



Inquiry into Social Media and Online Safety

eSafety Commissioner Submission

January 2022

Contents

Key Points	3
1. Introduction	5
A. The Online Safety Act (OSA)	5
B. eSafety’s approach to online safety	7
2. The range of online harms facing Australians on social media.....	13
A. Cyberbullying	14
B. Illegal and Restricted online content	16
C. Image-based abuse	19
D. Adult cyber abuse	21
E. Abhorrent violent material and online crisis events	23
F. Online harms typologies	25
3. Evidence of the potential impacts of online harms on the mental health and wellbeing of Australians.....	27
A. At-risk groups and intersectional factors	28
B. Children and young people	30
4. The transparency and accountability required of social media platforms and online technology companies regarding online harms experienced by Australians	36
A. Safety by Design.....	37
B. International and domestic efforts to enhance transparency	37
C. Gaps in transparency efforts	38
D. Basic Online Safety Expectations	39
5. Evidence of the extent to which algorithms used by social media platforms permit, increase or reduce online harms to Australians	45
A. Moderation algorithms	46
B. Recommendation algorithms	47
C. Algorithmic transparency and regulation	48
6. Evidence of existing ID verification and age assurance policies and practices and the extent to which they are being enforced.....	51
A. Identity verification and anonymity	52
B. Age Assurance	61

7.	Other related matters: Horizon scanning and international engagement	67
A.	Horizon scanning	67
B.	International engagement	68
8.	Conclusion.....	70

Key Points

1. The internet offers tremendous benefits, but it also carries the risk of a broad range of harms. The eSafety Commissioner (eSafety) knows that certain groups and individuals, such as children and young people, and those who are more vulnerable in the offline world, are at greater risk. We need to continue conducting research and building evidence to understand the problems we are trying to solve, and to make sure we tailor solutions that solve those challenges. eSafety will continue striving for better online outcomes for Australians in a way that is balanced, proportionate, coherent, coordinated, and evidence based.
2. In addition to the regulatory powers granted to eSafety by the Government, which includes a legislative mandate to coordinate education and online safety activities across the Commonwealth, we take a holistic approach beyond prevention and protection. This includes significant efforts to achieve greater proactive and systemic change across the technology ecosystem. The *Online Safety Act 2021* (Cth) (OSA), which comes into force on 23 January 2022, will strengthen this proactive and systemic approach.
3. eSafety looks forward to the commencement of the OSA and we will assess and evaluate its impact over time. This is the same process which informed the current reforms based on the threat trends eSafety was seeing, and which allowed us to identify and fill gaps in the legislative and regulatory framework. eSafety will report on operations under the OSA annually, and there will be an independent three-year review. We will keep a watching brief on our new and substantially changed schemes – namely, the Adult Cyber Abuse Scheme and the Online Content Scheme – to determine whether any change is necessary.
4. Through the Basic Online Safety Expectations and the mandatory industry codes, the OSA provides important new options for eSafety to regulate online services' systems and processes, based in part on systemic issues identified through our individual complaints schemes. This will help make sure we can lift safety standards for all Australians and shift the burden for safety from children and families to primarily industry.
5. The Basic Online Safety Expectations are a helpful and flexible tool to drive transparency and accountability on several existing and emerging safety issues, including the impacts of algorithms on user safety. eSafety needs time to implement industry reporting against the expectations before Government determines if there is a need for specific algorithmic auditing powers to combat online harms or other issues. Such powers would need to be carefully crafted and resourced to reflect technical talent requirements and other complexities.
6. The online environment is constantly evolving. eSafety continually monitors for emerging issues and global developments to stay ahead of trends and to pivot accordingly. In addition to following tech trends, challenges and technology paradigm shifts, like the metaverse and decentralisation, we also monitor overseas law reform efforts.

7. While Australia remains an early mover and leader in online safety, there are major legislative movements afoot, particularly in the United Kingdom and the European Union, to regulate the systems and processes of online services more robustly (without the individual complaint schemes eSafety provides). These proposals will come with major penalties for non-compliance.
8. We will need to review eSafety's regulatory toolkit, governance model and capacity over time if we are going to remain 'on par' with these legislative directions. As these potentially far reaching new legislative powers emerge in other jurisdictions, we will want to make sure we have comparably robust means to protect Australians. Of course, we are part of an international regulatory ecosystem, and we welcome new regulators and novel approaches being tested for important regulatory outcomes. We work with many overseas jurisdictions to share what we learn – and to learn from their planned approaches. This allows us to create formal cooperative agreements so we can achieve global impact together.
9. Given the many equities involved in addressing online harms, there is a risk of overlap and duplication. To keep up with the rapid evolution of the online environment, there is also a risk of rushing to impose blunt policy solutions that are not consistent with a balanced and evidence-informed approach. This can result in confusion for stakeholders as well as unintended negative consequences for the public.
10. eSafety is proud to be at the vanguard of so many of these local, national and global efforts. As an organisation, we have developed a level of online safety subject matter and operational expertise that does not exist anywhere else in the world. We are looking forward to releasing our updated eSafety Strategy in the coming months and we will continue striving for a joined-up approach to reducing online harms across Australia. We believe eSafety should lead any future work that comes from this inquiry to take Australia to the next level of online safety, in line with our national coordination mandate.

1. Introduction

The eSafety Commissioner (eSafety) is Australia's independent regulator and educator for online safety. eSafety promotes online safety for all Australians, leads online safety efforts across Australian Government departments and agencies, and works with online safety stakeholders around the world to extend our impact across borders. Established in 2015, our mandate is to make sure Australians have safer and more positive experiences online.

As the world's first government agency dedicated to fostering a safer and more positive online environment, we welcome the opportunity to contribute to this inquiry.

Our submission sets out some of the major risks and harms we have observed in the seven years since eSafety was established. It explains the diverse and important work eSafety is doing and gives insight into what is on the immediate horizon for us – implementing the new *Online Safety Act 2021* (OSA), which takes effect 23 January 2022.

Since 2015, eSafety has become a world-leading authority and regulator of online harms. The world looks to us for thought leadership about online safety – from small economies like Fiji to larger economies like the United Kingdom (UK) and European Union (EU).

Our staff numbers and budgets have grown significantly with the Government investing heavily in this policy space. Our services are in demand because we are successful in helping thousands of Australians to manage their online lives. As a small agency, we are flexible, agile and a unique entity in the Australian Government architecture. We expect demand for our expertise to grow, and as our functions and workstreams evolve, we note there will be a need to consider future resourcing.

We are refreshing our 2019-22 Strategy¹ to reflect the introduction of the OSA and related developments over the past several years. Building on this strategy, and in line with our national coordination mandate to coordinate Government activities relating to online safety for Australians,² we suggest eSafety should lead any national work relating to online safety which may arise from this inquiry, with appropriate resourcing.

A. The Online Safety Act (OSA)

The OSA will give eSafety improved powers to help protect all Australians from the most serious forms of online harm. This will enhance our ability to provide people-focused services and support, in line with our core functions. These functions include supporting and encouraging the implementation of measures to improve online safety for Australians; conducting and publishing research; supporting, encouraging, conducting, accrediting and evaluating educational, promotional and community awareness programs; making grants; and formulating guidelines and statements recommending best practices.

¹ Office of the eSafety Commissioner (eSafety) (October 2019) eSafety strategy 2019-2022 <https://www.esafety.gov.au/sites/default/files/2019-10/eSafety%20Strategy%20Plan.pdf>.

² *Online Safety Act 2021* (Cth) s 27(1)(d).

The key changes resulting from the OSA include:

- **a new, world-first Adult Cyber Abuse Scheme** – enabling eSafety to require the removal of cyber abuse material targeting an Australian adult where the Commissioner is satisfied the material is posted with the likely intention of causing serious harm
- **an enhanced Cyberbullying Scheme for Australian children** – enabling eSafety to require the removal of material from a broader range of online services where under 18s are now spending time
- **a strengthened Image-Based Abuse Scheme** – enabling eSafety to rapidly address the non-consensual sharing of intimate images
- **a modernised Online Content Scheme** – enabling eSafety to act against seriously harmful online content, such as child sexual exploitation material and pro-terror content, no matter where it is hosted
- **a regime for the development of new industry codes or standards** – to create mandatory measures for a broad array of online services to prevent and address the distribution of seriously harmful online content, and to keep children safe from unsuitable content
- **the determination of Basic Online Safety Expectations** for online services – to help make sure those services are safe for Australians to use and to provide greater transparency and accountability in relation to their safety features, policies and practices, and
- **stronger information-gathering powers** – including the ability to obtain identity or contact information about end-users where relevant to the operation of the OSA.

The operation of the OSA is subject to an independent review within three years after commencement. In implementing the OSA, eSafety will report annually on the operation of our schemes. We will also monitor global and domestic regulatory shifts so we can continue to lead and pivot, as necessary.

Regulatory developments in other jurisdictions, particularly Europe and the United Kingdom,³ are focused on greater oversight of online services' systems and processes, as opposed to the individual day-to-day experiences of their citizens. eSafety's model is unique in enabling us to remediate individual harms as well as to identify and seek to address the systemic challenges these complaints may reveal. The OSA strengthens our ability to do both, and particularly lifts our ability to address several systemic problems we have observed through our complaints schemes at-scale through the Basic Online Safety Expectations and the industry codes. As explained in the next section, we believe a combination of regulatory efforts across the pillars of prevention, protection and proactive and systemic change will continue to be necessary to remain at the vanguard of online safety policy.

³ UK Department for Digital, Culture, Media & Sport, *Draft Online Safety Bill*, CP 405, May 2021, <https://www.gov.uk/government/publications/draft-online-safety-bill>; European Commission, *The Digital Services Act: ensuring a safe and accountable online environment* https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en.

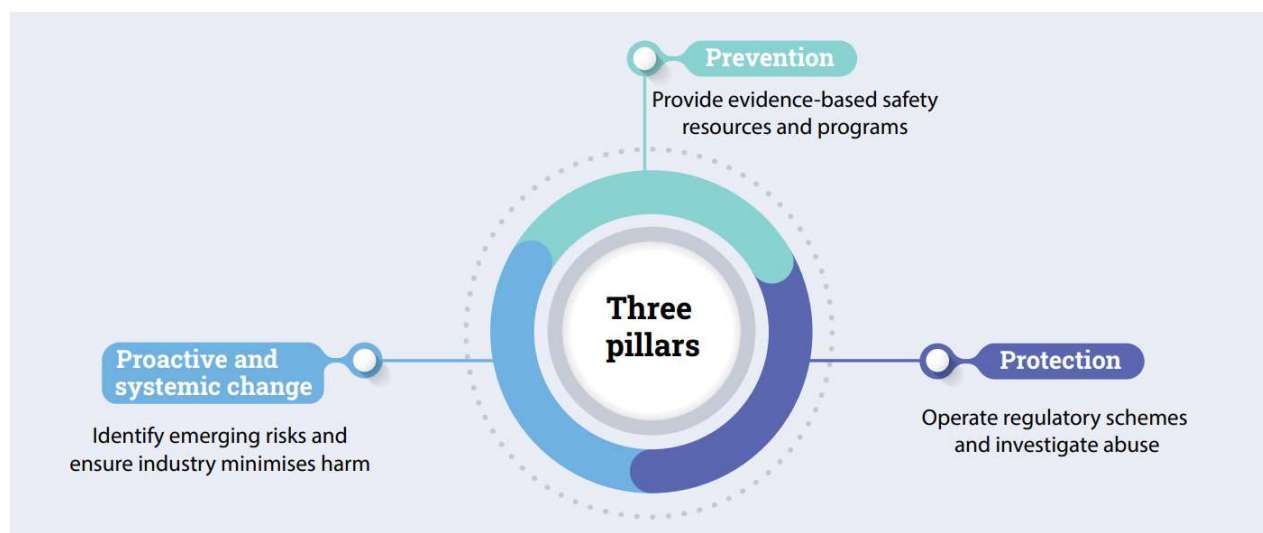
B. eSafety's approach to online safety

eSafety recognises that combating online harm is a global challenge and therefore we work as part of a cross-agency, cross-sector and multi-jurisdictional online safety ecosystem.

Our aim is to prevent and remedy harm, enhance transparency and accountability, and examine the effectiveness and impact of what services are doing to keep their users safer online.

We take a risk and harms-based approach to our work, which complements the role other agencies play in investigating and prosecuting crimes perpetrated online.

Our core mission, safeguarding Australians at risk of online harm, is the foundation that underpins our three pillars of purpose: prevention, protection and proactive and systemic change. These pillars reflect our broad and holistic legislative remit. Each of our functions serves a main purpose but is also connected to the others, supporting them and in turn receiving support.



Prevention

While our powers allow us to act as an important safety net for Australians online, our primary goal at eSafety is to prevent the online harms we see every day from happening in the first place.

We believe this a lifelong educational journey that should begin as early as possible.⁴ This is because of the 81% of Australian parents who say their child uses the internet, 42% of their children now have access to an internet-connected device by the age of 2 years – increasing to 94% by the age of 4 years.⁵ This education should continue well into the senior years,⁶ to make sure older Australians are supported to remain safely connected to their community and family at a time when digital forms of connection are becoming increasingly vital.

⁴ <https://www.esafety.gov.au/parents/children-under-5>.

⁵ eSafety (September 2019) Digital parenting Supervising pre-schoolers online <https://www.esafety.gov.au/sites/default/files/2019-09/digital-parenting-supervising%20preschoolers%20online.pdf>.

⁶ <https://www.esafety.gov.au/seniors>.

The inquiry's terms of reference include "the effectiveness and impact of industry measures to give parents the tools they need to make meaningful decisions to keep their children safe online".⁷ eSafety agrees that industry has an important role to play in equipping parents, carers and other adult supporters – as well as children and young people themselves – with the information and tools they need to stay safe online. As set out in section 2B, the industry codes will require online services to do much more on this front.

It is also important to make sure that online safety information is easily accessible from a variety of other sources and tailored to meet the needs of diverse groups. We strive to achieve this by conducting research⁸ and developing evidence-based education resources and community programs designed for specific audiences, giving people of all ages and backgrounds the right tools to protect themselves online. This includes:

- parents and carers as the front lines of defence, particularly in the early years⁹
- educators and schools¹⁰ to develop students' critical skills across the 4 Rs¹¹ (respect, resilience, responsibility and reasoning), to manage online safety incidents that may arise within the school community,¹² and to support best practice online safety education¹³
- domestic and family violence frontline workers to upskill people who support those experiencing technology-facilitated abuse, and¹⁴
- Specific diverse and vulnerable communities that our research shows are more likely to experience online harms.¹⁵

We are also heightening our efforts to reach young people in a way that resonates with them by developing a youth engagement strategy¹⁶ as well as a new Online Safety Youth Advisory Council.¹⁷

Over the last several years, we have started to see evidence of real change to behaviours and attitudes, with children and young people taking multiple actions and accessing a range of tools and tactics in response to negative experiences. For example, eSafety's 2021 survey of 3,600 young Australians aged 8-17 years-old found that 64 per cent of young people who have experienced negative online behaviour blocked or unfriended people who had bullied them online – a significant increase on 46 per cent of young people in 2017. The research also found that young people are increasingly reaching out to their parents and friends. Sixty-six per cent of young people

⁷ Parliament of Australia, Select Committee on Social Media and Online Safety Terms of Reference https://www.aph.gov.au/Parliamentary_Business/Committees/House/Social_Media_and_Online_Safety/SocialMediaandSafety/Terms_of_Reference.

⁸ <https://www.esafety.gov.au/research>.

⁹ <https://www.esafety.gov.au/parents>.

¹⁰ <https://www.esafety.gov.au/educators>.

¹¹ eSafety (10 November 2017) The 4 Rs of Online Safety <https://www.esafety.gov.au/newsroom/blogs/4-rs-online-safety>.

¹² eSafety Toolkit for Schools <https://www.esafety.gov.au/educators/toolkit-schools>.

¹³ eSafety Best Practice Framework for Online Safety Education <https://www.esafety.gov.au/educators/best-practice-framework>.

¹⁴ <https://www.esafety.gov.au/key-issues/domestic-family-violence/training-for-frontline-workers>.

¹⁵ <https://www.esafety.gov.au/diverse-groups>.

¹⁶ Young and Resilient Research Centre, Western Sydney University (2021) Office of the eSafety Commissioner Youth Engagement Strategy https://www.westernsydney.edu.au/young-and-resilient/projects/current_projects/office_of_the_esafety_commissioner_youth_engagement_strategy.

¹⁷ eSafety (15 December 2021) New advisory council to give young Australians a voice in online safety <https://www.esafety.gov.au/newsroom/media-releases/new-advisory-council-give-young-australians-voice-online-safety>.

who have experienced negative online behaviour told their parents (up from 55% in 2017) and 60% told their friends (up from 28% in 2017)¹⁸.

We are evaluating many of our programs, with very positive results. For example, phase 1 of an evaluation of the eSafety Women domestic and family violence training program (unpublished) found a large majority (87%) of respondents felt the eSafety training was useful and positively impacted the services they provided their clients. Around two-thirds of respondents reported changing how they work with their clients, allowing them to listen for a client's potential online safety needs, incorporate online safety questions into risk assessment and safety planning, and link clients to supporting resources. In addition, participants in the evaluation of the eSafety Commissioner's Teacher Professional Learning program completed on 30 November 2021 reported increased confidence in dealing with online safety issues. They also reported an understanding of tangible steps they could take to support students, including reporting to eSafety.

Protection

Where online harm does occur, eSafety offers tangible, rapid redress. We do this through accessible and trustworthy complaints mechanisms, which allow us to investigate and take action to remove certain types of content to reduce ongoing trauma and re-victimisation. This includes cyberbullying material targeting children, intimate images shared without consent and other forms of seriously harmful online content, such as child sexual exploitation or pro-terror material. As of 23 January 2022, it will also extend to adult cyber abuse. We provide more information about each of these schemes in section 2.

eSafety works with online service providers, including mid-tier and major social media companies, to get quick and positive outcomes for victims and survivors of online harm. We also provide broader support for complainants. This can include making referrals to law enforcement, mental health providers or legal services. It may also include providing practical tips on how to mitigate further harm to empower victims of online abuse to feel more resilient and in control.

Our overarching compliance and enforcement policy, as well as specific regulatory guidance for each of our schemes, are on our website.¹⁹

These schemes serve as an essential safety net for Australians experiencing harm and give eSafety insights for our systemic work.

Proactive and systemic change

With the rapid evolution of technology – and resulting threat environment – and the myriad ways humans have found to weaponise online platforms, we will never create a safer online world if we wait for things to go wrong and then react. Instead, mitigating emerging online risks involves a proactive and anticipatory mindset. If governments and online platforms are not thinking and being proactive about online harms, then we are accepting the *status quo*.

¹⁸ eSafety (4 November 2021) Australian young people learning to push back against online bullies <https://www.esafety.gov.au/newsroom/media-releases/australian-young-people-learning-push-back-against-online-bullies>.

¹⁹ eSafety Regulatory schemes <https://www.esafety.gov.au/about-us/who-we-are/regulatory-schemes>.

We are operating in an extremely complex and inter-linked technology ecosystem, which requires sophisticated systems thinking to address online harms at-scale. The OSA recognises there are eight different sections within the online industry that have interconnected dependencies. The internet involves a global ecosystem, with most of eSafety's regulatory targets based overseas and contending with many government organisations across the world that have some interest in or remit over the technology sector.

Therefore, proactive and systemic change is a pivotal pillar in helping drive eSafety's overall efforts – and this works in tandem with our pillars of prevention and protection.

In fact, it is the complaints we receive from individuals which enable eSafety to identify emerging trends and consider the extent to which there may be more systemic online safety problems at play and to formulate guidance on best practice. eSafety also engages in consultation and environmental and horizon scanning to understand evolving online threats so we can stay a step ahead of technology trends and challenges.

We have developed several position statements²⁰ outlining the safety and regulatory implications of on-the-cusp issues. These include the emergence of convincing deepfake technologies,²¹ the move to immersive virtual reality environments like the 'metaverse',²² and the growing interest in a more decentralised internet or Web 3.0.²³

The most critical efforts we are undertaking on this front involve working with tech companies to alter their design ethos from 'moving fast and breaking things' to anticipating, detecting and eliminating online threats before they occur through Safety by Design.²⁴ This will involve massive cultural change in the tech industry design ethos, concerted prioritisation and leadership from tech CEOs, and far more pre-emptive action taken by industry to harden technology platforms and services from known online harms and risks.

Research and consultation for our Safety by Design initiative began in 2018. It now includes:

- a set of principles and methodology that positions user safety as a fundamental design consideration²⁵
- interactive assessment tools for enterprise and start up technology companies²⁶
- resources for investors and financial entities²⁷
- engagement with the education sector to embed Safety by Design into curricula in Australian schools and universities, and around the world.²⁸

²⁰ <https://www.esafety.gov.au/industry/tech-trends-and-challenges>.

²¹ eSafety (11 May 2020) Deepfake trends and challenges – position statement <https://www.esafety.gov.au/industry/tech-trends-and-challenges/deepfakes>.

²² eSafety (10 December 2020) Immersive technologies- position statement <https://www.esafety.gov.au/industry/tech-trends-and-challenges/immersive-tech>.

²³ eSafety (29 July 2021) Decentralisation – position statement <https://www.esafety.gov.au/industry/tech-trends-and-challenges/decentralisation>.

²⁴ <https://www.esafety.gov.au/industry/safety-by-design>.

²⁵ <https://www.esafety.gov.au/industry/safety-by-design/principles-and-background>.

²⁶ <https://www.esafety.gov.au/industry/safety-by-design/assessment-tools>.

²⁷ <https://www.esafety.gov.au/industry/safety-by-design/investors>.

²⁸ <https://www.esafety.gov.au/industry/safety-by-design/tertiary-education>.

Our current efforts include encouraging adoption of Safety by Design by governments, industry and organisations globally. The principles were designed with industry, for industry, because if the technology sector doesn't start building in safety protections from the ground-up, we will never achieve the level of safety online we need and deserve.

The commencement of the OSA will allow us to build on this foundational Safety by Design work and move into the realm of formally regulating the systems and processes of online services through the co-regulatory industry codes (discussed in section 2B) and the Basic Online Safety Expectations (discussed in section 4D).

The way these elements inter-relate, and our regulatory priorities²⁹ for the current financial year, are depicted in Figure 1 below.

²⁹ eSafety (November 2021) Regulatory Posture and Regulatory Priorities 2021-22 <https://www.esafety.gov.au/sites/default/files/2021-11/OSA%20-%20Regulatory%20Posture%20and%20Priorities.pdf>.

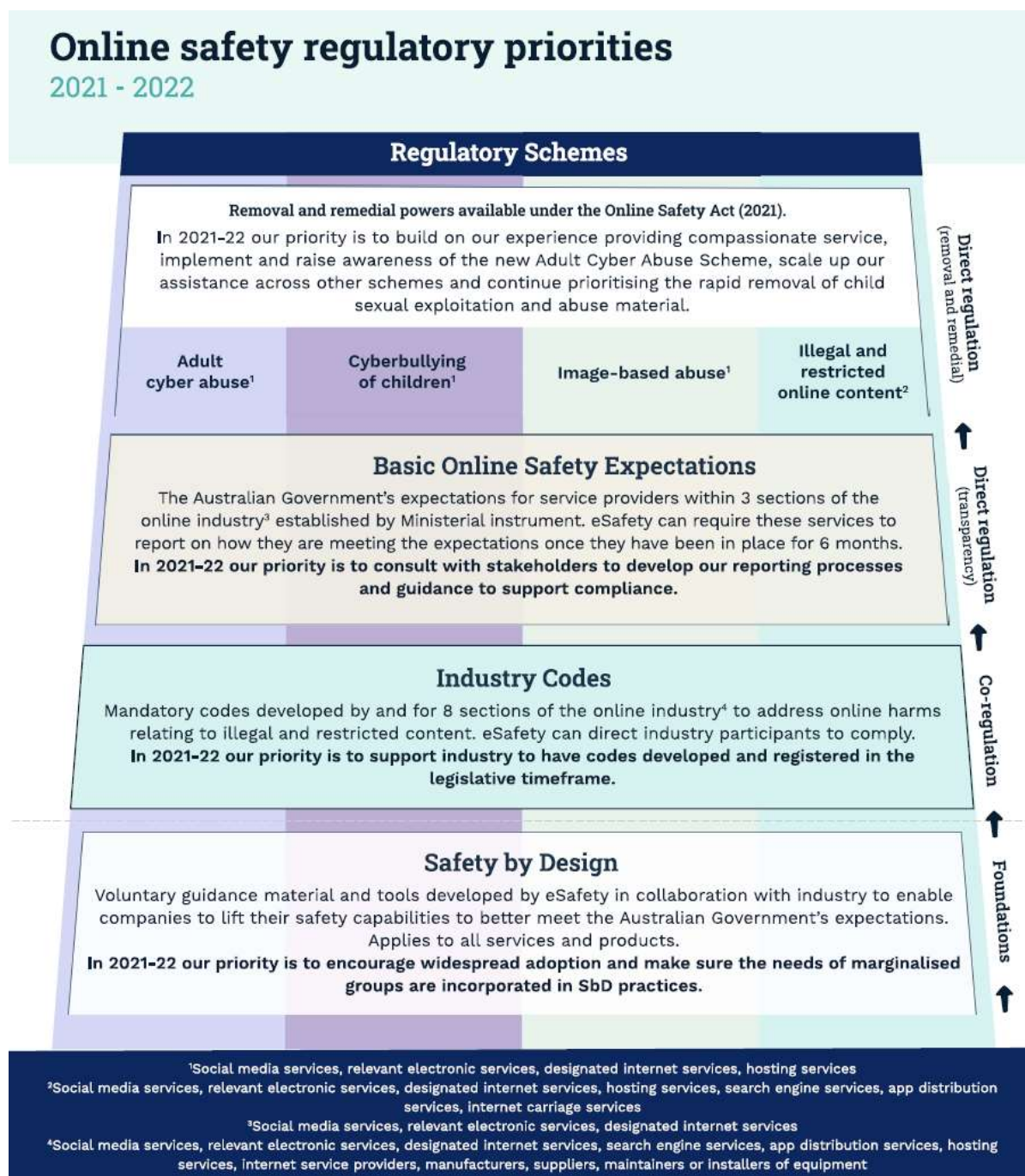


Figure 1

2. The range of online harms facing Australians on social media

Key Points

1. While there are many online risks and harms, it is important to remember the range of benefits the internet brings, and to take an approach which seeks to empower people to have safer, more positive online experiences.
2. There is no international consensus on how to define and categorise common types of harm, but eSafety has developed a typology based on emerging taxonomies, our research and experience. This typology seeks to frame online harms through a human rights lens wherever possible, emphasising the impacts on humans.
3. eSafety's remit includes several specific types of online harm, as defined in the OSA. These include cyberbullying of children, illegal and restricted online content ('class 1' material includes child sexual exploitation material and pro-terror content; 'class 2' material includes material that may be unsuitable for children, such as pornography), image-based abuse, adult cyber abuse, and material which promotes, incites, instructs in or depicts abhorrent violent conduct.
4. The Cyberbullying Scheme and the Image-Based Abuse Scheme have had high rates of success in removing content to remediate harm, and the OSA makes only minor changes to these schemes. We will keep a watching brief on the Online Content Scheme, which has been more substantially revised, and the Adult Cyber Abuse Scheme, which is new. As we begin to receive and investigate complaints under the Adult Cyber Abuse Scheme, we will evaluate the application of the legislative definition of adult cyber abuse, in line with concerns raised about the very high threshold.
5. Unlike the other schemes, the Basic Online Safety Expectations are not limited to the specific forms of harm defined under the OSA. This means, the Basic Online Safety Expectations give us the greatest opportunity to contend with a broader array of harms that may not meet the thresholds of these other schemes – see section 4D for details.

While today's internet has become a highly enabling environment for many forms of harmful content and conduct, it is also important to acknowledge that the online world is an amazing place which has brought myriad benefits.

Social media connects people with the world around them, as well as with their communities and their families. The stark and isolating nature of the pandemic has crystalised the need to access these channels. Forty-nine percent of Australians were either born overseas or have families

overseas, and the online world can help keep them connected.³⁰ Similarly, online connectivity remains critical for regional and remote communities, and for our Aboriginal and Torres Strait Islander communities who want to remain connected to country and culture. In addition, we know the benefits to neurodiverse young people from engaging online are evident both inside and outside of formal education. Being online helps them to develop social skills and offers ways to expand and enrich offline interests.

eSafety is committed to supporting and empowering all Australians to enjoy these benefits of digitally enabled lives. To do so, we need to combat the risks and harms they currently face online more effectively.

The following section sets out the range of online harms that fall within eSafety's legislated remit, which has expanded significantly since our office was formed. It also outlines the online harms typology developed for our Safety by Design initiative. This typology draws on our research and experience supporting those who have experienced online abuse and builds upon existing and emerging international approaches to understanding online harms.

A. Cyberbullying

eSafety was initially established as the Children's eSafety Commissioner to take a national leadership role in online safety for children, including through the implementation of a complaints scheme to address cyberbullying of children.³¹

Definition

Under our enabling legislation, cyberbullying material is online communication to or about an Australian child which is seriously humiliating, harassing, intimidating or threatening. This can include abusive texts and emails; hurtful messages, images or videos; excluding others; spreading nasty gossip and chat; or creating fake accounts to trick or humiliate someone.

Prevalence

Our research indicates that cyberbullying is relatively commonplace.³² In the 12 months to June 2017, one in five Australian young people reported being socially excluded, threatened or abused online. In addition, one in five Australian young people (15% of kids, 24% of teens) admitted to behaving in a negative way to a peer online – such as calling them names, deliberately excluding them or spreading lies or rumours. Of these, more than 90% had had a negative online experience themselves.³³

³⁰ Australian Bureau of Statistics (2017) Cultural Diversity in Australia, 2016
<https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2071.0~2016~Main%20Features~Cultural%20Diversity%20Article~60>.

³¹ Explanatory Memorandum, *Enhancing Online Safety for Children Bill* 2014
https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r5387_ems_d2990172-46c7-4b35-9eff-364175a7fbef/upload_pdf/399598.pdf;fileType=application%2Fpdf.

³² <https://www.esafety.gov.au/key-issues/cyberbullying>.

³³ eSafety (2018) State of Play – Youth, Kids and Digital Dangers <https://www.esafety.gov.au/sites/default/files/2019-10/State%20of%20Play%20-%20Youth%20kids%20and%20digital%20dangers.pdf>.

The current scheme

eSafety's Cyberbullying Scheme generally serves as a safety net for children and young people who have not been able to resolve their issue via the relevant social media service's reporting functions.

Since the start of our Cyberbullying Scheme in July 2015 to 31 December 2021, eSafety has helped over 3,600 children and their families with cyberbullying complaints.

Cyberbullying reports fall mainly into five categories:

- nasty comments / serious name calling (48%)
- offensive / upsetting pictures and/or videos (20%)
- fake accounts and impersonations (17%)
- threats of violence (13%), and
- unwanted contact (12%).³⁴

In many cases, cyberbullying is an extension of in-person bullying, particularly bullying that occurs at school. Where appropriate, eSafety works with schools to resolve complaints. We provide advice and emphasise the school should follow its cyberbullying policies. We also connect children and young people with counselling and support services, such as Kids Helpline.

While we have powers to send formal removal notices to people responsible for cyberbullying, we consider this option only in the most egregious cases, as cyberbullying is most often carried out by other children. eSafety is acutely aware of the psychological and emotional impact a notice could have on a child and considers the use of an end-user notice should be discretionary and proportionate in the circumstances.

Additionally, in most cases, we can get cyberbullying material removed most quickly and easily by reaching out to relevant social media services on an informal basis. We find that about 85% of the material is removed when we provide additional context that may help the service to understand the broader circumstances and impact on the victim.

In this way, the Cyberbullying Scheme largely works as a cooperative model between Government and industry, with eSafety encouraging social media services to remove cyberbullying material in the first instance in accordance with their own policies. This encourages responsiveness and action from the services themselves, rather than compelling this behaviour from them. eSafety can compel social media services to remove this material if the cooperative approach is not utilised.

Changes under the OSA

Under the OSA, the Cyberbullying Scheme will remain largely the same, with some enhancements to improve its reach and effectiveness. For example, whereas the previous scheme was limited to 14 specific social media services across two 'tiers' – one voluntary and one mandatory – the

³⁴ Complaints often involve more than one category, which is why the percentages add up to more than 100%.

enhanced scheme will apply to all social media services, as well as a range of other services where cyberbullying can occur, such as messaging and gaming services.

You can find more information in our [Cyberbullying Scheme Regulatory Guidance](#).³⁵

Using Online Defamation Action vs the Cyberbullying Scheme

eSafety noted some discussion through the Inquiry proceedings about how the proposed *Social Media (Anti-Trolling) Bill 2021* (Cth) may address online abuse targeting Australian children and young people, but no discussion of the Cyberbullying Scheme already under effective operation in this context. Based on our operational experience and engagement with young people through the scheme, they are often dealing with nasty comments, offensive images, fake accounts or threats. Where those behaviours meet the legislative threshold for cyberbullying, we would suggest this would be an effective way of addressing what is a common but damaging form of youth-based social discourse online and would provide an effective set of remedies. Of course, a family could choose to commence defamation proceedings against another child or their parents in some cases, but we would respectfully suggest this would be a much more taxing and expensive course of action and may have much broader impacts across the school community.

B. Illegal and Restricted online content

When the Children's eSafety Commissioner was initially formed in 2015, eSafety also assumed responsibility for the Online Content Scheme set out in Schedules 5 and 7 to the *Broadcasting Services Act 1992* (Cth), which was previously administered by the Australian Communications and Media Authority.

Scope

Among other things, the Online Content Scheme gives eSafety the power to regulate the hosting of prohibited content in Australia. Whether content is 'prohibited' is a decision made with reference to the National Classification Scheme guidelines. Material hosted in Australia that is classified Refused Classification (RC) or X18+ is prohibited and subject to takedown, while material classified R18+ is only prohibited if it is not placed behind a restricted access system that seeks to prevent children's exposure.

The current scheme

As a result of the strong regulatory framework in Australia, prohibited material is rarely hosted here. Accordingly, since 2015, eSafety has issued only a single takedown notice in relation to Australian-hosted prohibited material, where R18+ material was provided via an Australian-hosted adult website. However, we have joined global efforts to combat the forms of content causing the most serious harm.

While the National Classification Code and associated guidelines focus on content that may be offensive, more than harmful, eSafety applies a risk-and harms-based lens to its work, consistent

³⁵ eSafety (November 2021) Cyberbullying Scheme Regulatory Guidance <https://www.esafety.gov.au/sites/default/files/2021-11/OSA%20-%20Cyberbullying%20Regulatory%20Guidance%20V3.pdf>.

with its statutory remit. Accordingly, we prioritise the investigation of complaints about material that causes the greatest detriment to victim-survivors and society more broadly. This includes child sexual exploitation material, material that advocates committing a terrorist act and material that promotes, instructs or incites in matters of crime and violence.

Harm associated with online content can result from:

- the **production** of content – for example, where a perpetrator makes contact with a victim in an attempt to groom, coerce or force them into the production of content, or where coerced sexual activity or abuse is recorded
- the **distribution** of content – for example, where abusive material is posted, reshared or live-streamed online, which can compound the trauma experienced by victims harmed in the production of content
- the **consumption** of content – for example, where a person's behaviour, emotions, mental health, attitudes or perceptions are negatively impacted as a result of access or exposure to harmful content.

Over the more than 20 years of the Online Content Scheme's operation, complaints by the public have seen a steady increase. During the first full year of the Scheme's operation, 201 public reports were received about a variety of potentially prohibited content types. In the financial year 2020-2021, eSafety received more than 23,500 public reports. This was an increase of more than 60% on the previous financial year. Overwhelmingly, public reports concern child sexual exploitation material.

Global cooperation

Of the investigations we carry forward from these reports, 99% concern overseas-hosted child sexual exploitation material. eSafety facilitates the rapid removal of this material by notifying it to the International Association of Internet Hotlines (INHOPE) network for takedown within the host jurisdiction – typically within three business days.

INHOPE is a membership organisation consisting of 46 hotlines around the world working to combat online child sexual exploitation. The association works closely with domestic and international law enforcement to share intelligence, contribute to victim identification efforts and alleviate harm to victim-survivors who experience re-traumatisation because of the images of their abuse being circulated online.

The Australian Government has been a member of INHOPE for more than 20 years and is currently represented by eSafety. Being a member of INHOPE allows eSafety to overcome jurisdictional challenges and contribute to global efforts to eradicate child exploitation.

Changes under the OSA

Some of these jurisdictional challenges will also be alleviated through the OSA, which will modernise the Online Content Scheme and extend its reach – both across borders and to a broader array of service providers within the digital ecosystem.

The OSA creates two new classes of content linked to the National Classification Scheme. Class 1 material is content which has been or is likely to be classified RC, including child sexual exploitation material and pro-terror content. Class 2 material is content which has been or is likely to be classified X18+ or R18+, including real, non-violent sexual activity or anything that is unsuitable for a minor to see.

Whereas the current scheme centres around making sure service providers do not host prohibited content in Australia, the OSA will empower eSafety to give removal notices in respect of class 1 material regardless of whether the content is hosted in, or provided from, Australia or overseas. If a service ignores a class 1 removal notice, eSafety has additional regulatory options. These include seeking to have links that provide Australian access to the material removed from search engine results or removing apps from app stores.

In respect of class 2 material, eSafety can issue removal or remedial notices where the content is provided from, or hosted in, Australia. Remedial notices provide a more narrowly tailored option than removal, allowing the service provider to place material behind a restricted access system to prevent children's exposure.

You can find more information in our [Online Content Scheme Regulatory Guidance](#).³⁶

Industry codes

The OSA also sets out a regime for modernised industry codes or standards, expanded to additional sections of the online industry to make sure the whole digital ecosystem is playing its part. This includes social media services, messaging services, search engines, app stores, internet service providers, hosting service providers and manufacturers, suppliers and installers of equipment.

The codes will be mandatory, and eSafety will be empowered to direct industry participants to comply with the codes. Failure to comply with a direction may attract a civil penalty of 500 penalty units (up to \$111,000 for individuals and up to \$555,000 for companies).

The codes will require service providers to take more proactive steps to prevent and address the harms associated with class 1 and class 2 material. This could include, for example, providing information and tools to consumers so they are better equipped to protect themselves and their families from exposure to harmful content. For some industry sections, it will also include having appropriate systems and processes in place to detect and remove child sexual exploitation and pro-terror material, as well as to shield children from age-inappropriate content like online pornography.

In September 2021, eSafety released our Industry codes position paper to help guide industry bodies and associations to develop these codes.³⁷ This process is under way, led by a steering committee of industry associations.

³⁶ eSafety (December 2021) Online Content Scheme Regulatory Guidance <https://www.esafety.gov.au/sites/default/files/2021-12/eSafety-Online-Content-Scheme.pdf>.

³⁷ eSafety (September 2021) Industry Codes Position Paper <https://www.esafety.gov.au/about-us/consultation-cooperation/industry-codes-position-paper#:~:text=The%20new%20Online%20Safety%20Act,eSafety%20to%20register%20the%20codes.>

The paper sets out our preferred outcomes-based codes model, which includes three objectives:

- **Objective 1: Industry participants will take proactive steps to create and maintain a safe online environment**

Outcomes under this objective include proactively detecting, preventing and/or limiting relevant material; preventing Australian children from accessing age-inappropriate material; consulting and collaborating with other industry participants; and communicating and cooperating with eSafety. “Proactive” is an important qualifier here. Companies need to do more than they are now – accepting the status quo will not help lift safety across the ecosystem.

- **Objective 2: Industry participants will empower people to manage access and exposure to class 1 and class 2 material**

Outcomes under this objective include providing robust and effective reporting and complaints mechanisms; responding effectively to those complaints; and, relevantly to the inquiry’s terms of reference, providing tools and/or information to limit access and exposure to harmful material.

As highlighted in the paper, eSafety’s research reveals that the top online safety information Australian adults need is advice on where to report negative online incidents (45%), closely followed by how to use safety and privacy features on devices (43%).³⁸

- **Objective 3: Industry participants will strengthen transparency of, and accountability for, class 1 and class 2 material**

Outcomes under this objective include providing clear and accessible information about this type of material and publishing periodic reports about codes compliance.

The OSA allows eSafety to impose industry standards in certain circumstances, including where eSafety has made a request for code development and the draft code does not contain appropriate community safeguards.

You can find more information in our [Industry codes position paper](#).

C. Image-based abuse

In September 2018, a legislated scheme commenced to address image-based abuse, sometimes called ‘revenge porn’, enabling eSafety to investigate and take action in response to complaints about the sharing (or threatened sharing) of intimate images without consent.³⁹

Definition

Under the legislation, an intimate image is one that shows a person’s genital area or anal area (whether bare or covered by underwear); a person’s breasts (if the person identifies as female, transgender or intersex); private activity (for example, a person undressing, using the bathroom,

³⁸ eSafety (September 2020) Building Australian adults’ confidence and resilience online <https://www.esafety.gov.au/about-us/research/adults-confidence-and-resilience>.

³⁹ *Enhancing Online Safety (Non-consensual Sharing of Intimate Images) Bill 2018*
https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bld=s1113.

showering, bathing or engaged in sexual activity); or a person without attire of religious or cultural significance if they would normally wear such attire in public.

Prevalence

According to eSafety research commissioned in 2017, 11% of Australians aged 18 and over have experienced image-based abuse. While women aged 18 to 24 are more likely to be targets, image-based abuse affects people regardless of their age, race, religion, gender, sexual orientation, education or income.⁴⁰

The current scheme

In recognition of the seriously harmful nature of image-based abuse, eSafety has the power to give enforceable removal notices to service providers and end-users to facilitate the rapid removal of intimate images and offer victims relief. eSafety can also issue a remedial direction, which is a written communication that requires the recipient to take specific action aimed at preventing, or preventing further, non-consensual sharing of intimate images. eSafety can hold perpetrators of image-based abuse accountable – and help deter future abuse – using a range of enforcement actions which complement relevant state/territory and Commonwealth criminal offences.⁴¹ These include formal warnings, infringement notices, and seeking an injunction or civil penalty order from a court.

Between October 2017, when eSafety started to play a role in combatting image-based abuse, and 31 December 2021, we received 8,102 image-based abuse complaints concerning more than 13,500 URLs. Similar to the Cyberbullying Scheme, eSafety has been able to achieve removal in the vast majority (85%) of cases through informal engagement with the relevant service where the material is located.

We have also taken firmer, formal action to get important regulatory outcomes. This includes 11 removal notices (three issued to websites, seven issued to hosting services and one issued to the person responsible for posting the material); three remedial directions (requiring the recipient to take specific action aimed at preventing further non-consensual sharing of intimate images); and nine formal warnings (issued to people to deter them from sharing or threatening to share images).

Most complaints (about 75%) concern female victims, except in cases involving sexual extortion where a blackmailer targets a victim, generally for financial gain. Over half of all complaints received involve sexual extortion. Victims of sexual extortion who contact eSafety for help are predominantly male (about 67%).

About one quarter of complaints concern victims who were under 18 at the time the image or video was taken. Of these complaints from under 18s, only 8% concern peer-group sharing. Most relate to online child sexual exploitation where young complainants have been coerced into sharing naked images of themselves. Once a young person has sent an image to this type of offender, they receive threats to share their images and demands for more images. We have developed procedures which make sure eSafety is a safe place for children and young people to come for

⁴⁰ <https://www.esafety.gov.au/research/image-based-abuse>.

⁴¹ <https://www.esafety.gov.au/key-issues/image-based-abuse/legal-assistance/law-in-my-state-territory>.

help with these matters. The procedures align with our obligations to give relevant information to police, typically to the Australian Centre to Counter Child Exploitation.

When responding to complaints of image-based abuse, eSafety seeks to make sure the complainant is safe, supported and informed of our actions. Consistent with our Cyberbullying Scheme, we refer highly distressed victims to an appropriate counselling service, and if we are concerned the victim's personal safety is at risk, we help connect them with police or family and domestic violence support services.

Changes under the OSA

Under the OSA, the Image-based Abuse Scheme will remain largely the same, with only minor adjustments to bring it into alignment with the other schemes. This includes a reduced time period for online service providers to respond to formal removal notices – this will now be 24 hours across all the schemes, with discretion to extend the time where appropriate in the circumstances. The OSA also gives eSafety the discretion to release a public statement where a service provider repeatedly fails to deal with image-based abuse and the other forms of online harm falling within eSafety's complaints schemes, provided certain criteria are met.

You can find more information in our [Image-Based Abuse Scheme Regulatory Guidance](#).⁴²

D. Adult cyber abuse

In 2017, when the Children's eSafety Commissioner became the eSafety Commissioner – with responsibility for the online safety of Australians of all ages – eSafety also began accepting informal reports of adult cyber abuse.

The current situation

From June 2017 to 31 December 2021, eSafety has helped more than 4,700 adults with cyber abuse – 66% of whom are women. The most serious types of abuse targeting adults include impersonation accounts sending abusive messages to victims' contacts; technology-facilitated stalking (for example, through trackers, hacking, surveillance and keyboard monitoring); publishing private or identifying information about an individual with malicious intent to cause serious harm; and 'volumetric' attacks, also known as pile-ons or brigades.

To date, in the absence of a formal legislated scheme, eSafety has leveraged our relationships with social media services to advocate for the removal of cyber abuse material in the most serious cases involving vulnerable and highly distressed individuals.

Changes under the OSA

The OSA will establish a new Adult Cyber Abuse Scheme, giving eSafety the authority to require online service providers and people to remove cyber abuse that targets an Australian adult with the intention of causing serious harm.

⁴² eSafety (November 2021) Image-Based Abuse Scheme Regulatory Guidance <https://www.esafety.gov.au/sites/default/files/2021-11/OSA%20-%20IBA%20Scheme%20Regulatory%20Guidance.pdf>.

The threshold for what constitutes ‘cyber abuse’ under the legislation is high, with two parts.

1. The abuse must be intended to cause ‘serious harm’, which means serious physical harm or serious harm to a person’s mental health - like threats intended to cause serious psychological harm or serious distress that goes beyond ordinary fear
2. The abuse must also be menacing, harassing or offensive in all the circumstances.

Under the OSA, the term ‘adult cyber abuse’ is reserved for the most severely abusive material intended to cause serious psychological or physical harm. This would include material which sets out realistic threats, places people in real danger, is excessively malicious or is unrelenting.

Somebody finding something offensive or harassing is not enough to be adult cyber abuse. The content must also be intended to cause serious harm to that individual.

This threshold is significantly higher than the threshold for the Cyberbullying Scheme to reflect the expectation that adults are generally more resilient than children. The content in question must meet high thresholds consistent with those in the Criminal Code before we regard it as adult cyber abuse.

Examples likely to reach the threshold are publishing private or identifying information about an individual with malicious intent to cause serious harm; encouraging violence against a specific Australian adult based on their religion, race or sexuality; and threats of violence that make a person afraid they will suffer physical harm.

We recognise the threshold is high and there may be instances where we cannot take regulatory action. Every situation is unique and eSafety intends to help all Australians. Even if a matter does not meet the threshold, we will offer information and guidance to make sure that person feels supported and is aware of other options they might be able to take.

You can find more information in our [Adult Cyber Abuse Scheme Regulatory Guidance](#).⁴³

Online hate and volumetric attacks

eSafety notes that stakeholders have raised concerns through the course of the inquiry about a lack of effective remedies to combat online hate speech and ‘volumetric attacks’ or ‘pile-ons’ involving many people targeting a single person, often across multiple platforms.

eSafety is working to better understand and address online hate through a variety of research⁴⁴ and other initiatives⁴⁵ aimed at protecting voices at risk online. This includes education resources for young people, developing tailored resources for individuals from culturally and linguistically diverse backgrounds, and work to support Islamic women in dealing with image-based abuse that may involve a religious or cultural dimension. In addition, there will be an upcoming focus within our Safety by Design work on diverse, marginalised and at-risk groups to make sure their needs

⁴³ eSafety (December 2021) Adult Cyber Abuse Scheme Regulatory Guidance <https://www.esafety.gov.au/sites/default/files/2021-12/ACA%20Scheme%20Regulatory%20Guidance%20%20FINAL.pdf>.

⁴⁴ <https://www.esafety.gov.au/research/online-hate-speech>.

⁴⁵ <https://www.esafety.gov.au/sites/default/files/2020-12/Protecting%20voices%20at%20risk%20online.pdf>.

are effectively considered, incorporated and actioned in the design, development and deployment of online products and services.

Hate speech directed at groups (rather than an individual) will be beyond the scope of the new Adult Cyber Abuse scheme. However, hate speech targeting a specific person may meet the adult cyber abuse threshold. For example, when investigating adult cyber abuse, eSafety will consider whether menacing, harassing or offensive content was targeted at someone because of their racial or cultural background. eSafety will also consider whether posts are designed to generate volumetric attacks in determining whether serious harm is intended.

This will allow eSafety to consider hateful and dehumanising commentary and incitement to abuse where it involves the targeting of a particular Australian adult. Where the threshold is met, eSafety will be empowered to issue removal notices to relevant services or to people responsible for the abuse.

In addition, hate speech that promotes, incites or instructs in matters of crime or violence may fall within the Online Content Scheme (discussed above at section 2B).

Where complaints reveal systemic issues – such as a failure by services to enforce their terms of use (including terms of use relating to online hate or dehumanisation) or cooperate with other services (including to detect high-volume cross-platform attacks) – eSafety will have the power to drive services to do better through the Basic Online Safety Expectations, discussed below at section 4D.

As noted above, we will monitor the implementation of the OSA and evaluate the impacts of the Adult Cyber Abuse Scheme once it takes effect. We look forward to continuing discussions with community groups, academics, non-government organisations, and government bodies with responsibility for combatting discrimination and promoting human rights and equal opportunity to make sure we are working together to address these issues.

E. Abhorrent violent material and online crisis events

Following the horrific live-streamed terrorist attack in Christchurch, New Zealand in March 2019, the Australian Parliament passed the *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth). The abhorrent violent material (AVM) regime gives eSafety the power to issue a notice to any website publishing AVM and/or the service that hosts that website. The notices do not require the AVM to be removed. However, if a service is later prosecuted for failing to remove or cease hosting AVM, the notice can be used in legal proceedings to show recklessness regarding the AVM.

AVM definition

AVM is defined as audio and/or visual content produced by a perpetrator or accomplice of a terrorist act involving serious physical harm or death, murder or attempted murder, torture, rape or kidnapping involving violence. It does not include bystander coverage, and there are several exceptions for legitimate journalism, research, advocacy or artistic purposes.

The current scheme

Since April 2019, eSafety has issued 24 notices to a variety of content and hosting services in relation to 15 items of content depicting fatal terrorist violence, murder and torture. Content has been removed by a service in 93% (14 out of 15) of matters where eSafety issued an AVM notice.

Removing this material from online access prevents a range of social harms. These include the trauma and suffering of victims and their family members, the radicalisation of other potential perpetrators, and the use of AVM to threaten, harass or intimidate Australians or specific community groups.

The Parliamentary Joint Committee on Law Enforcement recently reviewed the operation and effectiveness of the *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth). The Committee found the legislation provides an effective and appropriate framework “for ensuring quick action is taken in relation to the AVM, both by deterring misconduct by industry and by providing regulators and law enforcement with important tools in their efforts to respond to cases where extremists or terrorists have sought to exploit online platforms to promote violence.”⁴⁶

Online crisis events

In addition to passing the AVM legislation, the Government created a Taskforce to Combat Terrorist and Extreme Violent Material Online (Taskforce). The Taskforce recommended that eSafety consider the temporary use of an existing power under subsection 581(2A) of the *Telecommunications Act 1997* (Cth) to direct ISPs to continue blocking domains known to be hosting the Christchurch attack footage and the attacker’s manifesto. It also recommended that eSafety work with Communications Alliance to develop a protocol for use of this blocking power in the circumstances of another ‘Online Crisis Event’ involving the viral and seriously harmful distribution of this type of content.

On 6 September 2019, eSafety issued the *Telecommunications (Protecting Australians from Terrorist or Violent Criminal Material) Direction (No. 1) 2019* (the Direction). The Direction formalised blocking action voluntarily taken by ISPs against websites providing access to the Christchurch material. This was a temporary, six-month direction put in place as an interim measure to stem the viral spread of the material and prevent it from causing harm to Australians, while the Government considered longer-term options for addressing misuse of online platforms by perpetrators of terrorism and violent extremism.

In December 2019, in conjunction with Government and industry and in line with the Taskforce’s recommendations, eSafety finalised a protocol governing ISP blocking in an online crisis event (the Protocol). The Protocol establishes detailed criteria, high thresholds and checks and balances to make sure eSafety’s powers are used only in very limited and very serious circumstances. Any blocking direction made under the Protocol would only be in place for a limited time, to be

⁴⁶ Parliamentary Joint Committee on Law Enforcement (14 December 2012) Media Release https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Law_Enforcement/AVMAct/Media_Releases; eSafety’s submission can be found at: https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Law_Enforcement/AVMAct/Submissions.

determined on a case-by-case basis. Following the initial blocking period, eSafety could take further action to address the relevant material, in consultation with the ISPs and affected websites.

Changes under the OSA

The OSA does not affect the AVM notice power within the Criminal Code. However, in line with the recommendations of the Taskforce, it does establish a more specific and targeted power for eSafety to direct ISPs to block certain domains containing terrorist or extreme violent material, for time-limited periods. This includes material that promotes, incites, instructs in or depicts violent kidnapping, rape, torture, murder, attempted murder or terrorist acts being disseminated online in a manner likely to cause significant harm to the Australian community. eSafety is working with Communications Alliance and ISPs to update the Protocol to reflect the new legislation.

You can find more information in our [Abhorrent Violent Conduct Powers Regulatory Guidance](#).⁴⁷

F. Online harms typologies

While the OSA sets the parameters for the specific types of online harm eSafety regulates through its complaints and removal schemes, there is no international consensus on how to define or categorise common types of harm.

Much of the research in relation to online risks has centred on children and young people, with several classification models and theories emerging.⁴⁸ Other approaches look to broadcasting regulatory frameworks and research into consumer experience of harm.

At this stage, it is largely up to individual online service providers to establish rules and guidelines for the type of activity and content that is or is not permitted on their platforms within community guidelines or terms of service. However, these can diverge significantly across services.

To support early-stage companies to understand the range of online harms that may occur on their services – as well as their impacts on victims, the different modes of abuse, and factors that may contribute to harm – eSafety developed an online harms typology as part of our Safety by Design assessment tools.⁴⁹

Given that our award-winning Safety by Design initiative focuses on people's rights, our typology seeks to frame online harms through a human rights lens wherever possible, emphasising the impacts on people. The categories are not exclusive and often overlap. They include, for example:

⁴⁷ eSafety (December 2021) Abhorrent Violent Conduct Scheme Regulatory Guidance <https://www.esafety.gov.au/sites/default/files/2021-12/OSA-AVCP-Regulatory-Guidance.pdf>.

⁴⁸ S Livingstone and M Stoilova, 'The 4Cs: Classifying Online Risk to Children' *CO:RE Short Report Series on Key Topics*, Hamburg: Leibniz-Institut für Medienforschung, Hans-Bredow-Institut (HBI), CO:RE - Children Online: Research and Evidence, 2021, [doi:10.21241/ssoar.71817](https://doi.org/10.21241/ssoar.71817); S Livingstone, G Mascheroni and E Staksrud (2015) Developing a framework for researching children's online risks and opportunities in Europe, EU Kids Online, https://www.researchgate.net/publication/285593314_Developing_a_framework_for_researching_childrens_online_risks_and_opportunities_in_Europe; Youth Protection Roundtable Toolkit (2009) www.kijkwijzer.nl/upload/download_pc/74_final_YPRT_Toolkit.pdf; B O'Neil, 'Research for CULT Committee – Child safety online: definition of the problem' European Parliament, Policy Department for Structural and Cohesion Policies, Brussels, 2018 [https://www.europarl.europa.eu/thinktank/pl/document/IPOL_IDA\(2018\)602016](https://www.europarl.europa.eu/thinktank/pl/document/IPOL_IDA(2018)602016); M Teimouri, S R Benrazavi, MD Griffiths, and M S Hassan, 'A model of online protection to reduce children's risk exposure: Empirical evidence from Asia'. *Sexuality & Culture*, 2018 22(4), 1205-1229, [doi:10.1007/s12119-018-9522-6](https://doi.org/10.1007/s12119-018-9522-6); UNICEF (December 2017) The State of the World's Children 2017: Children in a Digital World <https://www.unicef.org/reports/state-worlds-children-2017>.

⁴⁹ <https://sbd.esafety.gov.au/startup>.

- **Personal safety harms** – for example, direct and indirect threats or facilitation of violence; intimidation and harassment; viral challenges.
- **Health and wellbeing harms** – for example, self-harm and suicide material; material that promotes eating disorders; children’s exposure to developmentally inappropriate content.
- **Harms to dignity** – for example, insulting and demeaning comments; trolling⁵⁰ to provoke and disturb other users.
- **Privacy harms** – for example, doxing; sexual extortion; image-based abuse.
- **Harms involving discrimination** – for example, hate speech; racism; misogyny; sexual harassment; homophobia and transphobia.
- **Harms involving deception and manipulation** – for example, mis/disinformation; scams; catfishing; recruitment to extremism; grooming of children.

eSafety has conducted research on many of these harms and developed evidence-based resources tailored to specific audiences, as set out in sections 1B and 3A.

⁵⁰ We note that our definition of ‘trolling’ (see <https://www.esafety.gov.au/young-people/trolling>) differs from the concept of trolling as a synonym for defamation per the *Social Media (Anti-Trolling) Bill 2021* (Cth).

3. Evidence of the potential impacts of online harms on the mental health and wellbeing of Australians

Key Points

1. Online harms can be seriously damaging, especially for those most at-risk. It is important to continue building the evidence base on the intersectional factors that can contribute to greater harm and to provide tailored resources for those at risk. eSafety continues to do this through our research and Safety by Design work.
2. It is important we continue to research the impact of social media and online engagement on children and young people, as the current evidence base is mixed.
3. The internet offers many benefits to children and young people, including opportunities to learn, connect, be creative and seek help. It is important that interventions aimed at keeping children safe online are crafted to avoid the unintended consequence of reducing their ability to access these benefits.
4. An abstinence-based approach of cutting off young people from the internet is likely to have adverse consequences for the mental health and wellbeing of children and young people not from supportive homes with engaged parents. Moreover, it could shut vulnerable young people off from support services and affinity groups that could help them achieve a sense of understanding and belonging. Such blunt force approaches could also prevent young people who are excluded from developing the key skills they will need to navigate the online world safely as adults.
5. We need to take an approach which incorporates the voices of children and young people and seeks to raise their awareness to prevent harm; empower them to deal with issues effectively when they arise; and make sure industry is doing more to shoulder responsibility – as set out in further detail in section 6B.

Harmful online content and behaviour can be seriously damaging, especially for those most at-risk. The social, emotional, psychological and even physical impacts of online harms can be immediate, experienced over time and/or enduring. They can also be experienced both online and offline. Impacts can include:

- **Personal safety impacts** – fear of psychological violence, physical violence and murder.
- **Emotional and social impacts** – annoyance, anger, humiliation, shame, guilt, self-blame, deception, betrayal and/or fear.
- **Financial impacts** – ability to work and earn an income, loss of financial security, restricted access to or knowledge of personal finances.
- **Health and wellbeing impacts** – anxiety, aggression, depression, self-destructive behaviour, physical health problems, intimate relationship difficulties, re-victimisation, disassociation, loss

of self-esteem and confidence, withdrawal from social activities, lack of trust, substance abuse, ongoing trauma, self-harm and suicide.

eSafety research on adult's negative experiences online in 2020 found that, of the 67% who had a negative experience online, 40% reported a range of negative impacts as a result including mental or emotional stress (25%), financial loss (9%), relationship problems (6%) or reputational damage (6%).⁵¹

eSafety's research on negative impacts

eSafety's online hate speech research found that 58% of people who experienced hate speech reported a negative impact from their experience – typically mental or emotional stress, relationship problems or reputational damage.⁵²

eSafety's research with victims of image-based abuse (non-consensual sharing of intimate images) showed that 65% felt annoyed, 64% felt angry, 55% felt humiliated, 40% felt depressed and 32% felt afraid for their safety. It negatively affected the self-esteem of 42%, the mental health of 41% and the physical wellbeing of 33% of victims.⁵³

A. At-risk groups and intersectional factors

eSafety's research and reporting trends show certain individuals and groups are disproportionately at risk of online harm or face additional barriers to protecting themselves from harm or accessing support.⁵⁴

We recognise that many intersecting factors influence risk levels and individual experiences of online harm. We shape and prioritise our programs and resources to support, protect and build the capacity of those who are most at risk. Groups who are more likely to be subject to online abuse, compared to the rest of the Australian population, include:

- children and young people
- older Australians
- women
- people with disability
- Aboriginal and Torres Strait Islander peoples
- people from culturally and linguistically diverse communities, and
- people who identify as LGBTIQI+.⁵⁵

⁵¹ eSafety (August 2020) Adults' negative online experiences <https://www.esafety.gov.au/sites/default/files/2020-07/Adults%27%20negative%20online%20experiences.pdf>.

⁵² eSafety (2020) Online hate speech – Findings from Australia, New Zealand and Europe <https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf>.

⁵³ eSafety (October 2017) Image-Based Abuse National Survey: Summary Report <https://www.esafety.gov.au/sites/default/files/2019-07/Image-based-abuse-national-survey-summary-report-2017.pdf>.

⁵⁴ <https://www.esafety.gov.au/diverse-groups/protecting-voices-risk-online>.

⁵⁵ <https://www.esafety.gov.au/diverse-groups/protecting-voices-risk-online>.

Protecting voices at risk online

eSafety research⁵⁶ has found that:

Aboriginal and Torres Strait Islander peoples experience online hate speech at more than double (33%) the national average in Australia (14%).

Those who identify as LGBTIQ+ also receive elevated levels of targeted abuse, with 30% experiencing hate speech (compared to 14% for the rest of the population).

People from culturally and linguistically diverse backgrounds experience online hate speech at higher levels (18%) than the national average in Australia (14%).

For people with disability, abuse disproportionately tends to target their disability and/or physical appearance.

Women face disproportionate levels of online abuse that is sexualised and violent. Women and girls make about two-thirds of eSafety's cyberbullying, image-based abuse and informal adult cyber abuse complaints.

Women and their children experiencing domestic and family violence almost always experience technology-facilitated abuse designed to extend coercion and control over their lives.

Furthermore, the impact of negative experiences is more prevalent for these groups, which in turn increases the chances of them being the target of more serious online issues. For example, Aboriginal and Torres Strait Islanders and those identifying as LGBTIQ+ are more likely to be targets, with the majority (6 in 10) reporting adverse impacts as a result.⁵⁷

At eSafety we tailor our strategies, programs and resources to the needs of diverse people and communities.

eSafety recognises we need to remain innovative, agile and culturally responsive to meet the changing online safety needs of Australia's diverse and continually evolving population.

We have therefore adopted a multi-layered approach to continuously improving eSafety's ability to meet the needs of at-risk groups.

- We undertake whole-of-agency and individual training.
- We conduct in-depth research and keep up to date with key findings from other national and international research.
- We analyse trends identified through our reporting schemes and investigations.
- We test programs and resources in the field and modify them according to feedback.
- We consult target audiences and co-design resources with them.

⁵⁶ eSafety (2020) Online hate speech <https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf>

⁵⁷ eSafety (August 2020) Adults' negative online experiences <https://www.esafety.gov.au/sites/default/files/2020-07/Adults%27%20negative%20online%20experiences.pdf>.

- We share knowledge and best practice with other organisations in the online safety sector and those working with at-risk groups.
- We will be able to consider “intersectional characteristics” that may be driving targeted online abuse as factors in evaluating complaints around our cyberbullying and adult cyber abuse schemes.

B. Children and young people

While eSafety’s remit now includes promoting online safety for Australians of all ages, we maintain a strong focus on children and young people. We acknowledge and share the growing concern about their mental health and wellbeing.

Young people’s mental health and wellbeing

Young people have experienced more psychological distress than other age groups for some time. This has been exacerbated through the COVID-19 pandemic, with 74% of headspace (National Youth Mental Health Foundation, funded by the Australian Government Department of Health) clients reporting that their mental health was either a little (47%) or a lot (27%) worse since the pandemic.⁵⁸ While there is no evidence of an increase in suicide deaths among young people, young people are accessing services at greater rates for their self-harm and suicidality compared to before the pandemic and their suicidality may be more severe.

There are multiple reasons for increased distress among young people. This includes isolation, loneliness and increased uncertainty about the future. There are significant interdependencies between young people’s online and offline experiences.⁵⁹ Digital technologies can amplify or intensify both risk and protective factors, but the root causes of young people’s mental health difficulties tend to lie deeper and most likely stem from their (non-digital) personal circumstances.

The evidence is mixed

It is important to take a nuanced and balanced view of children’s and young people’s experiences online and avoid drawing causal lines where they are not supported by evidence. The evidence before us suggests the relationship between mental health issues and social media use is complex. In fact, some usage can be positive and beneficial to mental health and wellbeing, while other usage patterns and experiences can be harmful.⁶⁰

Most research exploring the intersection between social media and mental health notes there are mediating factors ranging from personality, underlying mental health issues, age, gender, socio-economic background, ethnicity, level of parental engagement, and a person’s level of self-regulation in social media use.

⁵⁸ Headspace (August 2020) Coping with COVID: the mental health impact on young people accessing headspace services <https://headspace.org.au/assets/Uploads/COVID-Client-Impact-Report-FINAL-11-8-20.pdf>.

⁵⁹ M Stoilova, C Edwards, K Kostyrka-Allchorne, S Livingstone and E Sonuga-Barke, (2021) Adolescents’ mental health vulnerabilities and the experience and impact of digital technologies: a multimethod pilot study. London School of Economics and Political Science and King’s College London, doi:10.18742/pub01-073.

⁶⁰ A Radovic, T Gmelin, B D Stein and E Miller, ‘Depressed adolescents’ positive and negative use of social media’, *Journal of Adolescence*, 2017, 55, 5-15 doi: [10.1016/j.adolescence.2016.12.002](https://doi.org/10.1016/j.adolescence.2016.12.002).
; M Stoilova et al, (2021) Adolescents’ mental health vulnerabilities and the experience and impact of digital technologies:

A study of more than 400,000 children by the Oxford Internet Institute found no consistent change over time in the relationship between technology and mental health. The study did suggest young people may be using technology more often as part of social support seeking and emotional coping processes.⁶¹

These findings are supported by a US longitudinal study of about 75,000 adolescents, which concluded there was no compelling evidence that social media use meaningfully increases adolescents' risk of depressive symptoms.⁶²

That said, for those with pre-existing mental health conditions, some negative online experiences and usage patterns can be harmful and negatively impact their symptoms.⁶³

Negative effects on wellbeing and the importance of resilience-building

The negative online experiences children and young people may face on social media are broad and varied. eSafety sees the lived experiences and mental health impacts of online harms through our regulatory complaints and investigation schemes, as well as through our research and training.

Our research shows that 45% of Australian young people have reported being treated in a hurtful or nasty way online. These experiences have led young people to feel sad, angry, socially isolated, helpless, have lower self-esteem and generally not feel good about themselves.⁶⁴

Other issues linked to young people's mental health on digital environments include time spent online, digital literacy and identifying misinformation, peer pressure (including body image), and unwanted or unsafe contact.

Factors that may generate negative mental health impacts from using social media include individual personality, underlying mental ill-health, age, gender, socio-economic background, ethnicity, level of parental engagement, and level of self-regulation in social media use.

Notably, eSafety research has also found that negative online experiences can also build resilience and protective mechanisms – with nine in ten teens motivated to engage in more positive online behaviours and build more inclusive online environments after negative experiences made them more aware of the impact of their actions. Additionally, many teens understand how to manage their exposure to negative experiences – with 54% blocking an account, 43% speaking to family/friends and 40% reporting it.⁶⁵

The evidence shows that the children and young people at risk in the 'real' world are also the ones most at risk online. Accordingly, we need to make sure we equip all children – and the adults who support them – with the skills they need to get the full benefits the online world has to offer them.

⁶¹ M Vuorre, A Orben, A K Przybylski, 'There Is No Evidence That Associations Between Adolescents' Digital Technology Engagement and Mental Health Problems Have Increased' *Clinical Psychological Science*, 2021, 9(5), 823-835. doi: [10.1177/2167702621994549](https://doi.org/10.1177/2167702621994549).

⁶² N Kreski, J Platt, C Rutherford, M Olsson, C Odgers, J Schulenberg and K Keyes, 'Social Media Use and Depressive Symptoms Among United States Adolescents' *Journal of Adolescent Health*, 2021, 68(3), 572-579 doi: [10.1016/j.jadohealth.2020.07.006](https://doi.org/10.1016/j.jadohealth.2020.07.006).

⁶³ J A Naslund, A Bondre, J Torous, 'Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice' *Journal of Technology in Behavioral Science*, 2020, 5, 245–257; doi: [10.1007/s41347-020-00134-x](https://doi.org/10.1007/s41347-020-00134-x); A Radovic, et al, 'Depressed adolescents' positive and negative use of social media'.

⁶⁴ eSafety, Youth Survey 2021, forthcoming release.

⁶⁵ eSafety (February 2021) The digital lives of Aussie teens <https://www.esafety.gov.au/research/digital-lives-aussie-teens>.

Positive effects on wellbeing, including help-seeking

There are also acknowledged benefits of social media for young people, including building and strengthening relationships, peer support and immediate relief from the emotional load from people with shared experiences; bolstering formal education (e.g. through forums, discussion boards, blogs or video tutorials) and informal education (e.g. through news or DIY videos); providing a safe place to find support and legitimisation for their identities (e.g. cultural, sexual, ethnic), and experiences (e.g. illness, disability); and allowing lonely young people to feel less shy by chatting online and feeling they belong to a group. eSafety notes that during the recent hearings by the Committee, expert witnesses also articulated these positive effects.

Numerous studies suggest social media has a positive effect on young people's mental health and wellbeing, including that social media and other online peer-to-peer connections can advance efforts to promote mental and physical wellbeing among people with serious mental illness.

As well as its broader social-emotional benefits, the online world provides crucial help-seeking avenues for those experiencing distress – including eSafety's own reporting schemes.

Youth mental health and support services have increasingly used social media platforms to raise awareness of their services and connect with young people through the medium they most actively participate in. By meeting children and young people where they are, mental health services are overcoming barriers to help-seeking and can provide self-help content young people can save for future reference and share with their peers.

What is powerful about social media platforms in this regard is they allow these services to have deeper engagement with young people. These services have told us that engaging with young people on social media helps them achieve their prevention and early intervention strategies.

Kids Helpline and Headspace offer online counselling services, acknowledging that young people prefer to initially engage with their services online – before transitioning to phone or in-person contact. Kids Helpline has told us that demand for webchat and email has significantly increased in the last five years. Last year, online support accounted for 44% of contact attempts to the service.⁶⁶ Following current trends, WebChat counselling could take over as the preferred contact type by 2023.⁶⁷

These services and others, like Reachout and the Butterfly Foundation, have built strong communities through online platforms (70k-170k followers each on Facebook alone). These platforms allow them to share advice, resources and events on their pages and passively through young peoples' news feeds. They also enable services to refer young people to crisis support through comment moderation.

Kids' Helpline, Headspace and Orygen also have their own online social networks and apps that connect people with similar mental health concerns, access peer-to-peer support, and participate

⁶⁶ Yourtown (2020) Kids Helpline Insights 2020: National Statistical Overview https://publications.yourtown.com.au/khl_insights-2020-statistical-summary-for-australia/.

⁶⁷ Yourtown (2020) Kids Helpline Insights into young people in Australia <https://www.yourtown.com.au/sites/default/files/document/Kids-Helpline-Insights-2020-Report-Final.pdf>.

in counsellor facilitated group work. These platforms help break down stigma and normalise help-seeking.

These examples show us that social media and online peer-to-peer connections can have a positive effect on people experiencing mental ill-health advancing efforts to promote mental and physical wellbeing.

How young people use social media

Young people use social media platforms differently to adults. For young people, these platforms are the infrastructure of everyday life. They take them for granted as the routine means to sustain relationships, express identities, and build networks.

On average, young people use four platforms simultaneously to carry on conversations. Young people also use social media to connect to retail, education, employment, like-minded communities and build their personal identity/brand. They turn to the internet (48.2%), mobile apps (25.8%) and social media (20%) for support on the important issues in their lives.⁶⁸

Adults tend to have a more segmented approach to social media – generally with their preferred social media platform for networking, getting news updates and watching DIY videos. Activities such as job hunting, or online shopping are likely to see adults go to specific websites.

When considering mental ill-health and social media, we need to understand these differences in use – particularly with young people relying on social media to connect with peers and their support community.

Many young people seem to curate their social media feeds to make sure they are being exposed to helpful and positive content. Young people who say social media usually makes them feel better report exposure to funny or inspirational content. Others describe being able to self-regulate their social media use during times they say they are feeling depressed, stressed, or anxious.⁶⁹ Behaviours such as self-curation, being an upstander rather than a bystander, using reporting and blocking tools, and the use of online services and communities to support their mental health and wellbeing, demonstrate a pro-active approach to mitigating negative experiences and mental ill-health.

Being reminded of this helps us to reflect on where and why young people experience negative behaviours and makes sure we apply youth-driven perspectives and solutions to generate positive online experiences and behaviours.

As highlighted in our submission to the consultation for the *Online Privacy Bill 2021* (Cth), interventions which may have the effect of excluding children and young people who are unable to either verify their age (if they are 16+) or their parental consent to participate (if they are under 16) risk eliminating the many benefits of online participation – including for connection to friends, family, culture, entertainment and support services.

⁶⁸ E Tiller, J Fildes, S Hall, V Hicking, N Greenland, D Liyanarachchi, and K Di Nicola (2020) Youth Survey Report 2020, Sydney, NSW: Mission Australia <https://www.missionaustralia.com.au/publications/youth-survey>.

⁶⁹ V Rideout, S Fox and Well Being Trust, 'Digital Health Practices, Social Media Use, and Mental Well-Being Among Teens and Young Adults in the U.S., Articles, Abstracts, and Reports. 2018, 1093. <https://digitalcommons.psjhealth.org/publications/1093>.

Youth consultation and engagement

In 2021, eSafety commissioned the Young and Resilient Research Centre at Western Sydney University to run a Living Lab process with children and young people to guide the development of eSafety's Engagement Strategy for Young People. The process used youth-centred, participatory co-research and co-design methods to explore young people's insights about online safety and to develop recommendations for eSafety's online safety messaging and resources, and our ongoing engagement with children and young people.

The research found that young people's main online safety concerns include interactions with others online (e.g., catfishing, fake accounts and contact from unknown people), privacy issues (exposure of personal information, photos, and stolen identities), and security issues (hackers, scams, and malware). Other key concerns include sexual exploitation (grooming, predators), accessing or being exposed to inappropriate content (pornography, violence), misinformation and fake news, commercial advertising (sexual or false advertising, sale of illegal or inappropriate goods), receiving judgment from peers about their opinions online, and the heightened vulnerability of particular groups (e.g., minorities) to a range of online safety issues.

Young people also told us they prefer to seek help from trusted adults in the first instance about an online safety issue. They will also seek professional counselling services, use reporting mechanisms on social media platforms, talk to peers, or consult online forums. However, there are many factors which prevent young people from seeking help. These include uninformed adults, adults not respecting boundaries, uncertainty about when help is needed, threats and blackmail, fear of punishment, stigmatisation, and victim blaming.

The research also found that young people prefer to get online safety messages through channels such as YouTube, Instagram, TikTok and Spotify, with interactive resources, personal stories, and with content and ads featuring other young people. They also want the practical skills to deal with negative online behaviours and mental health concerns, but they want information and advice delivered in a way that speaks to them. The report will be published by eSafety in January 2022.

Future directions

Orygen (the National Centre of Excellence in Youth Mental Health) has previously stated that social media use by young people is neither inherently good nor bad, but a balancing act unique to every young person and their needs and priorities.⁷⁰

We share this view at eSafety. We exist because negative behaviours such as online abuse, bullying, harassment and stalking happen, and have a disproportionate impact on children. But eSafety also exists to prevent these incidents occurring in future, by providing national leadership in online safety education and working with industry to make sure we consider safety at product inception and throughout the product lifecycle. We focus these efforts on achieving behavioural change, as human behaviour is at the core of these issues.

⁷⁰ Orygen (2019) Social media + youth mental health fact sheet <https://www.orygen.org.au/Training/Resources/digital-technology/Clinical-practice-points/Social-media-youth-mental-health/orygen-social-media-YMH-factsheet-2019?ext=>.

While we will continue to work on making platforms safer for people's mental wellbeing, we also need to educate young Australians on communicating respectfully, showing care, empathising with one another, and understanding the impact our words and actions have on others. We take the position that exposure to risk does not always equate to harm, but rather, it can also build resilience and critical thinking.

We also need to further educate young people and parents on identifying broader online harms which may compound the adverse effects of mental ill-health and empower them to make positive choices online that enhance their protection from negative experiences. For example, in 2021, eSafety delivered a parent webinar titled "Digital technologies and mental health", which outlined scenarios related to young people's experiences online, such as a friend sharing content about an eating disorder, self-harm, or suicidal thoughts. The content covered practical strategies for starting conversations and links to mental health providers like The Butterfly Foundation and headspace for further support and resources.

The online world is an extension of the offline world. Just as in the offline world, we cannot shield young people from all its potential ills, but we can better equip them to deal with pressures and tackle issues when things go wrong, and we can put in the safety levers to help them navigate digital environments safely. Active but conscious engagement on social media will be important for making sure young people build digital resilience to flourish online and be good digital citizens when they enter adulthood.

4. The transparency and accountability required of social media platforms and online technology companies regarding online harms experienced by Australians

Key Points

1. Transparency and accountability are essential to lifting the standard of online safety, but there is no consistent understanding of what it means to be transparent and accountable, and this can be interpreted in different (and selective) ways. eSafety's conception of transparency and accountability is set out in our Safety by Design assessment tools.
2. We are seeing improvements in industry transparency, but significant gaps remain. For example, many transparency reports focus on the removal of specific forms of harmful content, to the exclusion of information about broader safety efforts including proactive risk management practices, investment, innovation, cooperation, leadership, governance and incident response.
3. There are many cross-sector global efforts underway to promote greater transparency, largely in relation to specific issues, such as terrorist and violent extremist content. We welcome and engage in these efforts, but also note the risk of fragmentation if multiple competing frameworks are developed.
4. eSafety will seek to build on these efforts to generate greater transparency and accountability through the Basic Online Safety Expectations. These expectations are broader than our other schemes and allow us to require companies to report on a wide array of existing and emerging systemic issues.
5. The Basic Online Safety Expectations are not, themselves, enforceable (though reporting notices issued to companies are enforceable and subject to civil penalties). We note that other systemic regulation across the domains of privacy, security and competition, both domestically and globally, tend to have stronger enforcement options. eSafety will monitor the effectiveness of the expectations over time, and if required, will make recommendations to the Government to consider whether there is a need for additional penalties to make sure the online industry prioritises safety.

Transparency and accountability are hallmarks of a robust approach to online safety, and a thread of most global principles.⁷¹ They not only provide assurance that online services are operating according to their published safety objectives, but they also help to educate and empower people about steps they can take to address safety concerns.

⁷¹ For example, the G7 [Internet Safety Principles](#) – endorsed by Digital and Tech ministers; the UN [Convention on the Rights of the Child](#): General Comment No. 25 on child's rights in relation to the digital environment; the Five Country Ministerial, WePROTECT Global Alliance – [Voluntary Principles](#) to Counter Online Child Sexual Exploitation and Abuse; the International Telecommunications Union (ITU) – [Child Online Protection Guidelines](#); the Council of Europe – [Guide to Human Rights for Internet Users](#).

A. Safety by Design

Transparency and accountability are key principles of our Safety by Design work, which seeks to place the safety and rights of users at the centre of the design, development and deployment of online products and services.

The transparency and accountability module of our assessment tool highlights the importance of:

- Making information about safety policies, standards and processes publicly available;
- Establishing a dedicated section to house safety information, updating users on safety measures or policies and making the information accessible and inclusive;
- Building in measures to review the accuracy, impact and effectiveness of safety features and incorporating the results into key performance indicators or objectives;
- Undertaking internal reviews both routinely and after safety incidents occur;
- Having safety policies and standards independently reviewed and evaluated, and seeking to include a broad range of views and voices;
- Hosting forums for people outside the organisation to raise human rights concerns and ethical or safety issues; and
- Publishing regular user safety and transparency reports.

The module also includes a series of videos highlighting leading transparency and accountability practices by Snap (making safety policies and standards publicly available), TikTok (safety-related information and reporting), Google (engaging with independent experts) and Facebook, Instagram, Roblox, Tumblr and Twitch (engaging with a wide base of users and external stakeholders).

B. International and domestic efforts to enhance transparency

In addition to promoting transparency through the uptake of Safety by Design, eSafety is involved in several international collaborative efforts which seek to enhance transparency in the technology sector around particular threats to online safety. For example, we and the Department of Home Affairs have contributed to the OECD's multi-stakeholder project to develop a Voluntary Transparency Reporting Protocol (VTRP) in relation to terrorist and violent extremist content on their services.

We have also engaged with cross-sector transparency working groups led by the Global Internet Forum to Counter Terrorism, as well as the Center for International Governance Innovation's Global Platform Governance Network.

In addition, the eSafety Commissioner serves on the Board of the WeProtect Global Alliance, which has developed a Global Strategic Response to Online Child Sexual Exploitation and Abuse

calling on companies to publish regular transparency reports on the detection and removal of child sexual abuse material.⁷²

We are aware of numerous other worldwide efforts underway by civil society organisations, governments and industry to promote the uptake and consistency of transparency reporting.⁷³ Transparency requirements for service providers feature as key elements of the German NetzDG law, UK *Online Safety Bill*, EU *Digital Services Act* and in Canada's discussion paper that outlines its government's proposed approach to regulating social media and combating harmful content online.

Of note, a recent Joint Committee report scrutinising the UK *Online Safety Bill* recommended strengthening transparency requirements for service providers in the next iteration of the UK Bill, particularly around system design and automated content recommendation.⁷⁴ The Joint Committee contends that this will ensure the regulator and researchers can see what the platforms are doing, assess the impact it has and, in the case of users, make informed decisions about how they use platforms.

There are also several recent developments within Australia which involve reporting by technology companies. These include recommendations of the Taskforce to Combat Terrorist and Extreme Violent Material Online, as well as the Australian Communications and Media Authority's reporting and monitoring framework in relation to misinformation on digital platforms.

C. Gaps in transparency efforts

While we commend these international and domestic endeavours, we note that the proliferation of transparency initiatives presents the risk of a piecemeal approach. Few of these initiatives apply to the full range of online harms, and some have a narrow focus on particular issues. There also tends to be an emphasis on *content* risks, to the exclusion of *conduct* and *contact* risks—as well as a priority placed on detection and moderation of that content, over issues such as proactive risk management practices, investment, innovation, cooperation, leadership, governance and incident response.

Moreover, there is typically a spotlight on *removal* (and therefore an associated concern about censorship), which means that other interventions or tools to prevent or mitigate online harm tend to be overlooked. This, in turn, can lead to greater scrutiny of services which provide access to content on a large scale, while other types of services may go unnoticed, despite the serious safety risks they may present. It can also result in a conception of serious harm that is based on prevalence (with transparency metrics focused on the number of times content was viewed or reported) but which fails to consider the serious harm an individual might experience through

⁷² We Protect Global Alliance, Global Strategic Response to Online Child Sexual Exploitation and Abuse <https://www.weprotect.org/wp-content/uploads/WeProtectGA-Global-Strategic-Response-EN.pdf>.

⁷³ See eg <https://www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/>; <https://santaclaraprinciples.org/>; <https://inetco.org/mission>; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/708873/Government_Response_to_the_Internet_Safety_Strategy_Green_Paper_-_Final.pdf; <https://germanlawarchive.iuscomp.org/?p=1245>; <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>.

⁷⁴ Joint Committee on the Draft Online Safety Bill (14 December 2021) Draft Online Safety Bill Report of Session 2021-22 <https://publications.parliament.uk/pa/jt5802/jtselect/jtonlinesafety/129/129.pdf>.

abuse on a smaller scale. For example, cyberbullying or image-based abuse which is seen by few but may have a serious impact on the individual if it is not surfaced and addressed quickly and effectively.

The last several years have seen significant improvement in the extent and quality of transparency reporting in the online industry, as more services begin to develop transparency centres⁷⁵ and release reports on a routine basis.⁷⁶ eSafety welcomes these efforts. However, we would also like to see companies move beyond the current range of metrics offering 'selective' transparency and towards more 'radical' transparency into their individual and collective efforts to prevent and address the full range of online harms. We acknowledge there are legitimate concerns about the capacity of smaller companies to meet reporting requirements, the potential burden of inconsistent or duplicative reporting requirements across multiple jurisdictions, and the difficulty of developing metrics that can allow for consistency and comparison across different services. However, we do not believe these challenges are insurmountable.

D. Basic Online Safety Expectations

Developing a broad set of Basic Online Safety Expectations (expectations) under the OSA gives us an opportunity to explore and address some of these gaps, driving greater transparency and accountability across the online industry on a wide range of existing and emerging online safety issues.

Background

The expectations are a key element of the new Online Safety Act, and a vital, innovative and world-leading tool to add to our regulatory arsenal. They will articulate the Government's robust expectations for social media services, messaging services, gaming services and other apps and sites accessible from Australia, with a focus on making sure these services take reasonable steps to keep Australians safe.

The Minister for Communications, Urban Infrastructure, Cities and the Arts (Minister) will establish the expectations through a legislative instrument called a determination. The Department of Infrastructure, Transport, Regional Development and Communications held a public consultation on a draft determination from October to November 2021. Following consultation, the Minister will consider submissions and make a final determination. It is expected to come into effect in early 2022 at the same time as the OSA commences.

eSafety will have the power to require services to report on how they are meeting any or all of the expectations. The obligation to respond to a reporting requirement is enforceable and backed by civil penalties and other enforcement mechanisms. eSafety can also publish statements about the extent to which services are meeting the expectations. We believe this has great potential to help

⁷⁵ See, eg, Meta (<https://transparency.fb.com/en-gb/>), TikTok (<https://www.tiktok.com/transparency/en-us/>) and Twitter (<https://transparency.twitter.com/>).

⁷⁶ See eg New America's transparency report tracking tool (<https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/>) and the OECD's update on transparency reporting on terrorist and violent extremist content online (<https://www.oecd.org/digital/transparency-reporting-on-terrorist-and-violent-extremist-content-online-8af4ab29-en.htm>).

improve safety standards, and to bring greater accountability to services whose transparency to date has been highly selective and uneven.

Core expectations in the OSA

The Act specifies nine core expectations that must be included in the BOSE determination to create safer online environments for Australians:

Expectations regarding safe use

The provider of a service will:

1. Take reasonable steps to make sure people can use the service in a safe manner,
2. Consult the Commissioner in determining what are such reasonable steps.

Expectations regarding certain material and activity

The provider of a service will take reasonable steps to:

3. Minimise the extent to which the following material is provided on the service:
 - Cyberbullying material targeted at an Australian child
 - Cyber abuse material targeted at an Australian adult
 - A non-consensual intimate image of a person
 - Class 1 material
 - Material that promotes abhorrent violent conduct
 - Material that incites abhorrent violent conduct
 - Material that instructs in abhorrent violent conduct
 - Material that depicts abhorrent violent conduct, and
4. Make sure technological or other measures are in effect to prevent access by children to class 2 material provided on the service.

Expectations regarding reports and complaints

The provider of a service will make sure the service has clear and readily identifiable mechanisms that enable people to report, and make complaints about

5. Any of the following material provided on the service:
 - Cyberbullying material targeted at an Australian child
 - Cyber abuse material targeted at an Australian adult
 - A non-consensual intimate image of a person
 - Class 1 material
 - Class 2 material
 - Material that promotes abhorrent violent conduct

- Material that incites abhorrent violent conduct
- Material that instructs in abhorrent violent conduct
- Material that depicts abhorrent violent conduct; and

6. Breaches of the service's terms of use.

Expectations regarding dealings with the Commissioner

The expectation that the provider of a service will comply within 30 days if the Commissioner, by written notice, requests the provider to give the Commissioner:

7. a statement that sets out the number of complaints made to the provider during a specified period (not less than 6 months) about breaches of the service's terms of use;
8. a statement that sets out, for each removal notice given to the provider during a specified period (not less than 6 months), how long it took the provider to comply with the removal notice; or
9. specified information that explains what the provider does to make sure people can use its service safely.

Additional expectations in the draft determination

The draft BOSE determination released for public consultation includes 11 additional expectations (which could be subject to change through the consultation process) that service providers will:

1. Proactively minimise unlawful or harmful material or activity on the service.
2. Have regard to eSafety guidance in determining reasonable steps to ensure safe use.
3. Develop and implement processes to detect and address unlawful or harmful material or activity on encrypted services.
4. Prevent anonymous accounts from being used for unlawful or harmful material or activity.
5. Consult and cooperate with other services to promote safety.
6. Have (a) terms of use; (b) safety policies and procedures; and (c) policies and procedures to deal with user reports; and (d) standards of conduct.
7. Make information on how to make a complaint to the Commissioner accessible to users.
8. Ensure that information is (a) readily accessible to users; (b) accessible at all points in the user experience; (c) regularly reviewed and updated; and (d) written in plain language.
9. Regularly remind users of information and provide updates when there are changes.
10. Keep records of user reports for 5 years.
11. Designate a contact person to eSafety for purposes of the Act.

Reasonable steps

The draft determination includes examples of reasonable steps online services may take to meet relevant expectations. These are intended to provide services with guidance and to signal some of the matters on which eSafety may ask them to report.

Examples from the draft determination include:

- undertaking assessments of safety risks and impacts, and implementing safety review processes, throughout the design, development and deployment of the service
- making sure the default privacy and safety settings of services targeted at or used by children are robust and set to the most restrictive level
- working with other service providers to detect high volume, cross-platform attacks (also known as volumetric or 'pile-on' attacks)
- having processes that require verification of identity or ownership of accounts, and
- implementing age assurance mechanisms.

The reasonable steps provided as examples in the determination are not mandatory, and services may choose to undertake different steps. Services should be prepared to report on these steps, why they are reasonable, and how they are effective at meeting the relevant expectation(s) and keeping people safe.

Services are expected to consult with eSafety and refer to any guidance published by eSafety in deciding which reasonable steps are most suitable.

Once the determination is finalised, eSafety will begin producing guidance on the expectations and reasonable steps to meet them, in close consultation with stakeholders. Where there is overlap between the expectations and other eSafety workstreams – such as the industry codes and the age verification roadmap (explained in section 6B) – we will work to ensure alignment and consistency across the different elements and to cross-pollinate learnings from different engagement processes.

Reporting

There are three different ways eSafety will be able to seek information from services regarding compliance with the expectations.

1. eSafety may request information about terms of use complaints, the timeframe for responding to removal notices, or measures taken to make sure people can use the service in a safe manner. Failure to comply would give the Commissioner discretion to prepare a statement.
2. eSafety may give a reporting notice to a service provider requiring them to produce a report about their compliance with any or all of the expectations. These notices are enforceable, backed by civil penalties and other enforcement mechanisms, and can require non-periodic (one-off) reporting or periodic reporting over a specified timeframe of six to 24 months. In deciding whether to give such a notice, eSafety must consider several factors, including the number of complaints it has received under the OSA in relation to the service in the previous

12 months, any deficiencies in the provider's safety practices or terms of use, and any previous contraventions of civil penalty provisions relating to the expectations.

3. eSafety may make a legislative instrument requiring periodic or non-periodic reporting for a specified class of services. Like the reporting notices, these determinations are enforceable and backed by civil penalties and other enforcement mechanisms for failure to report.

In line with the minimum six-month reporting period established under the Act, eSafety will not require reporting until the expectations have been in effect for at least six months, though we may request reporting if a serious issue emerges during that time.

In the interim, our focus is on raising awareness of the expectations among the service providers, as well as consulting with stakeholders to develop regulatory guidance and to build their capacity to comply. We are also encouraging services to use our Safety by Design resources, particularly the self-assessment tools, to help them audit and improve their current safety practices and position themselves to meet the expectations.

For several years, eSafety has been discussing with industry some of the common safety problems and challenges we have identified across their services through our reporting trends and investigative insights. During 2022, eSafety will examine our evidence base through the lens of the expectations, deciding the key data points and information we may request or require from different services to assess how well they are meeting the expectations. We will also develop systems and processes to request or require reporting from services, with an emphasis on transparency and accountability for both industry and ourselves.

Through provider reporting, we hope to get a better sense of what services are (or are not) doing to keep people safe, acknowledge and highlight good practices, identify the gaps, and discern how they are measuring the effectiveness of their safety interventions. This will help us gauge the relative safety of specific services, as well as the broader online safety environment, and devise recommendations for continuous improvement.

Comparative approaches

Unlike many of the current and proposed regulatory tools at the disposal of other domestic and international regulators, elements of the expectations are not enforceable. We believe it will be important to monitor the implementation of the expectations and industry reporting over time to observe whether these provisions need to be stronger.

Overseas, the UK and the EU are both introducing heavier penalty schemes than Australia to regulate the systems and processes of online services. Mandatory risk assessments will accompany greater transparency requirements, followed by technical audits and significant fines for non-compliance, in the order of six to ten per cent of annual income.⁷⁷

⁷⁷ UK Department for Digital, Culture, Media & Sport, Draft Online Safety Bill, CP 405, May 2021, Clause 73 <https://www.gov.uk/government/publications/draft-online-safety-bill>.

Closer to home, the Australian Competition and Consumer Commission (ACCC) can apply significant financial penalties to regulate competition issues relating to big tech.⁷⁸ The *Online Privacy Bill 2021* (Cth) proposes to increase penalties available for the Office of the Australian Information Commissioner (OAIC), in line with those available for competition and consumer law.⁷⁹

If companies are to take safety seriously, it is important the incentives for investing and the penalties for failing to do so are commensurate across the realms of safety, privacy, security and competition.

⁷⁸ Australian Competition and Consumer Commission, Fines and Penalties <https://www.accc.gov.au/business/business-rights-protections/fines-penalties>.

⁷⁹ Attorney-General's Department, Online Privacy Bill Exposure Draft <https://consultations.ag.gov.au/rights-and-protections/online-privacy-bill-exposure-draft/>; Office of the Australian Information Commissioner (25 October 2021) Higher penalties to help protect Australians' privacy <https://www.oaic.gov.au/updates/news-and-media/higher-penalties-to-help-protect-australians-privacy>.

5. Evidence of the extent to which algorithms used by social media platforms permit, increase or reduce online harms to Australians

Key Points

1. The online world runs on algorithms. As basic computing instructions, they are often considered inherently neutral, with the human element of their design a key factor that can lead to further complexity and deeper ethical considerations.
2. Online services can use algorithms to reduce online harms (for example, by helping them to detect and filter out seriously harmful content at-scale), but algorithms can also create or contribute to a variety of harms. This can result from intentional efforts to make services ‘sticky’ so people stay engaged, or it may be an unintended consequence of human input – for example, where a curious click leads a person to increasingly extreme content.
3. The Basic Online Safety Expectations will allow eSafety to drive greater algorithmic transparency and accountability by enabling us to require reporting on how algorithms are mitigating or contributing to online harms. It is important to understand that “breaking the black box” should not necessarily be the goal – reducing the harmful outcomes algorithms may engender on online services should be the goal.
4. Algorithmic transparency and regulation are also being considered internationally. eSafety will continue to keep a watching brief on these developments and to engage in cross-jurisdictional and multi-disciplinary dialogue.
5. There are many complexities to auditing algorithms, and these – along with associated resourcing implications – would need to be carefully considered before introducing reforms.

In the context of social media, an algorithm is a coded sequence of instructions that prioritises what content a user will see. Simply put, algorithms are everywhere, and their use is only growing. The human aspect behind their design, and how they are coded and trained, can introduce complexity that can lead to both intentional and unintentional harms by their creators.

Algorithms are helping drive the digital economy by creating speed and efficiency through their ability to collate and disseminate large volumes of data and provide tailored user preferences which can help get services into the hands of those who need them more quickly. For example, a social media ad or post for low-cost child-care services targeted to a user identified as a parent could help provide a necessary and useful outcome for the parent and increase business revenue for a vital community service.

Algorithms can also help identify areas of need within a community that might inform future decision making such as identifying optimal areas for increased government funding to service a particular need, or the best location for a new business.

While these could be beneficial uses of algorithms as useful automations for aggregating factual community-based data, they may also end up demonstrating harmful biases because of human-coded inputs that may use indicators such as race, shopping preferences or political views to come to such conclusions. Such examples demonstrate the nuance of algorithms which can be developed with the best intentions but, without careful consideration, transparency and regular review of their outcomes can result in negative impacts.

Multiple algorithms, or machine learning instructions, may be active within a platform or service at any given time, all completing different tasks with different outcomes. As such, there is no single, fixed approach to considering the benefits and harms of an algorithm. To best understand those harms and the factors that contribute to them, it is essential to talk to the companies and engineers that develop the algorithms to understand how they have been designed, their policies and the intent behind the code.

Generally, discussion of ‘algorithms’ in the context of social media and other online services focuses on two areas:

1. content moderation algorithms, which seek to detect and take action to address content which may be harmful or problematic, and
2. content curation algorithms, which serve to recommend or amplify content which users may wish to see.

A. Moderation algorithms

Moderation algorithms are essential to enable services to moderate content at scale. They serve to reduce online harms by enabling services to proactively identify and filter out or remove seriously harmful content without having to wait for a user to report it or a human to review it. This can occur in the form of pre-moderation, where content is categorised and prevented from being visible to users before being posted; or post-moderation, where harmful content is flagged for review after it is already made available to the public.

Video or image content can be moderated through techniques such as image classification and object detection algorithms as well as image processing algorithms that can identify various regions inside the image and then categorise them based upon specific criteria. Text content can also be moderated through natural language processing algorithms that can conduct sentiment analysis and detect meaning to identify potential harms.

While moderation algorithms seek to reduce online harms, they can also present some challenges. Sometimes algorithms can get things wrong, and either fail to detect and address harmful content (for example, there are limitations in the ability of AI to identify harmful audio and video) or end up actioning content that should be permitted. For example, studies have found that LGBTQ+ sexual expression may be disproportionately flagged by moderation algorithms and taken down or

restricted.⁸⁰ Similarly, sensitive content warnings have been placed over photos depicting people with disability or physical differences, contributing to a sense of marginalisation.⁸¹

B. Recommendation algorithms

Curation algorithms sort through huge amounts of data to present content relevant to users. They help people discover new artists, friends, products, activities and ideas, and they help businesses and creators reach new audiences. Many people like having their online experience tailored to their preferences.

However, poor application of parameters within algorithms can cause issues – not just for user safety, but across a variety of domains.⁸² Concerns about curation algorithms include the potential for discrimination (for example, where recommendations are based on race or gender); the usage of personal data to produce recommendations in a way that may infringe on peoples' privacy; competition issues (for example, where certain products or services are given unfair preference); and the promotion of mis- or disinformation.

These algorithms may also connect users in inappropriate or dangerous ways. For example, this can happen by urging children to befriend adults, or by drawing people into 'click bait' or 'rabbit holes' of increasingly problematic content. In some cases, this may be intentional to produce 'stickiness' on the platform and keep people interested. In other cases, this may be an unintentional consequence of user input. Algorithms programmed to show more of the same content based on a mere view may be particularly susceptible, where a one-time curious click on a video may set off a flood of similar, or more extreme, content that could harm a vulnerable person.

Making the algorithm less sensitive to such inadvertent user behaviours, or to create an act of consent or confirmation of interest with the person before content is automatically generated into their feed, could be a useful mitigation strategy and an act of transparency that still preserves the objective of the platform to customise content to users' preferences.

Different inputs for recommendation algorithms can lead to different effects – good and bad. Prioritising time spent with a post or reactions to a post may result in people seeing more valuable information, or it may lead to an increase in inflammatory material being circulated. Promoting posts from family members and friends over businesses may help promote more meaningful interaction, or it may result in businesses and news sources generating heightened content to try to win back user engagement.

Algorithms will change over time responding to user needs, economic opportunities, regulation and public sentiment. Transparency by their creators in relation to what they can do, and are intended to do, is the most sustainable approach to enabling oversight and empowering people to make informed choices.

⁸⁰ A E Walderman, Disorderly Content (16 August 2021) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3906001.

⁸¹ C Findlay (19 October 2021) What it's like being disabled on the internet, Refinery 29 <https://www.refinery29.com/en-au/disabled-experience-online-abuse>.

⁸² Because of the range of potential issues associated with algorithms, many different Australian Government agencies have equities, including the Australian Human Rights Commission, the Australian Communications and Media Authority, the Australian Competition and Consumer Commission and the Office of the Australian Information Commissioner.

C. Algorithmic transparency and regulation

What eSafety is currently doing

Under the OSA, there are two specific regimes of relevance to algorithms.

The first is the development of industry codes to prevent and address harms relating to class 1 and 2 material. One of the measures recommended in eSafety's position paper guiding the development of codes is to reduce the promotion and reach of this content within algorithmic systems, including recommendation algorithms and choice architecture. Once finalised and registered, these codes will be mandatory, and eSafety will have the power to investigate potential breaches and direct services to comply.

The second is the power to require reporting from services about how they are meeting the Basic Online Safety Expectations. The effects of algorithms and how they relate to user-safety will be something we consider as eSafety implements the BOSE and reporting requirements. For example, information about the outcomes and intent of both moderation and curation algorithms would be relevant to the expectations for minimising harmful material and making sure people can use the service in a safe manner. eSafety can therefore request or require reports to help us understand the intent behind the relevant code, what inputs/outputs have been designed and the rationale for this.

In addition, our Safety by Design initiative aims to support industry to put user safety and rights at the centre of how they design and develop products and services. The next phase of our Safety by Design activities will consider the extent to which algorithms have the potential to contribute to online harms and highlight this as a factor for platforms to include in their risk-based thinking.

It is worth noting that Safety by Design applies across a broad and comprehensive set of technologies, processes and platforms and is technology-agnostic. We believe any future regulation that looks at emerging or novel technologies, whether AI, algorithms, encryption or virtual reality should also take a technology-agnostic approach in terms of combatting a variety of online harms.

Industry and international efforts

Initial limited efforts by industry to increase algorithmic transparency are a step in the right direction. However, they fall short of offering substantive explanations of the ways in which algorithms may or may not contribute to online harms.

eSafety is aware of international efforts to increase services' transparency and accountability in relation to algorithms. This includes algorithmic transparency working groups convened through cross-sector initiatives such as the Christchurch Call and the Global Internet Forum to Counter Terrorism, as well as proposals for regulation and law reform in various jurisdictions.

In the US, the Justice Against Malicious Algorithms Act has been proposed to enable services to be held liable for serious harm caused by algorithms. The EU's Digital Services Act will require greater transparency and consumer choice in relation to algorithms. Canada has developed an Algorithmic Impact Assessment tool to support regulators, advocates, public-interest technologists,

technology companies, and critical scholars who are identifying, assessing, and acting upon algorithmic harms.

The UK Joint Committee has recommended a stronger focus on the risks posed by algorithms in its *Online Safety Bill*. This includes recommending that the largest and highest-risk providers should be placed under a statutory responsibility to commission annual, independent audits of the effects of their algorithms. It also requires Ofcom to produce a mandatory Safety by Design Code of Practice, setting out the steps providers need to take to consider and mitigate risks created by a range of factors, including the nature of recommendation algorithms.

eSafety will continue to monitor these developments and work with domestic, international and industry stakeholders as we implement our own range of initiatives and regulatory options to promote greater algorithmic safety and transparency. This includes developing tools that will enable those developing algorithms to be able to anticipate and mitigate risks and build in protections prior to deployment.

Considerations for regulation

eSafety will also continue to consider what the next generation of technical regulation looks like.

There are significant practical, technical and regulatory challenges in explaining the functionality of complex algorithmic decision-making systems and their rationale in specific cases. We commend to the Committee the Ada Lovelace Institute's paper from December 2021 on technical methods for regulatory inspection of algorithmic systems, which explores six methods regulators can use and the benefits and challenges of each.⁸³

Manual review of code, as discussed in the Ada Lovelace paper, is highly onerous. Code bases can be extremely large and would require broad technical skillsets (including engineers, data scientists and technical auditors) to be able to review a wide array of algorithms, which are continuously being updated. Algorithms can be written in an endless number of coding languages which would be impossible for an internal team of technical reviewers to keep up with as a sustainable task. The paper recommends requesting pseudo-code (plain English descriptions of the code) from developers instead. While this could help mitigate the complexities of human code-base review, securing the technical and experiential skill sets to conduct technical audits would remain a key challenge.

Analysis of the source code would need to be accompanied by collection and review of potentially sensitive user data and training sets. Tactics such as site scraping to conduct reviews of the efficacy and impact of algorithms would therefore need to be supported by a clear legislative mandate and information-gathering powers.

Due to these complexities, a manual code review would be best approached if investigation of a platform's policies, procedures and engineering tactics revealed areas of concern or ambiguity, or observation of the outputs of an algorithm gave reason to believe there may be a significant problem. Code review as a supporting task done with just cause, but not as a standard practice.

⁸³ Ada Lovelace Institute (9 December 2021) Technical methods for regulatory inspection of algorithmic systems <https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/>.

Platforms and service providers may have valid security concerns around sharing their code bases requiring proper sharing arrangements (e.g., non-disclosure arrangements) to be in place.

We suggest this issue be considered in further detail when the operation of the OSA is reviewed, to determine whether the industry codes and Basic Online Safety Expectations have enabled sufficient transparency and accountability in relation to the safety impacts of algorithms. During this review, we recommend consideration be given to whether there are examples of effective global laws and policies that could be applied to the Australia context. In the interim, eSafety will continue to engage with and learn from stakeholders working in this space and assess our technical and resource needs in this area.

6. Evidence of existing ID verification and age assurance policies and practices and the extent to which they are being enforced

Key Points

1. It is important to identify the problems we are trying to solve through identity and age verification and make sure the solutions we put in place are balanced, coherent, coordinated and based on evidence. Requiring all users to be identifiable to other users – or to provide substantial identity information to verify age – does not achieve this. It also carries the risk of significant unintended consequences, noting the many benefits that a degree of anonymity offers to those seeking to protect their safety and privacy online.
2. Instead, we need to take a holistic and proportionate approach to making services age-appropriate and safe for children and young people, deterring bad behaviour, empowering victims to deal with issues, and holding perpetrators accountable.
3. The OSA provides options for eSafety to require services to address these issues. In relation to identity, the OSA creates an obligation for services to prevent the misuse of anonymous accounts on a systemic level through the Basic Online Safety Expectations, and on an individual level through our investigative schemes and associated information-gathering powers. In addition, age assurance measures will be required under the Basic Online Safety Expectations, the codes and the Restricted Access System, and eSafety is running a comprehensive consultation on this issue as part of our development of a roadmap for an age verification regime for online pornography.
4. Identity and age verification processes are also being considered through the *Anti-Trolling Bill* and *Online Privacy Bill*. It is important that Government takes a consistent and coordinated approach to these issues to avoid creating duplication and confusion for industry and the public.
5. There would be significant benefit in leveraging the findings of the age verification roadmap across related government processes, particularly given the extensive research and consultation that has informed the process and will continue to guide its development – and the importance of bringing public along on the journey.

There has been much discussion recently of the need for services to do more to protect children online. Debate has also focused on how to make sure people are not able to mask their identity to harm others. eSafety welcomes these important discussions.

With any online safety intervention, the problem must be clearly identified, the solution targeted, proportionate and effective, and the rights and best interests of users – especially children – paramount. So too, a person's right to be free from abuse online needs to be calibrated against a range of other imperatives, including the right to privacy online.

Privacy and protection from violence are both human rights which can be supported by the ability to maintain a degree of anonymity.

Therefore, it is important to emphasise the fact that an internet user who may not be identifiable to other users is not, in itself, a problem. The challenge is that real or perceived anonymity may contribute to a person's willingness and ability to abuse others – and may hinder efforts to hold them accountable for that behaviour. As set out below, making all users immediately identifiable to one another is not necessarily an effective way to address this issue, and carries a significant risk of unintended consequences which may make some people less safe.

Likewise, imposing a requirement for services to conduct age assurance or verification will not, by itself, serve to protect children from dangerous or inappropriate content, contact or conduct. Rather, age assurance processes are an important first step for services to be aware that a particular person is likely to be a child. It is important to consider what the most appropriate 'next steps' are once a person's age is estimated or verified to promote their safety and privacy. As explained below, eSafety has begun to explore this issue, at the Government's request, by developing a roadmap for a mandatory age verification regime for online pornography. Age assurance measures and 'next steps' for keeping children safe – such as imposing robust safety and privacy settings by default – are also key elements of the Basic Online Safety Expectations and the codes.

These are complex, interwoven issues that deserve deep and balanced consideration, which is why eSafety is conducting an extended consultation process on age verification⁸⁴ and taking a deeper analytical dive into anonymity and identity shielding.⁸⁵ From eSafety's perspective, taking the time to get these processes and roadmap right so they can be carried out effectively is preferable to moving quickly.

A. Identity verification and anonymity

Key issues and terminology

Before delving into the evidence, we consider there may be benefit in exploring some of the key issues and terminology in this space.

Anonymity and identity shielding

Anonymity and identity shielding allow someone to hide or conceal their personally identifying information online. This means that any data or information held has been treated so that a person can't be identified, directly or indirectly. It protects a person's identity from being shared.

⁸⁴ <https://www.esafety.gov.au/about-us/consultation-cooperation/age-verification>.

⁸⁵ eSafety (22 January 2021) Anonymity and identity shielding – position statement <https://www.esafety.gov.au/industry/tech-trends-and-challenges/anonymity>.

While there are many ways of appearing anonymous online, this does not mean a person is unidentifiable or incapable of being held accountable for misbehaviour. The spectrum of anonymity on social media and other online services can include:

- **Full anonymity** – interactions where people do not provide any personal information or identifiers, and neither the online service nor other users can identify the person at the time of the interaction or subsequently.⁸⁶

It is rare for a service to collect no information about their users, and so full anonymity generally requires a user to take active **identity shielding** steps to prevent the collection of their data, for example, by using a virtual private network (VPN) or other technologies to prevent disclosure of their geo-location or Internet Protocol (IP) address. For the average user, it is not easy to completely shield their identity (and achieve full anonymity).

It is important to note there are legitimate reasons for people to employ tools such as VPNs – for example, to keep their information secure when using public Wi-Fi.⁸⁷

- **Publicly anonymous** – interactions involving a person who may appear anonymous to others, however the service collects and holds some information about the person, such as their geo-location, IP address, the way they have engaged with the service and its other users, etc.

It is important to note there are legitimate reasons for people to be publicly anonymous – for example, to protect their privacy and confidentiality when seeking information and help online about sensitive topics.⁸⁸

- **Pseudonymity** – interactions involving a person whose registered username, handle or avatar is not their real name,⁸⁹ however, the service collects and holds some information about the person. For example, many services require people to provide an email address or phone number at sign-up.

While some services have ‘real name’ policies,⁹⁰ many allow the use of pseudonyms.⁹¹

It is important to note there are legitimate reasons for people to choose pseudonyms rather than using their real names online. For example, eSafety advises children not to use their real names online due to safety and privacy risks associated with sharing their personal details with people they do not know.⁹²

⁸⁶ OAIC, Chapter 2: APP 2 – Anonymity and pseudonymity (2.4): <https://www.oaic.gov.au/privacy/australian-privacy-principles-guidelines/chapter-2-app-2-anonymity-and-pseudonymity>.

⁸⁷ Australian Cyber Security Centre, Use a secure connection: <https://www.cyber.gov.au/acsc/view-all-content/guidance/use-secure-connection>.

⁸⁸ See, eg, ReachOut, Effectiveness of online help: <https://schools.au.reachout.com/articles/effectiveness-of-online-help>.

⁸⁹ OAIC, Chapter 2: APP 2 – Anonymity and pseudonymity (2.6): <https://www.oaic.gov.au/privacy/australian-privacy-principles-guidelines/chapter-2-app-2-anonymity-and-pseudonymity>.

⁹⁰ See, eg, Facebook: <https://www.facebook.com/help/112146705538576>.

⁹¹ See, eg, Twitter: <https://blog.twitter.com/common-thread/en/topics/stories/2021/whats-in-a-name-the-case-for-inclusivity-through-anonymity>.

⁹² eSafety, Sharing photos and my personal information online: <https://www.esafety.gov.au/kids/i-want-help-with/personal-information-online>.

- **Identity authentication** – interactions involving a person whose identity or contact details have been authenticated by the service in some way – for example, through an authentication code sent to the person's registered phone number or email address.
- **Identity verification** – interactions involving a person whose identity has been verified by the service through the provision of identity documents such as a passport or driver's licence.

There are also different ways to conceptualise a person's identity.

The social identity that an internet user establishes is often referred to as **online identity**. In offline interactions, there is a barrier between an individual's different social circles, interests and communities. Different online identities can help individuals translate this complexity to a digital space.

In contrast, **digital identity** often refers to either the verification of peoples' identity or to their data footprint online. A person's digital identity can comprise thousands of data points that make up a person's profile, preferences, and behaviours online. It can include emails, usernames, passwords and biometrics as well as usage patterns and purchases – essentially, factors which make people distinguishable from one another.

A person's digital identity can be verified using official documents or other identification credentials. This method of verification is generally accepted or required in certain contexts, such as accessing government, banking and gambling services online.

Digital identities can be supported through:

- **Two-Factor Authentication (2FA) and Multi-Factor Authentication (MFA)**. These methods require people to provide two or more authentication factors to verify themselves, such as a security token, a code from a smartphone, a fingerprint or other biometric identifier.
- **Blockchain-based identity management systems (BBIMS)** which store information entirely on a mobile device, secured with a private key and shared with designated others via the blockchain.
- **Digital signatures**. A digital signature uses asymmetric/public key cryptography to authenticate the sender of information and to verify and maintain the integrity of data, like a handwritten signature. On a platform, people could filter information which has not been provided with a digital signature.

It is also possible for digital identity systems to share specific personal attributes/credentials at the person's choosing (e.g., age). These tools can help confirm that people are real and not bots and can support online services to obtain appropriate 'basic subscriber information'. Of course, there are challenges with access and uptake of digital identity tools.

First, people may not have access to official and current government identity documents, they may be disenfranchised or already face barriers to engaging online. Second, it can unreasonably restrict people's right to expression and participation online. Requiring use of digital identity for online engagement risks excluding Aboriginal and Torres Strait Islander peoples, people

experiencing family and domestic violence, people experiencing homelessness and people who have come to Australia as refugees, from participation online.

Benefits to anonymity

There are many benefits and valid reasons for maintaining a level of anonymity or practicing identity shielding online, as set out in our position statement.⁹³ These include:

- Controlling how personal data is collected and stored as well as who can access and use it.
- Protecting users from unwanted contact and abuse. For example, a child using a username or nickname online which makes it more difficult for predators to interact with them. This is also important for people experiencing domestic and family violence, to make it more difficult for stalking or harassment to be perpetuated through technology.
- Engaging freely online without judgment. For example, it can allow people who are same sex attracted, intersex or gender diverse to explore their identity and to talk openly about their sexuality free from being 'outed' to their family and friends or harassed. It can be usefully applied to any situation where a person is at risk of being criticised or harmed for speaking out, such as:
 - allowing victims of domestic violence to participate online out without fear of being located;
 - providing a voice to those from diverse communities who are at risk online and who may otherwise be silenced or suppressed in public debate;
 - giving a voice to those living under oppressive or authoritarian regimes;
 - allowing individuals to express their religion, or lack thereof – without fear of persecution; and
 - helping to protect workers or whistle-blowers, who may fear that voicing an opinion publicly could cost them their jobs.

Anonymity can also enable people to seek mental health or other support services.

Participants in focus groups conducted by eSafety in 2021 stated:

'I don't think I would've gotten help for my mental health if the platform I used to reach out wasn't an anonymous platform'.

'People are embarrassed to discuss certain things due to social stigma or other reasons with friends and family, therefore reaching out to an online community under anonymity can give people answers and support when they have no one else to speak to'.

It is important to implement feasible, proportionate and privacy-preserving solutions to address clearly defined problems. This is so we avoid unintended consequences such as safety risks for

⁹³ eSafety (22 January 2021) Anonymity and identity shielding – position statement <https://www.esafety.gov.au/industry/tech-trends-and-challenges/anonymity>.

victims of abuse or oppressed communities; reluctance to seek help or information in relation to stigmatised issues; or the creation of ‘honey pots’ of sensitive user information.

Negative uses of anonymity

Notwithstanding these benefits, anonymity and identity-shielding can also be used as vectors for negative, harmful or illegal online behaviour. Anonymity has been linked to the creation and spread of disinformation, the cyberbullying of children using multiple accounts and the amplification of hate speech, among other forms of harm.

However, there is limited research exploring the relationship between anonymity and harm. In a scan of the evidence, eSafety found only a handful of harm perception studies focusing on cyberbullying and inflammatory comments by anonymous users. A review of this literature shows that the impacts resulting from anonymous accounts are not always clear and can vary according to the negative experience in question.

Several studies report that, although the perpetrators of cyberbullying were often perceived as anonymous, cyberbullying commonly takes place in the context of young people’s social groups and relationships from the ‘real’ world.⁹⁴ It was noted that where victims discover the identity of an anonymous perpetrator, and it turns out to be someone known to them, the impact can be more severe than if it were a stranger.⁹⁵

Some studies have also found that online discussions tend to be more polite on platforms with ‘real name’ policies than those where comments can be made anonymously or pseudonymously.⁹⁶

While perceived anonymity is likely to contribute to users’ willingness and ability to abuse others online, there are many other contributing factors which must be considered when exploring potential solutions. Experts have pointed to the ‘disinhibition effect’ of communicating online – including feeling invisible, not seeing other people react, not seeing the consequences of actions and/or forgetting that other users are real people not just on-screen profiles – as a factor in why individuals can cause harm to others online.⁹⁷ Aggressive expression online can also be a function of group norms on a particular service or platform.⁹⁸

In some cases, behaviour is not merely careless but calculated to cause harm. In these cases, the underlying issue is not necessarily that the person is acting anonymously, but that they are able to create multiple fake accounts to abuse a victim with impunity. We have seen this issue arise in our

⁹⁴ R Dennehy, S Meaney, J Walsh, C Sinnott, M Cronin and E Arensman ‘Young people’s conceptualizations of the nature of cyberbullying: A systematic review and synthesis of qualitative research’ *Aggression and violent behavior* 51 (2020) 101379 [doi: 10.1016/j.avb.2020.101379](https://doi.org/10.1016/j.avb.2020.101379).

⁹⁵ N Baas, M D T de Jong, and C H C Drossaert ‘Children’s perspectives on cyberbullying: Insights based on participatory research’ *CyberPsychology, Behavior & Social Networking*, 2013, 16(4), 248–253. [doi: 10.1089/cyber.2012.0079](https://doi.org/10.1089/cyber.2012.0079); K Naruskov, P Luik, A Nocentini, and E Menesini, E ‘Estonian students’ perception and definition of cyberbullying’ *TRAMES: A Journal of the Humanities & Social Sciences*, 2012, 16(4), 323–343. [doi: 10.3176/tr.2012.4.02](https://doi.org/10.3176/tr.2012.4.02).

⁹⁶ D Halpern, J Gibbs ‘Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression’ *Computers in Human Behaviour*, 2013, 29(3): 1159–1168, [doi:10.1016/j.chb.2012.10.008](https://doi.org/10.1016/j.chb.2012.10.008); I Rowe ‘Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms’ *Journal of Broadcasting & Electronic Media*, 2015, 59:4, 539–555, [doi: 10.1080/08838151.2015.1093482](https://doi.org/10.1080/08838151.2015.1093482).

⁹⁷ J Suler ‘The Online Disinhibition Effect’ *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society*, 2004, 7(3) 321–326. [doi:10.1089/1094931041291295](https://doi.org/10.1089/1094931041291295).

⁹⁸ L Rosner and N Kramer ‘Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments.’ *Social Media+ Society* 2.3, 2016 [doi:10.1177/2056305116664220](https://doi.org/10.1177/2056305116664220).

investigations, where new accounts have continuously sprung up to continue to bully or threaten a victim after they successfully block and report the initial offending account. Rather than requiring every user to verify their identity upon the creation of an account, services could improve their processes for detecting likely abuse, for example, where a proliferation of accounts are being created on a single device. Taking steps to authenticate those particular accounts offers a more targeted means of addressing abuse while allowing other rule-abiding users to continue to benefit from a level of anonymity.

Potential solutions and what industry is doing now

Applying a Safety by Design approach to the issue, eSafety believes services need to improve three facets:

- **Deterring misuse** – Preventing poor behaviour is the best option. Services should continue to innovate ways to reduce the sense of disinhibition online, to promote pro-social norms on their platforms and to create friction to make it more difficult to abuse others. An example of this is ‘nudge’ technology which sends a warning to users before they send or post a comment which is likely to be abusive or problematic, asking them if they are sure they wish to continue.
- **Empowering victims** – When problems do occur, users should have easy access to appropriate tools and information to keep themselves safe, and to seek swift and effective help from the services where the harm is happening. Blocking, muting and reporting mechanisms are basic examples of these types of ‘conversation controls’. Some services are exploring ways to provide enhanced tools to users who appear to be receiving high volumes of abuse, for example, the ability to block all messages from accounts that have only recently connected with the user.
- **Holding perpetrators accountable** – Where online abuse is serious and/or ongoing, there must be avenues to hold perpetrators accountable. This requires effective enforcement of consequences for terms of use violations, such as temporary account suspensions and permanent bans. This could include, for example, using device identifiers to block the creation of further accounts. It could also include requiring the account holder to verify certain identity or contact details before they are able to re-gain access to their account. Increasingly, we are seeing social media platforms and other online services deploy multi-factor authentication and other forms of identity verification, particularly in response to account violations, and this is a positive step. When serious transgressions occur, this information should be capable of being shared with relevant authorities (for example, law enforcement agencies and online safety regulators) through appropriate legal processes so they are able to pursue an investigation.

Through a combination of these measures, services can achieve a balance between safety, privacy and security for their users. eSafety does not believe that blanket bans of anonymous or pseudonymous accounts would be desirable or achievable, particularly when one considers the vast number of existing unverified accounts on the internet and the fact that most online services are global in nature (not all people who abuse Australians will be based here in Australia).

Proportionate and preventative measures to address online harms are available to industry. Such measures are focused on monitoring for threats (such as multiple account creation in quick succession) to prevent harms from occurring.

In addition, services can use individual online attributes, such as email addresses, phone numbers, IP addresses or mobile device identifiers (IMEIs) to identify users or link harmful behaviours to user accounts. These attributes, or other web or device identifiers, are often enough for platforms to take action to limit the ability of users to re-register or create multiple fraudulent accounts.

This removes the need for a service to gather large amounts of identifying information on all of their users and can instead focus on addressing the actual harm being perpetrated.

As noted above, services also have the option to escalate to requiring firmer identity verification where warranted – for example, where an account is found to be acting in an abusive manner.

Areas for improvement

Online service can do more to promote pro-social online behaviours and build safer platforms through enhanced content moderation, more robust account management, user empowerment and user accountability. To improve user safety, platforms should consider:

- Enhancing community guidelines and terms of service to better address different types of harmful behaviour and content.
- Developing methods to effectively enforce breaches of guidelines and terms of service. The consequences for harmful behaviour should be transparent and consistent.
- Promoting terms of service as they relate to user safety – whether through platform ads or notifications, having a dedicated and promoted safety resource centre, or running awareness-raising campaigns on appropriate and safe behaviour.
- Increasing transparency and data sharing between platforms to block abusers or to prevent abusers repeatedly recreating (phoenixing) accounts. Accounts being linked to a single account holder would allow eSafety to use more appropriate remedies, such as issuing an end-user notice to the person responsible, instead of relying on the platforms to repeatedly remove accounts reported for cyberbullying. Consistency of basic subscriber information collection could also help to address this, as discussed in more detail below.
- Enhancing user tools to protect themselves on platforms, including:
 - better and more transparent reporting procedures
 - default account settings which prioritise safety and privacy of people over settings which are more open to other unknown users by default
 - tools which empower people to control their experience, such as comment filters or customisable blocked terms to target specific words or issues that proactively screen messages out of chat or comment sections before they are seen. This can be effective for targeted and contextual harassment that would not otherwise be picked up by AI scanning technology

- timely and easy access to internal complaints procedures.

Platforms and governments could increase user awareness of existing online safety features as a proactive process, rather than as a reaction to a negative online experience. Measures may include:

- Platforms having responsibility to promote online safety resources or reporting schemes to Australians and refer people to counselling and support services.
- Platforms proactively educating people on acceptable behaviours and user expectations and following through with enforcement of those expectations if breached.
- eSafety's website contains a wealth of useful information to help Australians have safe, enjoyable experiences online. However, there is an opportunity to expand the prevention and inclusion work, as new audiences are emerging (e.g., victims of adult cyber abuse). This could take the form of webinars, podcasts or partnerships with industry and NGOs.

How the OSA intersects with identity verification

The OSA provides eSafety with several options to drive improvements across deterrence, empowerment and accountability. Most notably, as set out in section 4D, the Basic Online Safety Expectations will set a higher benchmark for the steps we expect services to take to keep people safe. The draft determination includes a specific expectation regarding anonymous accounts, requiring services to take reasonable steps to prevent those accounts from being used for harmful or unlawful material or activity. eSafety will be able to require services to clearly set out the steps they are taking, and to explain how they are adequate and effective to prevent and address online harm. This will enable us to address online harms – including those exacerbated by identity-shielding – on a systemic basis.

Information-gathering powers

In addition, the OSA strengthens eSafety's information-gathering powers. Under Part 13 of the OSA, eSafety can compel online service providers to produce information about the identity of an end-user or contact details of an end-user where it reasonably considers this information is relevant to the operation of the OSA. This will help eSafety to conduct investigations into harmful online behaviour, issue notices, and issue fines. If an online service provider ignores a notice, it could face enforcement action, such as a civil penalty.

However, we note based on our experience that obtaining basic subscriber information from a service is only a first step in opening further lines of inquiry to try to identify a person. It is not always the case that a person's identity or even contact details can be established through information held by a digital service. For example, while an Australian mobile phone number may be useful information, the identity of the subscriber can only be ascertained by querying restricted databases or seeking further end-user information from the relevant service provider.

Under Part 14 of the Act, eSafety can also require a person to appear before the Commissioner, produce documents, answer questions, or provide any other information in connection with an investigation.

By opening key lines of inquiry for eSafety to pursue through its regulatory investigations, these powers will enable us to hold individuals accountable for perpetrating harm, including those operating publicly anonymous or pseudonymous accounts.

Consistency across government

On 1 December 2021, the Attorney-General's Department released an exposure draft and explanatory paper for the *Social Media (Anti-Trolling) Bill 2021* (Cth), which seeks to address defamation on social media services. The Bill includes a conditional defence from defamation liability for social media services which adopt and comply with mechanisms that allow the services to disclose 'relevant contact details' in response to a request from a complainant. This can be done either with consent or in response to a court order. 'Relevant contact details', for these purposes, are the commenter's name or the name by which they are usually known, phone number and email address. Details which turn out to be fake or inaccurate, and so do not allow the complainant to effectively commence proceedings against the commenter, will not meet the definition.

Accessing the conditional defence will require a significant expansion of services' current practices for the collection and verification of Australian users' information. In eSafety's experience, there is little consistency in the type of basic subscriber information collected by services, and the quality of the data can vary significantly. For example, a person may provide an email address or mobile phone number at sign-up, but as these are not always routinely re-verified, the data may go out of date. Additionally, while temporary or 'throwaway' email addresses may be prohibited by platforms' terms of service, this is not enforced across the industry. In these circumstances, as the Bill is currently drafted, the service may not be able to access the conditional defence from defamation liability.

eSafety will make a separate submission on this proposal, drawing on our experience conducting regulatory investigations and requesting such information from services. The submission will outline some of the challenges with reliance on particular data points for purposes of identifying or 'unmasking' a person, as well as some of the potential unintended consequences of the Bill.

We particularly note the potential for confusion over what the Bill seeks to achieve and how it relates to eSafety's Adult Cyber Abuse Scheme, as the term 'trolling' is not generally understood as being synonymous with defamation.

It also appears to us that social media providers would be incentivised to take steps to retroactively collect and validate the relevant contact details for all of their existing Australian users, in case those users end up engaging in defamation in the future, if they wish to access the conditional defence from defamation liability. We believe some users may be hesitant to provide these details to services due to a lack of trust, and others may have difficulty complying – for example, low income users who may not have a static mobile phone number. We suggest there is benefit in considering a more targeted approach, involving the authentication or verification of those accounts which are identified as engaging in harmful behaviour, rather than all accounts.

As highlighted above, eSafety will be working on closely related issues, including investigating complaints of adult cyber abuse and seeking to address systemic safety issues – including the misuse of anonymous, pseudonymous or multiple accounts – via Safety by Design and the Basic

Online Safety Expectations. We look forward to continuing our discussions with the Attorney-General's Department to promote a combined approach to online harms across Government.

B. Age Assurance

Key issues and terminology

Age is one aspect of a person's identity.

Age assurance is an umbrella term for measures which determine the age or age-range of a user. Age assurance encompasses **age verification** where the age (or age range, such as 18+) of a user is determined to a high level of confidence – for example, verifying age against identity documents. Some, but not all, age verification measures also involve identity verification.

Age estimation measure estimate a user's likely age, often using AI or other behavioural analysis, but with a lower level of statistical certainty than age verification.

Age assurance tools include:

- Age screening and age-gating – relies on people self-declaring their personal details, for example checking a box declaring they are over 18 to view a site selling alcohol or entering a birthdate when creating a social media account.
- Behavioural signals technology – utilises AI and machine learning to determine age based on online behaviour.
- Age prediction and estimation tools – generally uses biometric data to assess age. For example, using a facial scan to assess that a person is over 18.
- Age-related eligibility assurance (AREA) tokens – a person purchases a token from a retailer, the same as purchasing cigarettes or alcohol. This token can then be used to access age-restricted content.
- Third party identification – third party companies collecting and verifying identities (such as credit checks).

Shifting the burden of responsibility for young users

As noted above, children and young people are at higher risk of a range of online safety issues. The internet was created by adults and for adults, but it is young people who are fully inhabiting the online world – and shouldering the primary burden of online risk and harms.

The burden of responsibility needs to be flipped so that large tech companies take responsibility for online safety and embed safety features in the design and development of their products. It should not be left to children, parents and members of vulnerable communities who experience online abuse at rates much greater than the general population to protect themselves online from harms that are enabled by the design of services.

Where the responsibility should be

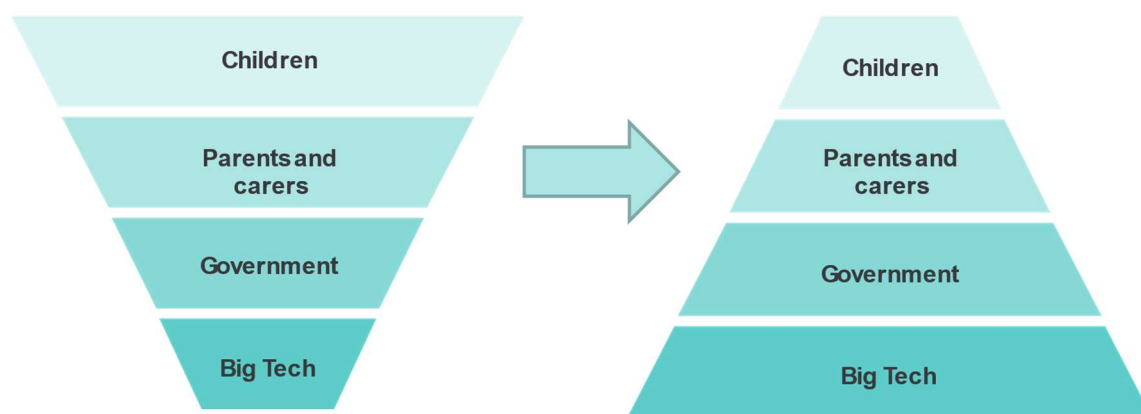


Figure 2

To put in place appropriate safeguards for young users, services must be able to ascertain which of their users are young, requiring some form of age assurance. Identifying the age of people can allow platforms to design and develop spaces and experiences which are safe, age-appropriate and protect children from content which they may not be developmentally ready for, or from being contacted by unknown adults. It can also be used to safeguard children's data from being used by advertisers or ensuring young people's accounts have the highest privacy settings by default.

Existing industry measures

Many social media services have a minimum age to use their platforms (often 13 years, consistent with US privacy law). While there is no single consistent approach to preventing underage users from accessing services, many major services use one, or a combination of, the following methods.

- Age ratings in app stores which enable parents to block downloads of age-inappropriate apps using device-level controls.
- Age gates at sign up, requiring people to enter a complete birthdate. If a person initially enters a birthdate which does not meet the age minimum of the site, the platforms can prevent that person (through device identification or IP address) from attempting to make another account to bypass the age gate.
- Age estimation technology, which allows safety and moderation teams to flag accounts that behave as though they are underage. This can be supported by reports from other people and monitoring of keywords to surface potential underage accounts.
- In some circumstances and jurisdictions, platforms verify peoples' ages through official identity documents.

Age verification roadmap and initial stakeholder views

In February 2020, the Standing Committee on Social Policy and Legal Affairs finalised an inquiry into age verification for online wagering and online pornography. One of its recommendations was to task eSafety with developing a roadmap to lay the necessary groundwork for a successful age verification regime.⁹⁹ The Senate Committee report specifically noted that a successful age verification regime would require robust privacy, safety and security standards, a legislative basis, the consideration of complementary non-technical solutions and, importantly, understanding and acceptance among the community.

Bringing the public on the journey and helping them to understand these technologies and how data is used was identified as critical in the Senate Committee's report. The need to raise awareness and support for age verification is supported by eSafety's research with Australian adults, which showed that awareness of age verification is not high (around 50%).¹⁰⁰ Further, about a quarter (24%) of respondents lacked confidence in the effectiveness of its design, implementation and operationalisation by government.

This is why, in part, the Australian Government has asked eSafety to conduct a thorough consultation and to develop a roadmap to lay the foundation for a workable age verification regime to protect children from exposure to pornography.

The age verification roadmap process includes research and consultation on proportionate age verification and age assurance measures to make recommendations which preserve privacy and offer data security. There are international standards and technical requirements currently in development for age verification and age assurance technologies, which eSafety will draw on to draft our report.

eSafety has until December 2022 to deliver the roadmap to ensure sufficient time to consult with industry, stakeholders and the public. This time is necessary to make sure age verification options are underpinned by a robust base of evidence and partnered with extensive community consultation.

This process began in August 2021 when eSafety issued a call for evidence seeking insights into effective age verification and age assurance techniques. We also called for evidence about the impact of online pornography on children and proven methods of educating young people about both respectful and harmful sexual behaviours. We have published a high-level thematic analysis of the evidence emerging from this first phase of input on our website.¹⁰¹

These submissions have informed the next phase of targeted consultation, which started in November 2021 to allow closer examination of the themes that emerged from the call for evidence. We are still in the early stages of the consultation process, and there will be more opportunities to contribute throughout 2022 for those yet to be involved.

⁹⁹ https://www.aph.gov.au/Parliamentary_Business/Committees/House/Social_Policy_and_Legal_Affairs/Onlineageverification/Report.

¹⁰⁰ eSafety, Public perceptions of age verification for limiting access to pornography, October 2021, <https://www.esafety.gov.au/research/public-perceptions-age-verification-for-limiting-access-pornography>.

¹⁰¹ <https://www.esafety.gov.au/about-us/consultation-cooperation/age-verification>.

In our initial phase of consultation, stakeholders have raised it is essential that any approach to age verification or age assurance is proportionate, viable and privacy and security-preserving.¹⁰² There is a risk that age assurance or verification measures, if not carefully considered and deployed, can have unintended consequences and erode the privacy of both adults and children.

Privacy and data collection

Stakeholders emphasise that age verification or age assurance measures should be privacy-preserving and minimise the data they collect. In many cases, there is no need to collect information beyond the attribute that a user is above a set age. Noting that one of the inquiry's terms of reference is the collection and use of relevant data by industry in a safe, private and secure manner, it is not presently clear to us that industry has the capability to collect and hold large amounts of personal information in a private and secure manner.

Multiple tools

Stakeholders have raised that age verification technology is not a silver bullet for addressing harms associated with access to pornography. It is only one tool of many to address the potential harm and needs to be supported with additional measures. Risks should be approached in a holistic manner to empower young people and their parents. This needs to be supported with education for children, as well as their parents and carers, on sexuality and online pornography, having conversations about these issues, and how to use filters and safeguards to curate a safe online experience. Pornographic content can be accessed on the internet on multiple platforms and in multiple formats, beyond dedicated pornography sites. A one-size-fits-all technological solution would not be effective.

Trust

The effectiveness of age verification measures also relies on significant amounts of trust. Users and industry must trust that the verification providers are accountable and that their data is protected and being used appropriately. To gain trust, age verification measures need to be proportionate, data-minimising, privacy-preserving and clearly and transparently communicated to the public.

Consultation

We consider extensive and wide consultation as vital to developing any age verification measures. Lack of public consultation and subsequent lack of public support was a key reason the United Kingdom abandoned its original age verification mandate.¹⁰³

It is also important to avoid imposing duplicative or inconsistent age assurance requirements across different online harms. Given that age assurance is relevant to multiple eSafety workstreams – including industry codes, the Basic Online Safety Expectations, and the Restricted

¹⁰² <https://www.esafety.gov.au/about-us/consultation-cooperation/age-verification>.

¹⁰³ In May 2021, the UK government released the Online Safety Bill 2021 (UK), which includes the requirement that if a platform can be accessed by children, it will have to comply with the safety duties for child protection. It also imposes duties of care in relation to content that is harmful to children, such as pornography. Age-verification is one of many measures that companies in scope may be required to implement to protect children from inappropriate content.

Access System declaration – we will work to ensure alignment across the different elements and to cross-pollinate learnings from different engagement processes.

Age assurance measures within eSafety's remit

Age verification roadmap: The roadmap will propose a suitable legislative and regulatory framework, with technical and non-technical tools to minimise the risks and harms associated with young people's exposure to online pornography. eSafety will present a proportionate, viable and effective approach to mitigating harms to children; respect the privacy, security and safety of Australians online; balance the regulatory burden and cost; and harmonise Australia's actions with international measures.

Basic Online Safety Expectations: A core expectation is that service providers will take reasonable steps to make sure technological or other measures are in effect to prevent access by children to class 2 material. Reasonable steps include implementing age assurance mechanisms. eSafety will have the power to require services to report on how they are meeting this expectation.

Codes: The OSA provides that industry-led codes are to be developed to address the harms associated with class 1 and class 2 material. eSafety recently released a position paper setting out our expectations for the codes. The paper suggested that industry should develop a code to address very serious class 1 material by July 2022, and another code to address children's access to class 2 material by December 2022, in alignment with the timing for the age verification roadmap so the learnings and connections made through that process could inform code development. Regardless of whether they take this phased approach, the code(s) must include measures to prevent children from accessing pornography, which would require some form of age assurance or verification.

Restricted Access System declaration: The OSA provides eSafety with discretion to issue a remedial direction to a service provided from Australia requiring them to place class 2 - R18 content behind a restricted access system (RAS). The requirements for a RAS, including any age assurance requirements, will be set out in a legislative instrument following consultation with industry.

Together, we believe these provisions can have significant impact over time, and we will seek to evaluate their effects in the months and years following the commencement of the OSA.

Consistency across government

On 25 October 2021 the Attorney-General's Department released an exposure draft of the *Privacy Legislation Amendment (Enhancing Online Privacy and Other Measures) Bill 2021* (Cth), which would enable the creation of a binding Online Privacy code for social media services, data brokers and other large online platforms with an Australian link. Under the code, social media platforms will be required to take all reasonable steps to verify their users' age and get parental consent from people under the age of 16. eSafety has made a submission emphasising the need for a consistent approach to age verification across Government.

Australia is one country in a global market and industry. It is important to ensure cohesive and connected government intervention to provide certainty to the public and industry, and to avoid the potential for confusion, duplication or the risk of working at cross-purposes. We would caution against introducing multiple identity and age verification or assurance requirements and would welcome efforts to build on eSafety's ongoing regulatory experience, research and stakeholder consultation across these issues.

eSafety looks forward to continued engagement in these processes and ongoing participation in discussions with the OAIC and the Attorney-General's Department to promote coordination and positive outcomes for Australians.

As noted in section 1, we suggest eSafety could coordinate any national work which may arise from this inquiry in line with our national leadership mandate and our ongoing stakeholder engagement and research for developing an age verification roadmap.

7. Other related matters: Horizon scanning and international engagement

Key Points

1. Given the rapid pace of change within the threat landscape, online safety regulators must have a proactive and anticipatory mindset. eSafety engages in consultation and environmental and horizon scanning to understand evolving online threats so we can stay a step ahead of technology trends and challenges.
2. Online safety issues are global challenges which require collaborative responses. eSafety contributes to cross-border and multi-disciplinary discussions, working to elevate the status of safety alongside privacy and security in the digital ecosystem. We work to lift the capacity of industry and governments to prevent and address online harm and create long-term systemic change to strengthen online safety across jurisdictions and sectors.
3. It is important to promote harmonisation and combined approaches to avoid a patchwork and fragmentation of online safety legislation which could end up working at cross-purposes.
4. eSafety continues to monitor international developments and engage with relevant government and non-government bodies globally. We are currently world leaders in online safety, but we also want to evolve with major paradigm shifts so we can maintain our global leadership well into the future.

Technological change will always outpace policy. To be effective, it is imperative to stay a step ahead of tech trends and challenges and ensure that a lens of 'safety' is applied to emerging issues. At eSafety, we take a balanced and nuanced view to emerging technologies and trends. We weigh up the risks and benefits those innovations could have for the safety and wellbeing of the public, but also provide a critical lens on how these changes could be used to abuse, harass or harm individuals and then point to solutions where they exist.

A. Horizon scanning

To keep ahead of the curve, we continually scan for new research, policy, legislative and technical updates which we cross-reference with our investigative insights as well as engagement with experts from around the world. We share these insights publicly to stimulate community debate on issues that might be on the periphery of our collective vision. We have developed several position statements¹⁰⁴ outlining the safety and regulatory implications of on-the-cusp issues. These include the emergence of convincing deepfake technologies,¹⁰⁵ the move to immersive virtual reality

¹⁰⁴ <https://www.esafety.gov.au/industry/tech-trends-and-challenges>.

¹⁰⁵ eSafety (11 May 2020) Deepfake trends and challenges – position statement <https://www.esafety.gov.au/industry/tech-trends-and-challenges/deepfakes>.

environments like the ‘metaverse’,¹⁰⁶ and the growing interest in a more decentralised internet or Web 3.0 where there are no longer central points of responsibility for the moderation of harmful content and activity.¹⁰⁷

It is this constant evolution of technology that makes it more important now than ever for governments to keep up to speed with these changes and review legislation to make sure it remains fit for purpose so we can limit online harms. We believe a principles-based approach – like that of Safety by Design and the Basic Online Safety Expectations – will allow us the flexibility to assess threats posed by major paradigm shifts, such as the metaverse and decentralisation, and be on the front foot to contribute to solutions.

B. International engagement

The internet, and its associated harms, know no geographic boundaries. Reducing online violence, stepping up the fight against child sexual exploitation and making sure children can safely navigate the digital environment is a global effort. This is why we are committed to collaborating internationally to help shape the evolving norms and standards for how technology is designed, deployed and regulated. We will only be able to shape the global landscape, and progress solutions, through collaboration, information and intelligence sharing and partnerships.

As an established online safety regulator, eSafety – and Australia – has a unique opportunity to shape and influence an emerging landscape.

Global leadership

There is widespread recognition that eSafety is a global thought leader in online safety and ahead of the curve in its approach and vision to address online harms. Through our international engagement, we are providing inspiration and guidance to governments and stakeholders around the globe – particularly through Safety by Design, which has been internationally recognised as a flagship initiative, as well as our evidence-informed, best practice education and prevention resources.

eSafety engages in several global alliances and initiatives, including the G7 and G20 online safety streams, in working groups convened by the OECD, World Economic Forum and ITU, and in leadership roles in the We Protect Global Alliance and INHOPE. We also regularly contribute to broader global debates, discussions and projects, advocating for a human rights-based approach to online safety in multilateral forums and encouraging take-up of Safety by Design as a model approach to online safety.

We work with the Department of Foreign Affairs and Trade (DFAT) to include online safety issues and actions in Australia’s International Cyber and Critical Tech Engagement Strategy and contribute to UN resolutions, including General Comment No. 25 (2021) on children’s rights in

¹⁰⁶ eSafety (10 December 2020) Immersive technologies- position statement <https://www.esafety.gov.au/industry/tech-trends-and-challenges/immersive-tech>.

¹⁰⁷ eSafety (29 July 2021) Decentralisation – position statement <https://www.esafety.gov.au/industry/tech-trends-and-challenges/decentralisation>.

relation to the digital environment.¹⁰⁸ In partnership with DFAT, we also deliver online safety capacity building projects in the Indo-Pacific to increase global cooperation and dialogue on these matters.

Changing regulatory landscape

The global online safety landscape is changing, with governments looking to fundamentally shift the way the digital environment operates by introducing legal requirements and regulatory frameworks for how platforms operate. Significant legislative reforms in the US, EU and UK will have the power to alter the global digital ecosystem. As outlined in section 4D, proposals in these countries involve more specific requirements than the OSA – for example, risk assessments and technical audits – as well as heavier penalties. We are leaders now in online safety, but we also want to make sure we evolve with major paradigm shifts so that we can maintain our global leadership well into the future.

eSafety notes the importance of a combined approach to reducing online harms. This includes promoting multi-stakeholder and multi-disciplinary partnerships and perspectives in formulating and developing policy responses. This will allow us to secure harmonisation across jurisdictions and avoid a patchwork and fragmentation of online safety legislation and governance arrangements.

eSafety looks forward to continued engagement in these processes and ongoing participation in global discussions to increase awareness of Australia's regulatory purpose and priorities – and to promote coordination and positive outcomes for Australians.

¹⁰⁸ United Nations (2021) General Comment No.25 (2021) on children's rights in relation to the digital environment <https://digitallibrary.un.org/record/3906061?ln=en>.

8. Conclusion

eSafety appreciates the opportunity to contribute to this important Inquiry.

eSafety believes our holistic remit across the pillars of prevention, protection and proactive and systemic change – strengthened by the OSA – will continue to give Australians safer, more positive experiences online by placing greater onus on services across the online ecosystem to prevent and address both individual and systemic harms.

We are aware of concerns raised by stakeholders who believe that some of our schemes do not go far enough, as well as global regulatory proposals which carry more significant penalties for non-compliance.

eSafety will continue to conduct research, extract insights from our investigations, engage in horizon-scanning, and work closely with stakeholders here and overseas to make sure we do everything we can to promote online safety for Australians in a way that is balanced, coordinated and evidence based.

eSafety will also monitor the implementation of the OSA, reporting annually on our efforts, and prepare for the independent three-year review.

eSafety is happy to respond to any questions the Committee may have and welcome the opportunity to take forward any work the Committee may recommend in line with our national coordination mandate on online safety.