can proactively address the challenges they encountered in implementing similar frameworks.

These reforms must also be supported by structural reforms to eSafety. The duty of care will bring new and complex regulatory challenges. It is exactly these types of challenges that the new Commission model of governance is intended to address, allowing for robust and collaborative decision-making. It follows that enabling eSafety to move to this new governance structure, comprised initially of a Chair, Deputy Chair and Commissioner, must also be progressed as a priority. eSafety will also need a critical injection of new funding to deliver the duty of care reforms, and any other new functions that are progressed as a priority by Government. Creating new regulatory schemes without ensuring the regulator can effectively implement, administer and enforce them will undermine the utility of important protections for the Australian public.

14.2 Measures to support Australians online (safety nets) must also be prioritised

Outside of the duty of care and associated reforms, the most pressing changes will be those that improve the experiences of Australians online. In this sense, improving the operation of some or all of the Act's four complaints schemes (child cyberbullying, adult cyber abuse, non-consensual sharing of intimate images, and the Online Content Scheme) will be key. These schemes are where people go for immediate action to address an online harm they have suffered, including online hate (which will also be addressed through the duty of care reforms). As important as it is for our regulatory framework to adopt a preventative approach to ensure that harms do not occur in the first place, it must still be able to respond effectively when harms do occur.

Other reforms will take considerable thought and transitioning arrangements to implement, especially transitioning eSafety to a standalone regulatory agency and implementing a cost recovery framework.

65

Recommendation 65:

That if required, the Government should prioritise implementation of the key reforms arising from this review that will provide the most substantial and immediate online safety protections for Australians, including in particular the new duty of care and associated reforms. This should coincide with eSafety moving to a Commission model of governance and appropriate resourcing to support the implementation of priority reforms.

© Getty Images. Credit: Strauss/Curtis.

14.3 The Online Safety Act must be regularly reviewed

This review has consistently highlighted the tremendous change that has occurred in the online environment since the commencement of the Act in January 2022. This trajectory of change is unlikely to slow. It is hard to say with any confidence what the digital landscape will look like in a few years' time, let alone a decade from now. None of this would be a concern but for two salient factors: first, that the rapid evolution of the digital environment creates new online harms, and amplifies old ones, and second, our increasingly ubiquitous presence in these digital spaces makes these harms almost unavoidable.

The Act must evolve to serve the needs of Australians now and into the future. I believe the recommendations contained in this report will help future-proof our regulatory framework and provide a pathway towards a modern, fit-for-purpose online safety regulator. However, as experience suggests, we must stay vigilant by constantly assessing the effectiveness of our regulatory settings.

Governments often amend legislation on an ad-hoc basis to address urgent or significant shortcomings, but the importance of systematically considering the operation of legislation is vital. It allows for a consideration of the Act as a whole, how the pieces fit together, creates space for big picture thinking, and provides opportunities for improving existing frameworks through minor but important amendments. A thorough review process also ensures meaningful consultation with industry and civil society so that the experiences of individuals, communities and industry can be considered.

Consistent with the approach adopted in existing section 239A of the Act, a review should be conducted within three years of the commencement of any key reforms to the Act or by 2029, whichever is the earliest. The review must test the effectiveness of any reforms to ensure it is meeting its objectives and providing appropriate protections for Australians.

66

Recommendation 66:

That the updated Act be subject to independent review within three years of the commencement of the key reforms to the Act, or by 2029, whichever is earliest.

14.4 On the horizon – the case for a Digital Services Commission

One of the quirks of Australia's federal online safety framework is the separation of various online harms into different ministerial portfolios and portfolio agencies. This has made the regulatory landscape an exceedingly complex one. For ordinary Australians who are not familiar with this bureaucratic labyrinth, it must be even more perplexing. For instance:

- The Minister for Communications is responsible for certain online safety harms covered under the Act with the relevant regulator being the eSafety Commissioner
- The Minister for Communications is also responsible for other harms such as the proposed mis/dis information scheme with the relevant regulator being the Australian Communications and Media Authority (ACMA)
- The Assistant Treasurer is responsible for the Australian Consumer Law and the Government's anti-scams agenda, with the relevant regulator being the Australian Competition and Consumer Commission (ACCC)
- The Attorney-General is responsible for the Privacy Act. Privacy regulation has and will continue to grow in its impact on the online world, with the relevant regulator being the OAIC
- The Minister for Industry, Science and Resources is responsible for Australia's approach to safe and responsible artificial intelligence and potential mandatory guardrails for artificial intelligence in high-risk settings; and
- The Minister for Cyber Security has primary responsibility for Australia's cyber policy coordination and setting the strategic direction of Government's cyber effort. A number of security and law enforcement agencies are involved in mitigating cyber security risks and managing incident response.

While these arrangements are acceptable, consideration should be given in the future to what the ideal online harms regulatory framework should look like. Options that could be considered, include consolidating online harms under one Minister, one portfolio or a single regulator, such as a broader independent Digital Services Commission. While these matters are beyond the scope of this review, the merits of such an approach should be investigated.

During the course of this review, I have engaged with numerous federal departments and agencies to understand the interaction and linkages between the Act and the work they do. Understanding these legislative frameworks and where the policy responsibility of one Minister ends, and another begins, has been challenging, particularly as the digital landscape continues to evolve. This makes coordinating policies and adopting consistent regulatory approaches to online harms more challenging, and limits the ability of Government to respond dynamically and nimbly to emerging harms and risks.

Indeed, the second interim report of the Joint Select Committee on Social Media and Australian Society appears to acknowledge the need for consolidation and improved coordination on online harms, recommending that "the Australian Government establish a Digital Affairs Ministry with overarching responsibility for the coordination of regulation to address the challenges and risks presented by digital platforms."[234]

Ministerial responsibilities aside, I do see particular benefit in there being one regulator responsible for regulating a range of online harms, if not all.

Having one central online harms regulator would strengthen regulatory expertise and bring more resources to bear that could be quickly drawn upon to respond to regulatory changes in the operating environment. The regulator would be able to accurately assess the impacts of these changes through multiple lenses, deepen its understanding of the sector it regulates and ultimately result in better decisions.

A single regulator with a clear set of objectives, regulatory positions and priorities will make better decisions and will become a source of confidence and trust for both industry and the public. It is also likely that with this weight of authority and trust, the regulator's actions and interventions

234 Joint Select Committee on Social Media and Australian Society, October 2024. Second interim report: digital platforms and the traditional news media, Recommendation 1.

will have a greater impact in changing industry behaviour for the better. This would be a significant improvement on the disparate and fragmented regulatory environment that currently exists which makes industry compliance more difficult than it needs to be, and adds layers of complexity.

Consolidating a range of online harms into one regulator also amplifies the effectiveness of the regulator and creates economies of scale. For instance, regulators generally have a range of non-regulatory functions, such as education, research, awareness-raising, providing customer-facing services, international engagement and advisory functions. Funding one regulator to undertake these activities would build capability and strengthen these functions.

This is not to say that our current network of online harms regulators do not have the relevant expertise or are failing to collaborate and coordinate their regulatory activities. Indeed, the Digital Platform Regulators Forum, comprised of the eSafety Commissioner, the ACCC, the ACMA, and the OAIC, are doing valuable work to share information and cooperate on cross-cutting issues and activities involving the regulation of online services. However, this collaboration is voluntary and none of the participants are funded to do this work. Moreover, the group cannot make binding decisions that would impose any obligations of any of its members. Ultimately, they operate under different legislative frameworks, answer to different Ministers, and their overarching policy objectives are not the same. If there was one central online harms regulator, this kind of collaboration would occur organically and would result in more consistent regulatory approaches.

There are also benefits to industry and the public in consolidating online harms regulation. For industry, it would lead to a more consistent regulatory approach, and provide a central point of contact for engagement on online harms, thereby reducing regulatory burden, though the regulator would need to take great care to avoid regulatory capture. For the public, the principal benefit is clarity – there will be a one-stop shop if they want to make complaints or receive assistance in relation to online harms. Victims of online harms can be distressed or find themselves in a vulnerable situation and nothing can be more frustrating and disheartening than being transferred from one government agency to another in an attempt to identify who can best resolve their concerns.

While establishing one overarching Digital Services Commission is likely to be a costly exercise in the short-term, it would provide significant savings in the long-run by reducing the number of regulators who have online harms functions. This centralisation of online harms regulation would create the types of efficiencies outlined above, and likely result in the more effective use of limited resources.

The consolidation of online harms regulation into one central regulator could be supported by consolidating various legislation relating broadly to online harms into one legislative framework. This framework would no doubt be considerable in size – covering potentially everything from the harms captured under the Online Safety Act, to the proposed mis/dis information scheme, and matters relating to consumer safeguards and privacy. However, there are inherent benefits to having one overarching legislation that all stakeholders (including the public) can go to understand their rights and responsibilities.

I appreciate that changes of this magnitude take an enormous amount of political will and implementing such reforms would likely take years. While not an immediate matter for consideration, I flag it as something that should be considered in the future.

67

Recommendation 67:

That the Government consider how its existing administrative arrangements relating to online harms are operating and whether there is a case for having a central online harms regulator. Given the level of change that needs to happen now to better protect Australians, this consideration may be best left to around the time of the next review.

APPENDIX A— GLOSSARY

Term	Meaning
ABC	Australian Broadcasting Corporation
Abhorrent violent conduct	Defined in section 474.32 of the Criminal Code. Includes a person engaging in a terrorist act, murder or attempted murder, or torture, rape or kidnapping of another person.
Abhorrent violent material	Defined in section 474.31 of the Criminal Code. Includes audio, visual or audio-visual material that records or streams abhorrent violent conduct engaged in by one or more persons, and that a reasonable person would regard in all the circumstances as being offensive. The material must also have been produced by a person or persons, each of whom is: • engaged in the abhorrent violent conduct, • conspired to engage in the abhorrent violent conduct, • aided, abetted, counselled, procured, or were in any way knowingly concerned in the abhorrent violent conduct, or • who attempted to engage in the abhorrent violent conduct. It is immaterial whether the material has been altered, or whether the abhorrent violent conduct was engaged in within or outside Australia.
ACCC	Australian Competition and Consumer Commission
ACMA	Australian Communications and Media Authority
AFL	Australian Football League
AFP	Australian Federal Police
AI	Artificial intelligence
App distribution service	Defined in section 5 of the *Online Safety Act 2021*. A service that enables end-users to download apps, where that download is by means of a carriage service. Examples include Apple App Store, and Google Play Store.
APS	Australian Public Service
Basic Online Safety Expectations	The Basic Online Safety Expectations are determined under the *Online Safety Act 2021* and set out the Australian Government's expectations of the steps that should be taken by providers of social media services, messaging services, gaming services, apps and certain other sites accessible from Australia to keep Australians safe online. The *Online Safety Act 2021* provides eSafety with powers to require services to report on their compliance with the Basic Online Safety Expectations.
Broadcasting Services Act	*Broadcasting Services Act 1992*

Term	Meaning
Caching service	A type of intermediary service that includes automatic, intermediate and temporary storage of information provided by a service recipient in a communication network for the purpose of making the onward transmission to other recipients on request more efficient. For example, a content delivery network (temporary storage or caching of files in geographically distributed servers to reduce the page loading time). (See Article 3 of the European Union's Digital Services Act).
Class 1 material – section 106 of the Online Safety Act 2021	Material that is or would likely be refused classification under Australia's National Classification Scheme, by reference to the National Classification Code. It includes material that: • depicts, expresses or otherwise deals with matters of sex, drug misuse or addiction, crime, cruelty, violence or revolting or abhorrent phenomena in such a way that they offend against the standards of morality, decency and propriety generally accepted by reasonable adults to the extent that they should not be classified • describes or depicts in a way that is likely to cause offence to a reasonable adult, a person who is, or appears to be, a child under 18 (whether the person is engaged in sexual activity or not), or • promotes, incites or instructs in matters of crime or violence. Class 1 material includes, for example, child sexual exploitation and abuse material and pro-terror material.
Class 2 material – section 107 of the Online Safety Act 2021	Material that is, or would likely be, classified under Australia's National Classification Scheme, by reference to the National Classification Code as either: • X 18 + (or, in the case of publications, category 2 restricted), or • R 18 + (or, in the case of publications, category 1 restricted), which is legally restricted to adults. Class 2 materials include, for example, pornography and other high impact material such as R 18 + video games.
Classification Act	*Classification (Publications, Films and Computer Games) Act 1995*
Commissioner	eSafety Commissioner
Criminal Code	*Criminal Code Act 1995* (Commonwealth)
Designated internet service	Defined in section 14 of the *Online Safety Act 2021*, and only to the extent that material on the service is accessible to or delivered to one or more end-users in Australia. A service (other than a social media service, relevant electronic service, or on-demand program service) that allows end-users to access material on the internet using an internet carriage service or a service that delivers material to persons by means of an internet carriage service. This includes most apps and websites accessed by Australian end-users including retail websites, information apps (such as train timetables), and adult websites.
EU Digital Services Act	Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

Term	Meaning
Hosting service	Defined in **Australia** as a service that hosts stored material that has been provided on a social media service, relevant electronic service, or designated internet service (see section 17 of the *Online Safety Act 2021*). Includes, for example Amazon Web Services. Defined in Europe as a type of intermediary service that stores information provided by a service recipient at their request. For example, cloud services. (See Article 3 of the European Union's Digital Services Act).
Internet carriage service	Defined in section 5 of the *Online Safety Act 2021* as "a listed carriage service that enables end-users to access the internet." The internet carriage service is provided to the public by an internet service provider. Examples include Optus, Telstra, and TPG Telecom Limited.
Internet search engine service	A service designed to collect, organise and/or rank material on the internet, that have the sole or primary purpose of allowing end-users to search the service's index of material for results in response to the end-user's queries, and the service returns search results in response to the query. Examples include Google Search, Microsoft Bing and Yahoo! Search.
ISP	Internet service provider.
Manufacturers, suppliers and installers of equipment	Definition applies where equipment is for use by end-users in Australia in connection with a social media service, relevant electronic service, designated internet service or internet carriage service. Examples include Apple and Samsung.
Material that depicts abhorrent violent conduct	Defined in section 9 of the *Online Safety Act 2021* as audio, visual or audio-visual material that records or streams abhorrent violent conduct. It is immaterial whether the material has been altered or who produced it.
Mere conduit service	A type of intermediary service that transmits information provided by a service recipient in a communication network, or provides access to a communication network. For example, virtual private networks, internet exchange points, or domain name system services. (See Article 3 of the European Union's Digital Services Act).
NGO	Non-government organisation
OAIC	Office of the Australian Information Commissioner
Ofcom	UK Office of Communications, the regulator for communications services in the United Kingdom.
Privacy Act	*Privacy Act 1988*
PGPA Act	*Public Governance, Performance and Accountability Act 2013*
Relevant electronic service	Defined in section 13A of the *Online Safety Act 2021*, and only to the extent that material on the service is accessible or delivered to one or more end-users in Australia. A service that allows end-users to communicate with other end-users by means of email, instant messaging, short message service (SMS), multimedia message service (MMS), chat service or online game. Examples include Roblox, Gmail, and WhatsApp.
Regulatory Powers Act	*Regulatory Powers (Standard Provisions) Act 2014*
SBS	Special Broadcasting Service

Term	Meaning
Sections of the online industry (defined in Part 1 of the Online Safety Act 2021)	Groups consisting of the providers of: • Social media services • Relevant electronic services • Designated internet services • Internet search engine services • App distribution services • Hosting services • Internet carriage services So far as those services are provided to end-users in Australia; and • The group consisting of persons who manufacture, supply, maintain or install equipment for use by end-users in Australia in connection with a social media service, relevant electronic service, designated internet service or internet carriage service.
Social media service	Defined in section 13 of the *Online Safety Act 2021*, and only to the extent that material on the service is accessible or delivered to one or more end-users in Australia. A service that has the sole or primary purpose of enabling online social interaction between end-users, where end-users can also link to other end-users and post material on the service. Examples include Facebook, Instagram, Tik Tok, and YouTube.
Stevens Review	Review of Australian classification legislation, a 2020 report by Neville Stevens.
UK Online Safety Act	*Online Safety Act 2023* (United Kingdom Government)

APPENDIX B— TERMS OF REFERENCE

Terms of reference – Statutory Review of the *Online Safety Act 2021*

Context

Online interactions are a part of the everyday life of nearly all Australians. Spending time online provides opportunities to connect with each other and with community. While the online environment has enabled significant benefits across society and the economy, these technologies also provide avenues for malicious activities that can harm individuals and erode social cohesion.

Australia's *Online Safety Act 2021* (the Act), which commenced in January 2022, supports Australians online by providing the eSafety Commissioner (the Commissioner) with powers to address cyberbullying of children, cyber abuse of adults, illegal and restricted content, and the non-consensual sharing of intimate images. The Act makes online service providers more accountable for the online safety of Australians who use their services through the Basic Online Safety Expectations (BOSE) and the development of industry codes or industry standards.

The Act currently defines 'Class 1' and 'Class 2' material which underpin the industry codes and standards regime through reference to the Australian National Classification Scheme and the classification or likely classification of the material. The Government will conduct a separate public consultation process in the first half of 2024 to inform the development of options for the second stage of reforms to ensure the National Classification Scheme is fit-for-purpose in the modern media environment.

History of online safety legislation

On 1 July 2015, the *Enhancing Online Safety for Children Act 2015* commenced, establishing the Children's eSafety Commissioner as an independent statutory office holder, supported by the Australian Communications and Media Authority (ACMA), to take a national leadership role in online safety for children.

The Act was renamed to the *Enhancing Online Safety Act* in 2017 and the Children's eSafety Commissioner became the eSafety Commissioner following changes to the Act to broaden the Commissioner's role to online safety for all Australians, not just children.

In 2018, an independent review of the *Enhancing Online Safety Act 2015* and Schedules 5 and 7 of the *Broadcasting Services Act 1992* was conducted by Ms Lynelle Briggs AO (the 2018 Review). The 2018 Review recommended that there be a single up-to-date Online Safety Act that would allow key elements of the framework to be modernised and improved. In 2021, the Act passed Parliament, creating a modernised and fit for purpose regulatory framework that built on existing legislative regimes for online safety. The Act commenced on 23 January 2022.

The Act established the BOSE framework, which is a key part of the Act and underpin efforts to improve transparency and accountability of platforms and keep Australians safe from online harm. Industry is also expected to do more to keep its users safe, including by developing mandatory, enforceable industry codes designed to protect Australians from illegal and restricted online content (Online Content Scheme). If a code does not meet statutory requirements under the Act, the Commissioner can develop an industry standard for that section of the online industry instead.

On 24 January 2022, the *Online Safety (Basic Online Safety Expectations) Determination 2022* (BOSE Determination) came into effect. The Commissioner has powers to seek information from service providers about how they are meeting the expectations outlined in the BOSE Determination. Further information about the reporting notices issued by the Commissioner can be found at *Basic Online Safety Expectations | eSafety Commissioner*.

Industry bodies are developing codes in a two-phased approach. The first phase is focused on the most seriously harmful online content by reference to Class 1 of the National Classification Scheme, including Class 1A (child sexual exploitation material, pro-terror material and extreme crime and violence material) and Class 1B (crime and violence material and drug-related material).

The Commissioner registered six industry codes in 2023, covering social media services, internet carriage services, equipment providers, app distribution services, hosting services and internet search engine services. Five codes came into effect on 16 December 2023, and one will come into effect on 12 March 2024. The Commissioner declined to register two draft industry codes and is now developing industry standards for relevant electronic services and designated internet services.

eSafety will soon commence work with industry on the development of a second phase of codes which will focus on Class 1C material and Class 2 material (which include online pornography and other high-impact material). Further information about the development of industry codes including regulatory guidance outlining Class 1 and Class 2 materials can be found at *Industry codes | eSafety Commissioner*.

Legislative basis for the Review

Section 239A of the Act states:

239A Review of operation of this Act

(1) Within 3 years after the commencement of this section, the Minister must cause to be conducted an independent review of the operation of this Act.

(2) The Minister must cause to be prepared a written report of the review.

(3) The Minister must cause copies of the report to be tabled in each House of the Parliament within 15 sitting days of that House after the day on which the report is given to the Minister.

As part of the *Government's response to the House of Representatives Select Committee on Social Media and Online Safety Report*, the Government committed to undertaking and completing the Statutory Review earlier than required under the Act, and within this term of Government, so that the Act can keep pace with the evolving online environment.

Matters to be considered by the Review

The Act does not prescribe particular provisions to be examined by the Review. Accordingly, the Review will be broad ranging and include consideration of:

1. The overarching objects in section 3 of the Act, including the extent to which the objects and provisions of the Act remain appropriate to achieve the Government's current online safety policy intent.
2. The operation and effectiveness of the following statutory schemes and whether the regulatory arrangements should be amended:
    a. cyber-bullying material targeted at an Australian child[235]
    b. non-consensual sharing of intimate images[236]
    c. cyber abuse material targeted at an Australian adult[237]

235 The statutory scheme for cyberbullying material targeted at an Australian Child is described in Part 5 of the Act. 'Cyber-bullying material targeted at an Australian child' has the meaning given in section 6 of the Act.

236 The statutory scheme for non-consensual sharing of intimate images is described in Part 6 of the Act. 'Non-consensual intimate image of a person' is defined in section 16 of the Act.

237 The statutory scheme for cyber abuse material targeted at and Australian adult is described in Part 7 of the Act. 'Cyber-abuse material targeted at an Australian adult' has the meaning given by section 7 of the Act.

d. the Online Content Scheme,[238] including the restricted access system and the legislative framework governing industry codes and standards, and

e. material that depicts abhorrent violent conduct.[239]

3. The operation and effectiveness of the Basic Online Safety Expectations (BOSE) regime in the Act.

4. Whether additional arrangements are warranted to address online harms not explicitly captured under the existing statutory schemes, including:

   a. online hate

   b. volumetric (pile-on) attacks

   c. technology-facilitated abuse and technology-facilitated gender-based violence

   d. online abuse of public figures and those requiring an online presence as part of their employment

   e. other potential online safety harms raised by a range of emerging technologies, including but not limited to:

      - generative artificial intelligence
      - immersive technologies
      - recommender systems
      - end-to-end encryption,
      - changes to technology models such as decentralised platforms

5. Whether the regulatory arrangements, tools and powers available to the Commissioner should be amended and/or simplified, including through consideration of:

   a. the introduction of a duty of care requirement towards users (similar to the United Kingdom's *Online Safety Act 2023* or the primary duty of care under Australia's work health and safety legislation) and how this may interact with existing elements of the Act

   b. ensuring industry acts in the best interests of the child

6. Whether penalties should apply to a broader range of circumstances.

7. Whether the current information gathering powers, investigative powers, enforcement powers, civil penalties or disclosure of information provisions should be amended.

8. The Commissioner's functions and governance arrangements, including:

   a. the Commissioner's roles and responsibilities under the Act

   b. whether the current functions and powers in the Act are sufficient to allow the Commissioner to carry out their mandate.

9. Whether the current governance structure and support arrangements for the Commissioner provided by the ACMA are fit for purpose for both the Commissioner and the ACMA.

10. Whether it would be appropriate to cost recover from industry for eSafety's regulatory activities.

238 The Online Content Scheme is described in Part 9 of the Act.

239 The statutory scheme for material that depicts abhorrent violent conduct is described in Part 8 of the Act. 'Material that depicts abhorrent violent conduct' is defined in section 9 of the Act. 'Abhorrent violent conduct' is defined in section 5 of the Act as having the same meaning as in Subdivision H of Division 474 of the Criminal Code.

Process and timing

The Minister for Communications has appointed Ms Delia Rickard PSM to undertake the Review. Ms Rickard will be supported by staff from the Department of Infrastructure, Transport, Regional Development, Communications and the Arts.

The Review will involve a period of public consultation, commencing with the release of an Issues Paper in the first half of 2024. This will be accompanied by a call for public submissions, with the intention to conduct follow up stakeholder meetings as required. Subject to the discretion of the Reviewer, consultation may be conducted with relevant stakeholders, including but not limited to: industry, non-government organisations, community support groups, Members of Parliament, the Commissioner, ACMA, the Australian Federal Police and other law enforcement agencies, international regulatory bodies, Commonwealth, state and territory government agencies, and other interested groups and individuals.

The Final Report of the Review will be provided to the Minister for Communications by 31 October 2024, for tabling in Parliament within 15 sitting days as required by section 239A of the Act. Any recommendations made by the Review will be carefully considered by Government and responded to at the appropriate time.

APPENDIX C— STAKEHOLDER ENGAGEMENT

Meetings with individuals and organisations

Throughout the course of the review I met with the following individuals and organisations:

- Adobe
- Alannah and Madeline Foundation (AMF)
- Amazon
- ANU Tech Policy Design Centre
- Apple
- Attorney-General's Department
- Australian Catholic Bishops Conference
- Australian Christian Lobby
- Australian Communications and Media Authority
- Australian Federal Police (AFP) Commissioner
- Australian Human Rights Commission
- Australian Muslim Advocacy Network (AMAN)
- Brian Hay (Cultural Cyber Security)
- Dolly's Dream
- Department of Home Affairs
- Department of Social Services
- DIGI
- Domestic, Family, Sexual Violence Commission (Assistant Commissioner)
- eSafety Commissioner
- eSafety Youth Council
- European Commission
- Executive Council of Australian Jewry
- Federal Court and Family Court of Australia
- First Nations Digital Inclusion Advisory Group
- Frances Haugen
- Google
- International Justice Mission
- It's Time we Talked
- Judge John Cain, Victorian State Coroner
- Meta
- Microsoft
- Ofcom
- Online Hate Prevention Institute
- Online Safety Act Network (UK)
- People with Disability Australia
- Qoria
- Race Discrimination Commissioner
- Reset.Tech Australia
- Reset.Tech Youth Group
- Special Envoy to Combat Antisemitism
- The Honourable Robert French AC
- Snap Inc
- Synod of Victoria and Tasmania (Uniting Church in Australia)
- Tattarang
- Teach Us Consent
- Telstra
- TikTok
- Twitch
- University of Western Australia Tech and Policy Lab
- X Corp
- YouTube

Roundtables

I convened seven roundtables with civil society organisations and individuals comprising approximately 77 participants. The roundtables considered a range of online safety issues and were centred around specific themes and were attended by the following participants:

Date	Theme	Attendees
July 2024	Online safety issues for young people	• Alannah and Madeline Foundation (AMF) • Dolly's Dream • Carly Ryan Foundation • Centre for Excellence in Child and Family Welfare • Children and Media Australia • Daniel Morcombe Foundation • Domestic, Family and Sexual Violence Commission • Families Australia • Headspace • Minus18 • Multicultural Youth Advocacy Network • National Association for Prevention of Child Abuse and Neglect (NAPCAN) • NSW Office of the Advocate for Children and Young People • Our Watch • Project Rockit • Queensland Family and Child Commission • Reset.Tech Australia • Teach Us Consent • UNICEF Australia • Youth Law Australia
July 2024	Experiences of community groups disproportionately impacted by online harms	• ACON • Australian Muslim Advocacy Network • Australian National Imams Council • Commonwealth Attorney-General's Department *(observing only)* • Council on the Ageing • Disability Advocacy Network Australia • Executive Council of Australian Jewry • Federation of Ethnic Communities' Council of Australia • Islamophobia Register Australia • LGB Alliance Australia • LGBTIQ+ Health Australia • Muslim Women Australia • Online Hate Prevention Institute • South Australian Council on Intellectual Disability • Women with Disabilities Australia

Date	Theme	Attendees
July 2024	Technology-facilitated gender-based violence	• Australia's National Research Organisation for Women's Safety (ANROWS) • Domestic Violence Connect (QLD) • Full Stop Australia • Monash University • National Women's Safety Alliance • No to Violence • Our Watch • Royal Melbourne Institute of Technology (RMIT) • Safe and Equal • Safe Steps Family Violence Response Centre • University of Technology Sydney (UTS) • Wesnet
August 2024	Online safety issues for First Nations people	• Aboriginal and Torres Strait Islander Advisory Council on Family, Domestic and Sexual Violence • First Peoples Disability Network Australia • Tangentyere Women's Family Safety Group • Victorian Aboriginal Child Care Agency • Secretariat of National Aboriginal and Islander Child Care • The Healing Foundation • First Nations Digital Inclusion Advisory Group • 13YARN • Djirra
August 2024	Online child sexual exploitation and abuse	• NAPCAN • International Justice Mission • Body Safety Australia • International Centre for Missing and Exploited Children • Stop it Now • National Centre on Child Sexual Abuse • Australian Childhood Foundation • Domestic, Family and Sexual Violence Commission
September 2024	Online safety issues for parents and carers	• Australian Institute of Family Studies • Australian Parents Council • Australian Council of State School Organisations • Catholic School Parents Queensland • Council of Catholic School Parents NSW & ACT • Federation of Parents and Citizens Associations of NSW • Triple P
September 2024	Online abuse of public figures	• Australian Broadcasting Corporation • Australian Football League • Trawalla Foundation • Ginger Gorman, Journalist • name withheld, Journalist • name withheld, Journalist

Public Submissions

On 29 April 2024, the issues paper was released inviting feedback from the public. In total there were 168 substantive submissions[240] and over 2,100 comments received. The submissions listed below will be published on the Department of Infrastructure, Transport, Regional Development, Communications and the Arts website, and excludes those submissions that were marked private and confidential.

Submission 1 – Anonymous

Submission 2 – Leo A

Submission 3 – Anonymous

Submission 4 – Gordon

Submission 6 – Ms R Stirling

Submission 7 – Norma Braun

Submission 9 – James Longfield

Submission 10 – Ken Mewha

Submission 11 – Australians for Social Justice

Submission 12 – Judith Buchan

Submission 14 – Anonymous

Submission 16 – Jean Linis-Dinco

Submission 17 – James

Submission 18 – Steve Venning

Submission 20 – Associate Professor Marilyn Bromberg

Submission 21 – Asia-Pacific Development, Diplomacy & Defence Dialogue

Submission 22 – Michele Browne

Submission 23 – Commissioner for Children and Young People WA

Submission 24 – Anonymous

Submission 25 – David Ross

Submission 26 – Robert Ristevski

Submission 28 – Anonymous

Submission 29 – David A W Miller

Submission 30 – Professor Dan Jerker B. Svantesson

Submission 31 – Jane Munro

Submission 33 – Dr Stephen Jones

Submission 34 – LGB Alliance Australia

Submission 35 – National Women's Safety Alliance

Submission 36 – David Rohde

Submission 37 – Anonymous

Submission 38 – Internet Australia

Submission 39 – Alannah & Madeline Foundation

Submission 40 – Alcohol and Drug Foundation

Submission 41 – International Centre for Missing and Exploited Children Australia

240 Private or confidential submissions have been omitted from this list. One submission which contained a corrupted file was also omitted as we were unable to reach the submitter to receive an updated file.

Submission 42 – Queensland Human Rights Commission
Submission 44 – Peter Fam LLB (of Maat's Method) and Alex Hatzikalimnios LLB
Submission 45 – Greg Tannahill
Submission 46 – Alexander Hatzikalimnios
Submission 47 – Australian Feminists for Women's Rights
Submission 48 – Internet Society
Submission 49 – Dolly's Dream
Submission 50 – Queensland Family & Child Commission
Submission 51 – Qoria Limited
Submission 52 – Reddit, Inc.
Submission 53 – ACT | The App Association
Submission 54 – Alcohol Change Australia
Submission 55 – Relationships Australia National Office
Submission 56 – Lynzi Ziegenhagen, CEO, Bandio PCB
Submission 57 – Black Ink Legal
Submission 58 – Foundation for Alcohol Research and Education
Submission 59 – Eros Association
Submission 60 – Dr Martin Husovec and Souha Al-Fihri (LSE)
Submission 61 – Allies for Children
Submission 62 – J Cameron and T Winning
Submission 63 – Free Speech Union of Australia
Submission 64 – Uniting Church in Australia, Synod of Victoria and Tasmania and Synod of Queensland
Submission 65 – Centre for Excellence in Child and Family Welfare
Submission 66 – Jonathan Green
Submission 68 – Online Safety Act Network
Submission 69 – Anonymous
Submission 70 – Reset.Tech Australia
Submission 71 – Centre for AI and Digital Policy
Submission 72 – Internet Association of Australia Ltd
Submission 73 – eSafety Commissioner
Submission 74 – Australian Catholic Bishops Conference
Submission 75 – Executive Council of Australian Jewry
Submission 76 – Corynne McSherry
Submission 77 – Information and Privacy Commission NSW
Submission 78 – Amaze
Submission 79 – UNICEF Australia
Submission 80 – Crispin Rovere
Submission 81 – Prevention United
Submission 82 – Orygen
Submission 83 – Gary Christian

Submission 84 – yourtown

Submission 85 – Australian Child Rights Taskforce

Submission 86 – ARC Centre of Excellence on Automated Decision-Making and Society

Submission 87 – Food for Health Alliance

Submission 88 – Children and Media Australia

Submission 89 – Scarlet Alliance

Submission 90 – Microsoft

Submission 91 – Tiffany Burleigh

Submission 92 – Sandra Anthony

Submission 93 – David Johnson

Submission 94 – Anonymous

Submission 95 – International Justice Mission

Submission 96 – Global Network Initiative

Submission 98 – Australian Christian Lobby

Submission 100 – The Hon. Zoe Daniel MP

Submission 101 – SBS

Submission 102 – Consumer Electronics Suppliers Association

Submission 104 – Mega Limited

Submission 105 – Department of Industry, Science and Resources

Submission 106 – Rob Cover and Jennifer Beckett, RMIT Digital Ethnography Research Centre

Submission 107 – Institute of Public Affairs

Submission 108 – Anonymous

Submission 109 – Privacy and Digital Rights organisations (Joint Submission)

Submission 110 – Free TV

Submission 111 – Emma Baillie

Submission 112 – Digital Rights Watch

Submission 113 – Merillot Pty Ltd

Submission 114 – Arved von Brasch

Submission 115 – Hannah Petocz and Bridget Harris

Submission 117 – Australia New Zealand Screen Association

Submission 120 – Anthony

Submission 121 – Butterfly Foundation

Submission 122 – AAWAA (Affiliation for Australian Women's Action Alliances)

Submission 123 – Anonymous

Submission 124 – Mark Nottingham

Submission 125 – Per Capita

Submission 127 – Australian Gaming and Screens Alliance

Submission 129 – Our Watch

Submission 130 – Youth 4 Online Safety

Submission 131 – Advocate for Children and Young People

Submission 132 – National Office for Child Safety

Submission 133 – Epic Games

Submission 134 – UTS Centre for Media Transition

Submission 135 – Australian Human Rights Commission

Submission 136 – Tasmanian Government

Submission 137 – Collective Shout

Submission 138 – ABC

Submission 139 – AMAN Foundation Ltd (Joint Submission)

Submission 140 – The International Social Games Association

Submission 141 – Australian Federal Police

Submission 142 – Interactive Games & Entertainment Association (IGEA)

Submission 143 – The Hon. Kate Chaney MP

Submission 144 – DIGI

Submission 145 – X Corp

Submission 146 – Office of the Australian Information Commissioner

Submission 147 – Tech Council

Submission 148 – Tech Policy Design Centre

Submission 149 – Law Council of Australia

Submission 150 – TikTok Australia

Submission 151 – WESNET

Submission 152 – Tattarang

Submission 153 – AFL

Submission 154 – NSW Government

Submission 155 – Human Rights Law Centre

Submission 156 – Snap Inc.

Submission 157 – Communications Alliance Ltd

Submission 158 – Human Technology Institute

Submission 159 – Interactive Advertising Bureau (IAB) Australia

Submission 160 – Google

Submission 161 – Youth Law Australia

Submission 163 – American Chamber of Commerce

Submission 164 – LinkedIn

Submission 165 – University of Western Australia Tech and Policy Lab

Submission 166 – Meta

Submission 167 – Online Hate Prevention Institute

Submission 168 – Telstra

Submission 169 – First Nations Digital Inclusion Advisory Group

APPENDIX D—COMPLAINT AND CONTENT-BASED SCHEMES

Overview of complaint and content-based removal schemes

	Image-based abuse	Child cyberbullying	Adult cyber abuse	Illegal and restricted content
Online harm	Posting or threatening to post an intimate image depicting another person without that person's consent (irrespective of whether the image has been altered).	Online material that is likely intended to have an effect on an Australian child, and likely to have the effect of seriously threatening, seriously intimidating, seriously harassing or seriously humiliating the Australian child.	Online material that is likely intended to have the effect of causing serious harm to an Australian adult and would reasonably be regarded as menacing, harassing or offensive in all the circumstances.	Online material that is Class 1 or Class 2 material (determined by reference to Australia's National Classification Code).
Who is protected?	Person depicted (or purported to be depicted).	A targeted child (who is ordinarily resident in Australia).	A targeted adult who is ordinarily resident in Australia.	End-users in Australia.
Link required to Australia	The person depicted, or the person who posted or threatened to post, is ordinarily resident in Australia (or, for objection notices only, the image is hosted in Australia).	Material is targeted at a child ordinarily resident in Australia ('Australian child').	Material is targeted at an adult ordinarily resident in Australia ('Australian adult').	Material suspected to be accessible to Australians online. Class 1 material can be hosted anywhere, but Class 2 material must be provided by a service in Australia or hosted in Australia.
Who can make a complaint?	The person who has reason to believe an intimate image depicting them has been shared without consent (or that a threat to share such an image has been made); a person authorised by the depicted person; or a parent or guardian of the depicted person.	The targeted Australian child or a parent, guardian or responsible person authorised by the child or an adult who was an Australian child.	The targeted Australian adult or responsible person authorised by the Australian adult.	A person who resides in Australia, or an entity that carries out activities in Australia, or an Australian Government. (Note, eSafety can investigate material within this scheme without receiving a complaint).
Does complainant need to report to the service provider before a removal notice can be issued?	No.	Yes.	Yes.	No (can be reported anonymously).

Commissioner's complaint scheme compliance and enforcement powers

	Image-based abuse	Child cyberbullying	Adult cyber abuse	Illegal and restricted content
Formal warning to person who posts or threatens to post image	Yes	No	No	No
Removal notice to service provider/hosting service provider	Yes	Yes	Yes	Yes
Removal notice to end-user	Yes (a 'removal notice')	Yes	Yes (a 'removal notice')	No
Remedial direction to end-user	Yes	No	No	No
Remedial notice to service provider	No	No	No	Yes (Class 2 only)
Service provider notification	Yes	Yes	Yes	No
Service provider statement	Yes	Yes	Yes	Yes
Link deletion notice	No	No	No	Yes (Class 1 only)
App removal notice	No	No	No	Yes (Class 1 only)
Federal Court order to cease providing service	No	No	No	Yes (in exceptional situations)
Alternative enforcement arrangements	Formal warnings, enforceable undertakings, court injunctions, infringement notices, civil penalty orders and financial penalties.			

These compliance and enforcement powers involve the following:

- **Removal notice:** Notice requiring recipient (end-user, service provider or hosting service provider) to remove material or stop hosting material within 24 hours or longer period the Commissioner allows (civil penalty for non-compliance).
- **End-user notice:** Notice requiring person who posted cyberbullying material targeted at a child ordinarily resident in Australia to: remove the material and/or refrain from posting cyberbullying material targeting the child and/or apologise for posting the material (enforceable by injunction).
- **Remedial direction:** Direction to end-user who has posted or threatened to post intimate images without consent to take specified remedial action to prevent future contraventions (civil penalty for non-compliance).
- **Remedial notice:** Notice requiring the recipient to remove Class 2 material or to make the material subject to a Restricted Access System (civil penalty for non-compliance).[241]
- **Link deletion notice:** Notice requiring internet search engine provider to cease providing a link to Class 1 material where material subject to a removal notice in the last 12 months has not been removed and the link has been used to access the material at least twice in a 12-month period (civil penalty).
- **App removal notice:** Notice requiring an app distribution service to cease the ability for end-users in Australia to download an app used to facilitate the posting of Class 1 material at least twice in a 12-month period and where the material has not been removed following a removal notice issued in the past 12 months (civil penalty).
- **Service provider notification:** Notification to service provider advising that material on the service has been found to be the specific harmful material defined under the complaint scheme.
- **Service provider statement:** A statement that the Commissioner has found multiple occurrences of harmful material being posted/hosted on the service within a 12-month period, in contravention of the service's terms of use, and that may be published on the Commissioner's website.
- **Federal Court order:** In the most exceptional situations, the Commissioner can apply for Federal Court orders for a social media service, relevant electronic service, designated internet service or internet carriage service to cease providing their service in Australia. An order can only be made where a service provider has, on two or more occasions in the past 12 months, contravened a civil penalty provision in the Online Content Scheme and as a result the continued operation of the service represents a significant community safety risk.
- **Alternative enforcement arrangements:** The Act adopts the enforcement arrangements set out in the *Regulatory Powers (Standard Provisions) Act 2014* for civil penalties, infringement notices, enforceable undertakings and injunctions. The Commissioner can issue formal warnings, issue an infringement notice, accept an enforceable undertaking, seek a court ordered injunction, and/or pursue civil penalties for non-compliance with a requirement under the Act depending on the case circumstances. eSafety's Compliance and Enforcement Policy outlines the matters considered when determining what the preferred compliance and enforcement actions are in a particular situation.

241 The Commissioner may declare by legislative instrument that a specified access control system is a restricted access system (*Online Safety Act*, section 108).

APPENDIX E—VARIATIONS IN HATE SPEECH PROTECTIONS ACROSS AUSTRALIA

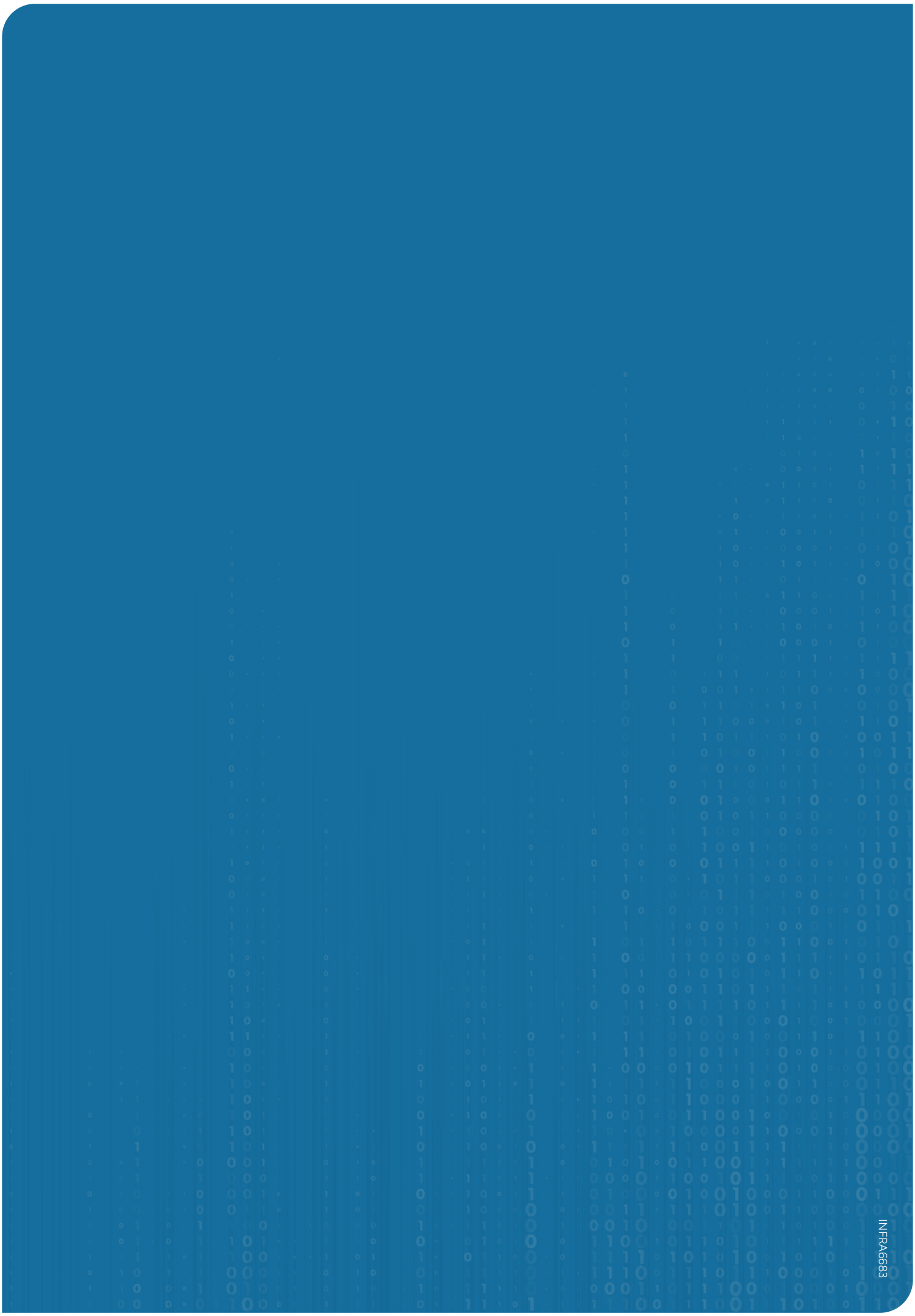
Matrix of hate speech protections and legislation at State and Federal levels in Australia[242]

	COMMONWEALTH (FEDERAL)	ACT	NSW	NT
Legislation	Racial Discrimination Act 1975 Sex Discrimination Act 1984 Human Rights Commission Act 1986 Disability Discrimination Act 1992 Age Discrimination Act 2004 Fair Work Act 2009 Sex Discrimination and Fair Work Amendment Act 2021	Discrimination Act 1991 (ACT)	Anti-Discrimination Act 1977 (NSW)	Anti-Discrimination Act 1996 (NT)
Inclusions				
Race	✓	✓	✓	✓
Sex	✓	Under Federal	Under Federal	Under Federal
Age	✓	Under Federal	Under Federal	Under Federal
Disability	✓	✓	Under Federal	Under Federal
HIV/Aids	✓	✓	✓	×
Sexuality	×	✓	✓	×
Gender Identity	×	✓	✓	×
Intersex	×	✓	×	×
Religion	×	✓	✓	×

242 Purpose (2023). Online Hate Speech in Australia: The Role of News Media and Pathways for Change. Part Two: Curbing Dehumanising Hate Speech Online, 23. The reviewer does not guarantee, and accepts no legal liability for, the accuracy, reliability, currency or completeness of information included in this table.

Matrix of hate speech protections and legislation at State and Federal levels in Australia *cont.*

	QLD	SA	TAS	VIC	WA
Legislation	Anti-Discrimination Act 1991 (QLD)	Civil Liability Act 1936 (SA) Equal Opportunity Act 1984 (SA)	Anti-Discrimination Act 1998 (TAS)	Racial and Religious Tolerance Act 2001 (VIC) Equal Opportunity Act 2010 (VIC)	Equal Opportunity Act 1984 (WA)
Race	✓	✓	✓	✓	Under Federal
Sex	Under Federal	Under Federal	✓	Under Federal	Under Federal
Age	Under Federal	Under Federal	Under Federal	Under Federal	Under Federal
Disability	Under Federal	Under Federal	✓	Under Federal	Under Federal
HIV/Aids	×	×	×	×	×
Sexuality	✓	×	✓	×	×
Gender Identity	✓	×	×	×	×
Intersex	×	×	✓	×	×
Religion	✓	×	✓	✓	×



INFRA6683