Australian Government | eSafetyCommissioner

eSafety submission to the Joint Select Committee on Social Media and Australian Society

eSafety Commissioner

21 June 2024

Contents

Introduction

As Australia's independent regulator, educator and coordinator for online safety, the eSafety Commissioner (eSafety) aims to safeguard Australians from online harms and to promote safer, more positive online experiences.

eSafety welcomes the opportunity to provide a submission to the Joint Select Committee on Social Media and Australian Society.

eSafety shares the goal of preventing all Australians from experiencing online harm and recognises particular considerations apply to children. However, it is important that any responses seeking to address online harm are:

- supported by, and based in, evidence;
- reasonable, proportionate and take a balanced approach to benefits and risks;
- practical and capable of being implemented and enforced;
- designed to promote transparency and accountability among industry;
- respectful of the rights of individuals, including the rights of children; and
- centre the voices of those with lived experience in the decisions that affect them, as part of a whole-of-community and multidimensional approach to online safety.

Overview of submission

This submission begins with an overview of eSafety's regulatory approach and insights. This is to contextualise and support our responses to the Terms of Reference. It is also to show that for individual online safety initiatives and interventions to be effective, they need to form part of a multilayered and holistic approach to online safety.

The submission then focuses on the following terms of reference (TOR), as they relate to eSafety's remit:

- TOR (a) - the use of age verification to protect Australian children from social media;
- TOR (d) - the algorithms, recommender systems and corporate decision making of digital platforms in influencing what Australians see, and the impacts of this on mental health;
- TOR (e) - other issues in relation to harmful or illegal content disseminated over social media.

Overall, our submission underscores the critical role of eSafety in ensuring the online safety of Australians.

eSafety's approach to online safety

The *Online Safety Act 2021* (Online Safety Act) sets out our legislative functions. This includes administering investigations schemes for four types of online harms: cyberbullying of children, cyber abuse of adults, the non-consensual sharing of intimate images, and illegal or restricted online content. The Online Safety Act also provides eSafety with some powers to regulate digital platforms' broader systems and processes.

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

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

Our regulatory approach is comprised of the three pillars of prevention, protection, and proactive and systemic change. These pillars reflect our broad and holistic legislative remit and are underpinned by our core mission of safeguarding Australians at risk of online harm.

An overview of these pillars is provided below.

Prevention

While eSafety acts as an important safety net for Australians online, our primary goal is to prevent online harms from happening in the first place. This work falls under our prevention pillar.

Through research, education and training programs, we aim to build the capacity of Australians to interact safely online. We seek to provide Australians with the practical skills and confidence to be safe, resilient and positive users of the online world, and to know where to seek help if issues arise.

eSafety is firmly of the view that education and capacity building is a lifelong journey that should begin as early as possible. This is because research has shown 87% of children aged 4 to 7 years use the internet – with 16% having used the internet without permission – and 40% of children in this age group have access to a personal device. This journey should continue well into the senior years, 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.

We know that online safety information needs to be easily accessible from a variety of sources and tailored to meet the needs of diverse groups. eSafety works closely with Australian communities to understand their online experiences. We use these insights to inform our initiatives and to co-design meaningful resources and support so that they can engage safely and confidently online. We undertake research and develop evidence-based education resources and community programs designed for specific audiences. This gives people of all ages and backgrounds the right tools to protect themselves online. This includes:

- parents and carers as the first lines of defence, particularly in the early years;
- educators and schools, to develop students' critical skills across the 4 Rs (respect, resilience, responsibility and reasoning), to manage online safety incidents that may arise within the school community, and to support best practice online safety education;
- domestic, family and sexual violence frontline workers, to upskill people who support those experiencing technology-facilitated abuse; and
- specific diverse communities that our research shows are more likely to experience online harms.

We are also heightening our efforts to reach and engage with young people in a way that resonates with them. This includes by developing a youth engagement strategy, as well as supporting the eSafety Youth Council (discussed further below).

Over the last several years, we have started to see evidence of real change to behaviours and attitudes. For example, we see children and young people taking multiple actions and accessing a range of tools and tactics in response to negative experiences. eSafety's 2021 survey of 3,600 young Australians aged 8-17 years-old found that 64% of young people who have experienced negative online behaviour blocked or unfriended people who had bullied them online – a significant increase on 46% of young people in 2017. The research also found that young people are increasingly reaching out to their parents and friends. 66% of young people who have experienced negative online behaviour told their parents (up from 55% in 2017) and 60% told their friends (up from 28% in 2017).

Industry has an important role to play in the prevention space. This includes 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.

eSafety both supports and regulates industry to achieve this. We provide industry with a range of supportive measures through the voluntary Safety by Design initiative. We also hold industry to account with our industry codes and standards, which require online services to do much more on this front. For example, the Social Media Services Online Safety Code (Class 1A and 1B Material) (**SMS Code**) requires that certain social media services must provide easy-to-use tools that can help safeguard the safety of a young Australian child using the service, and information regarding the safety tools and settings, in a manner that is understandable to users of all ages allowed on the service.

Protection

Where online harm does occur, eSafety offers tangible, rapid assistance. This work falls under our protection pillar.

Our individual complaints mechanisms allow us to investigate and take action to remove certain types of content. eSafety administers investigations schemes to address:

- cyber abuse material targeted at Australian adults (adult cyber abuse);
- cyberbullying material targeted at Australian children (cyberbullying);
- non-consensual sharing, or threatened sharing, of intimate images (image-based abuse); and
- Class 1 and Class 2 material such as child sexual exploitation or pro-terror material (illegal and restricted content).

These schemes cover both natural and synthetic (including artificial intelligence (AI)-generated) online content.

Where the online material reported meets the relevant legislated thresholds, eSafety can issue a removal notice to the online service provider where the content is available and a removal to the hosting service provider that hosts the content. In some instances, other enforcement options are also available, as set out in our Compliance and Enforcement Policy.

eSafety works with online service providers to get quick and positive outcomes for victim-survivors of online harm. We will often approach online service providers informally to ask them to remove adult cyber abuse, image-based abuse or cyberbullying material in the first instance, as this generally results in faster removal of material compared to formal actions.

We also provide broader support for complainants. This can include making referrals to law enforcement or mental health providers, or suggesting other options that may be helpful, such as legal services. It may also include providing practical tips and strategies for how to mitigate further harm. This helps reduce ongoing trauma and re-victimisation.

The Online Safety Act commenced on 23 January 2022. The success of our reporting schemes in that time shows the important safety net eSafety provides.

Adult Cyber Abuse Scheme

During the period 23 January 2022 to 12 June 2024, eSafety received 6,695 complaints related to adult cyber abuse. For those complaints, eSafety issued 1,170 informal complaint alerts to platforms. Of those, 77.93% were removed by the platforms.

Children's Cyberbullying Scheme

During the period 23 January 2022 to 12 June 2024, eSafety received 5,200 cyberbullying complaints. For those complaints, eSafety issued 1,621 informal removal notices to platforms. Of those, 83% were removed by the platforms.

Image Based Abuse Scheme

During the period 23 January 2022 to 12 June 2024, eSafety received 18,489 image-based abuse reports. For those reports, eSafety requested removal in relation to 9,495 locations, generally URLs, with 89% of the reported locations being successfully removed.

Online Content Scheme

During the period 23 January 2022 to 12 June 2024, eSafety received 27,486 complaints about illegal and restricted content at 70,749 URLs. 84% of these complaints were about the online availability of child sexual exploitation and abuse material. Nearly 61,000 investigations into child sexual exploitation and abuse were completed, with over 90% of the material referred within 2 days to the global network International Association of Internet Hotlines (INHOPE) for removal.

The Online Content Scheme is discussed further under our response to the Terms of Reference.

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, eSafety knows we need to be at the forefront of anticipating, mitigating and responding to online harms. This work falls under our proactive and systemic change pillar.

We are operating in an extremely complex and interlinked technology ecosystem.

This requires sophisticated systems-thinking to address online harms at-scale. The online world is borderless, with most of eSafety's regulatory targets based overseas. We are also working with many government organisations across the world that have some interest in, or remit over, the technology sector.

We have a range of measures to address online harms systemically. We outline some of the key ones below.

Basic Online Safety Expectations

The Online Safety Act allows eSafety to require reports on services' implementation of the Basic Online Safety Expectations. These reporting powers aim to improve the transparency and accountability of online services. They have the goal of incentivising improvements in the safety of services on a broad range of harmful and unlawful material and activity. The obligation for services to respond to a reporting requirement is enforceable and backed by civil penalties and other enforcement mechanisms. Information obtained from the reporting notices are published on the eSafety website in transparency summaries, where appropriate.

The Basic Online Safety Expectations Determination establishes expectations that service providers will take reasonable steps to:

- ensure that end-users are able to use the service in a safe manner (s6(1)), to proactively minimise the extent to which material or activity on the service is unlawful or harmful (s6(2)), and to ensure that the best interests of the child are a primary consideration in the design and operation of any service likely to be accessed by children (s6(2A));
- minimise certain material, including Class 1 material and material that depicts abhorrent violent conduct (s11), and provide mechanisms to report and make complaints about this material (s13); and
- ensure they have terms of use, policies, and procedures in relation to the safety of end-users, as well as policies and procedures for dealing with reports and complaints, and that they detect and respond to breaches of these terms of use, policies and procedures (s14).

While the Basic Online Safety Expectations are not enforceable, eSafety has used its reporting powers, which are enforceable, to shine a light on what industry is, and is not, doing to implement the expectations and keep Australians safe online.

We have used these reporting powers to issue 19 notices, covering 30 services, to date.

The Basic Online Safety Expectation Determination was registered in 2022 and was amended by the Minister for Communications in May 2024.

The Basic Online Safety Expectations are discussed further under our response to the Terms of Reference.

Industry codes and standards

Industry codes and standards also apply to Class 1 and Class 2 material. They are designed to protect Australians from illegal and restricted online content, by setting obligations for relevant industry sectors to proactively deal with this material at a systemic level. They cover eight sections of the online industry and aim to drive systemic change to help address the harms of illegal content.

eSafety has the power to request the creation of, and then register, industry codes and/or develop standards for eight sections of the online industry outlined in the Online Safety Act.

Like our four investigations schemes, the industry codes and standards apply to online content whether or not it is generated by AI, though some do have requirements specific to generative AI services.

The industry codes are being developed in two phases. The first phase targets the most seriously harmful Class 1A and 1B material, which includes child sexual exploitation material, pro-terror material, extreme crime and violence material, and drug-related material.

To date, the eSafety Commissioner has registered six industry-developed codes for the first phase, which apply to social media services, app distribution services, hosting services, internet carriage services, equipment providers, and search engine services.

eSafety declined to register two codes developed by the industry associations. We have drafted industry standards for Relevant Electronic Services (such as messaging and gaming services) and Designated Internet Services (such as websites offering generative AI functionality which meet a certain threshold, online file and photo storage services, and websites offering high-impact material such as pornography). We have closely considered submissions received during public consultation and will be finalising the standards shortly.

In the future, there will also be industry codes or standards for high impact material, such as online pornography.

Industry codes and standards are discussed further under our response to the Terms of Reference.

Tech Trends

As part of our work as an anticipatory regulator, eSafety has a Tech Trends workstream where it conducts horizon scanning and works with subject matter experts to identify emerging technology issues and their impact on online safety. This allows eSafety to identify the online safety risks and benefits of emerging technologies and online behaviours, as well as the regulatory challenges and benefits they may present.

We have developed several position statements outlining the online safety and regulatory implications of emerging technology and issues, including the impacts of recommender systems, generative AI and immersive virtual reality environments like the metaverse. They summarise eSafety's approach to these systems, including proactive safety measures for industry and advice for users.

Safety by Design

Recognising that user safety should be at the forefront of design, we work with tech companies to shift their design ethos to anticipating, detecting and eliminating online threats before they occur. This is our Safety by Design initiative.

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

The initiative promotes online safety through three guiding principles:

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

A Safety by Design approach can seek to address numerous online safety issues by promoting a proactive approach to user safety. For example, it could include:

- having individuals or teams accountable for the creation, evaluation, and implementation of relevant policies;
- putting tools and processes in place for detecting and actioning content that violates these policies;
- ensuring that community guidelines and reporting processes are accessible and easy to understand;
- carrying out open and meaningful engagement with a wide user base including diverse and at-risk groups, independent experts and other key stakeholders;
- committing to consistently innovate and invest in safety-enhancing technologies;
- publishing information about safety tools, policies, and processes, and their impact and effectiveness; and
- ensuring that design features and functionality preserve fundamental user and human rights.

Safety by Design requires cultural change in the tech industry and concerted prioritisation and leadership from tech CEOs. It also requires pre-emptive action by industry to harden technology platforms and services against known and anticipated online harms and risks.

To help industry, we have developed practical resources via the Safety by Design assessment tools. This includes educative content on intersectional risk factors for online harms, insights into perpetrator motives and exploration of human rights in the digital context.

Informed by evidence and the voices of children

eSafety undertakes research to ensure our programs and resources are based on evidence.

We also recognise the importance of ensuring the voices and perspectives of those with lived experience inform the policies and programs that affect them.

This is part of our evidence-based and co-design approach.

The discussion on children's access to social media raises important questions. This includes the effects of social media on young people's mental health. eSafety is acutely aware of the safety and wellbeing risks social media can pose, and that parents and carers – and Australians more generally – are deeply concerned about what can be done to mitigate these risks.

We are also aware that developing robust solutions requires listening to children and young people and understanding their life experiences.

We know through our work with our Youth Council that young people themselves recognise the complexity of the issue and are motivated to tackle it in partnership with policymakers across government, non-government and digital technology companies.

eSafety Youth Council

eSafety's Youth Council, which is comprised of young people between 13 to 24 years, does just this. It provides young people with the opportunity to share their insights and perspectives and for these to inform online safety measures that impact them. In April 2024, the Youth Council shared with us their insights about social media and young people's wellbeing. They observed that limiting social media use can be difficult. They also observed that some young people may struggle even with the concept of an online world, because online spaces are integrated into the rest of their everyday lives. This can make keeping boundaries around time online, including social media, challenging.

The Youth Council considered how regulation might empower young people to support their wellbeing while navigating social media. They observed that Big Tech companies that lack policies and regulations to protect the wellbeing of young users pose a barrier to the online safety of young people. They felt that Big Tech companies should be held accountable for the wellbeing of their users, through clear regulations about acceptable and unacceptable conduct, and that these regulations should be enforced. The Youth Council also thought that government should act to hold Big Tech accountable and that this should include conducting regular reviews of social media regulations.

The Youth Council also highlighted the role of education in supporting healthy social media use amongst young people. They made several recommendations, including:

- educators should be informed about evolving technologies and the harms they may pose to young people;
- frequent, school-led workshops to teach young people wellbeing skills that support positive social media use; and
- parents and caregivers should be educated about the harms of social media and how to protect their children and young people.

The Youth Council will be making their own submission to this inquiry.

Research and evidence

An evidence-based approach means being informed by robust and rigorous research.

We therefore want to outline what is known about the relationship between young people and social media – and just as importantly, what is not known.

Research shows that most users' experiences of social media will be varied and multifaceted, with experiences ranging from neutral, to positive, to negative. For some young people, the harms may outweigh the benefits.

eSafety's research shows that 45% of Australian young people have reported being treated in a hurtful or nasty way online. Almost two-thirds of young people aged 14–17 were exposed in the

past year to negative content, such as content relating to drug taking, suicide or self-harm, or gory or violent material. 28% of young people aged 14-17 years were exposed to content that promoted unhealthy eating weekly or more often. This was significantly higher for females (35%) than males (19%).

Additionally, social media can pose immediate harms that can be facilitated online to young people. Acute risks and harms to children, such as online grooming, sexual extortion and the production of child sexual exploitation and abuse material, are increased where services do not build in Safety by Design.

However, it is imperative that a discussion about the risks of social media is balanced with a discussion of the benefits.

Social media may also provide a range of opportunities that are protective of mental health, such as inclusion, social connection and belonging. These benefits are especially important for young people who experience difficulties with participation and social inclusion in other contexts.

For example, eSafety's research into the online experiences of Aboriginal and Torres Strait Islander children, the digital lives of young people with disability, and our report on LGBTQI+ teens, highlights some of the ways online environments can help facilitate connection, support and cultural expression.

Our most recent research on young men shows that while young men's online experiences are marked by tensions, complexities and possibilities, overall the internet is a place they can explore and express their identities. It is also a source of community and belonging.

The relationship between mental health and social media is complex and the evidence base for this relationship is still evolving. The science is by no means settled.

Further, the impact of social media is not the same for everyone. Its effects accrue differently for different users. Young people vary substantially in how they use social media. This includes the platforms they access, the digital features they are exposed to, the content they consume and the communities they engage with.

The subsequent impact of online experiences or use is also highly individualised. It is influenced by a range of individual, familial, social and contextual factors and vulnerabilities. Consequently, restrictive measures that may benefit one child may be ineffective, or even harmful, for another.

This makes decisions about preventing or limiting children's access or participation online incredibly complex. These decisions require thorough consideration of solutions that do not inadvertently introduce negative outcomes.

Additionally, most of the evidence and recommendations available are based on international research. There is a need to review and weigh the Australian evidence base and consider the extent to which international evidence and advisories are generalisable to the Australian context.

Ultimately, the rights of a child include rights in relation to the digital environment and it is imperative this right is respected, protected and secured in discussions about their access to social media.

Age verification to protect Australian children from social media

Understanding social media

The discussion about preventing children from accessing social media through age verification implies social media is a discrete form of media that can be separated from the rest of the internet and modern media.

But social media forms part of a converged and integrated contemporary media environment. There is a fluid interplay between social media, websites, messaging apps, gaming platforms, dating apps, as well as ephemeral media.

For example, the harmful messages of many of the most controversial social media influencers can just as easily be found through an internet search as they can on social media.

Social media also influences, and is influenced by, a range of other media sources, including news, popular culture, marketing and advertising.

Even if social media could be demarcated and separated from other media, a primary concern is that children would migrate to other services and platforms with fewer safeguards. As outlined below, these services and platforms may not have the safety features of social media platforms or the beneficial information on these platforms.

Children accessing social media

Most major social media platforms require users to be at least 13 years to create an account. This includes Instagram, TikTok, Snapchat, Facebook, Pinterest and Twitter/X.

However, despite this age restriction international research and eSafety's own data, including insights from our investigation schemes, suggests that existing measures are not effective in preventing children under 13 from accessing social media.

Research from the UK communication services regulator Ofcom (2024) found that half of children aged 3 to 12 years of age use at least one social media app or service. The percentage of children using social media rises with age. One third of children aged 3 to 7 years (34%) use social media and this rises to 6 in 10 of eight-year-olds (63%) and most 12- to 15-year-olds (92%). This would also indicate that parents and carers are facilitating children under 13 to access social media.

Challenges of age-restrictions

A particular concern for eSafety is that restriction-based approaches may limit young people's access to critical support. This would potentially compromise the help-seeking responses that are protective of mental health and wellbeing outcomes. Research shows that the benefits of the online world can be especially profound for those who may not otherwise have support or resources available to them offline. It can be a critical source of health information and emotional support, as well as a place for fun, learning and community.

Additionally, if age-based restrictions are imposed, eSafety has concerns that some young people will access social media in secrecy. This may mean that they access social media without adequate protections in place and are more likely to use less regulated non-mainstream services that increase their likelihood of exposure to serious risks. Restriction-based approaches can also reduce young people's confidence or inclination to reach out to a trusted adult for help if they do experience harm, which is a key protective factor for safer internet use. When young people experience online abuse and are not able to reach out for support, they can experience heightened distress and greater long-term negative impacts.

Banning children of a certain age also doesn't work to build the capacity of young people to engage online safely. Bans also place the onus on children to keep themselves safe, rather than putting the onus on online platforms and services to keep young people safe.

This is a complex challenge requiring multiple, complementary solutions. Alongside examining the effects of technology and social media, we must also understand the varied and complex drivers of youth mental health. To that end, we have recently met with the National Mental Health Commission, which has recently completed its consultation on digital technologies and youth mental health and will commence a review of the Australian evidence this year. We understand that this will build on the work of inquiries conducted internationally and look to situate the evidence and subsequent recommendations in the Australian context.

eSafety is committed to ensuring all young people are having online experiences that are safe and promote their health and wellbeing. We remain focused on empowering young people with the essential knowledge and skills they need to do this. We provide advice and resources that build online safety awareness, and promote young people's resilience, respect, reasoning and responsibility. We also provide resources and guidance that help parents and carers to recognise and cultivate these skills in their children.

Age verification and age assurance

Age assurance is an umbrella term referring to various methods which are used to determine a person's age. These methods offer different levels of certainty. They can be used establish or predict the age (or age range) of an individual.

Age verification determines a person's age to a high level of certainty, typically by verifying data against an external source. Examples include government ID.

Age estimation measures infer an approximate age or age range without other confirmed sources of information about the individual. This can involve the use of biometric data such as facial scans or other information such as behavioural patterns to estimate a person's age or age range.

eSafety's work on age assurance

In March 2023, eSafety submitted to the Australian Government for consideration an Age Verification Roadmap (Roadmap) that addresses how to prevent and mitigate harms to children

from online pornography. eSafety presented the accompanying background report in August 2023.

The request to develop the Roadmap formed part of the then Government's response to the House of Representatives Standing Committee on Social Policy and Legal Affairs report, 'Protecting the age of innocence'. eSafety was tasked with considering if, and how, a mandatory age verification mechanism (or similar) could practically be achieved in Australia, as well as providing additional advice on legislative, regulatory and complementary measures to government.

The Roadmap was informed by eSafety's research, an independent technical assessment and consultations with stakeholders. It made recommendations for a possible way forward, including complementary measures for a holistic approach to children's safety online. These included:

- a call for research to better understand the experiences of younger children online;
- that the use of age assurance, if implemented, should be underpinned by robust international standards, a diverse market to enable consumer choice, and a framework for accreditation and oversight; and
- that addressing children's access to age-restricted content requires a digital-ecosystem response, which takes into account the roles of other services – such as search engines – in facilitating access to sites that may not have protections in place.

It is important to note that the Roadmap addressed children's access to pornography, which is content already legally restricted to people aged 18 or over. It did not relate to social media and its findings, as well as the findings of other age-restricted goods such as gambling and alcohol, cannot be directly applied to social media or age restrictions other than 18 or over.

This means that applying age assurance to restrict services to different age thresholds will require additional specific research and investigation.

For example, age assurance measures that rely on government identity information, such as driver's licences, may not be accessible for younger users. Other methods, such as facial age estimation, have limited independent testing for younger age groups.

Draft guidance released by Ofcom on protecting children online considered the use of age assurance for enforcing minimum ages on platforms. Ofcom determined that its use would not be proportionate, given limited independent evidence of age assurance distinguishing between children of different age groups and that other methods, such as passport matching, may not be accessible to children. It considered this could result in a serious impact on children's ability to access these services.

Age assurance can be an important tool in making children's online experiences safer and more age appropriate. However, it does not work on its own. Once a service determines the age of users, it then needs to provide a safe and age-appropriate experience using a range of complementary safer design practices, tools and policies. For example, if a service knows the age or age range of users, it can:

- apply age-specific default safety and privacy settings;
- adjust moderation or recommender algorithms;
- not collect users' data where they are too young to consent to its collection; and

- offer age-specific services such as groups, channels or functions designed specifically for teenagers.

Tech solutions must also go hand-in-hand with education for children and their parents or carers. As the Roadmap noted in the context of online pornography, age assurance is not a panacea for online harms. Educational and awareness-raising activities should target specific audiences with tailored content. Broadly, educational measures should:

- Support children and young people to critically think about the content they see online, including how to seek help and help themselves when they see unwanted content that makes them feel uncomfortable.
- Address some of the specific harms associated with the content sought to be restricted, so that children and young people have an enhanced capacity to engage critically with that type of content.
- Equip trusted adults, such as parents, carers, educators, frontline workers and others who work with and support children and young people, with the skills and knowledge to have conversations with children about the content they see online. Education should also cover how age assurance tools work, what measures are in place to protect user safety, privacy, and security. Parents and carers should be informed on how to access and apply safety technology and tools, such as parental controls and search filters.

Trial of age assurance technology

On 1 May 2024, the Government announced $6.5 million as part of the 2024-25 Budget to conduct a trial of age assurance technology to protect children from age-restricted online services.

The trial is being conducted by the Department of Industry, Transport, Regional Development, Communications and the Arts (DITRDCA).

The trial will occur in parallel with the development of the Phase 2 industry codes overseen by eSafety, which seek to address the access and harms of high-impact, but not illegal, material such as pornography. We expect these to be mutually reinforcing processes. eSafety will have visibility of the trial as a member of the government working group for the age assurance trial. The outcomes of the trial could support the development of guidance for industry about what steps are reasonable and appropriate to comply with age assurance expectations under the Basic Online Safety Expectations, as well as the assessment of the community safeguards in Phase 2 industry codes.

Algorithms, recommender systems and corporate decision making

An **algorithm** can be described as 'any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output'. Algorithms are fundamental computing instructions, are often considered inherently neutral, and have long been used to aid decision-making online and offline. However, recent developments such as widespread adoption of machine learning and an exponential increase in available data have created ever-more sophisticated and complex algorithmic decisions.

A **recommender system** is a system that prioritises content or makes personalised content suggestions to users of online services. A key component of a system is its **recommender algorithm**, the set of computing instructions that determines what a user will be served based on many factors. This is done by applying machine learning techniques to the data held by online services, to identify user attributes and patterns and make recommendations to achieve particular goals.

Algorithms used for **content moderation** and **targeted advertising** are also relevant to online safety. Information on these algorithms, and further information on algorithms and recommender systems generally, can be found in eSafety's position statement and the literature summary prepared by members of the Digital Platforms Regulators Forum (DP-REG).

Recommender systems

Recommender systems and their underlying algorithms are built into many online services. They sort through vast amounts of data to present content that is relevant to users. To personalise content to users, the algorithms analyse user data such as their interactions on the platform, and sometimes the interactions of their contacts and the user's off-platform activity.

Recommender systems are especially relevant in the context of social media and online harms because they directly affect the content users are shown. They can reduce or exacerbate online harms. For example, they can help to identify and downrank or filter out abusive and harmful material and bad actors. However, they can also present risks. This includes the potential to amplify harmful and extreme content, especially where they are optimised for user engagement.

In December 2022, we released a Tech Trends paper on recommender systems and algorithms, which summarises eSafety's approach to these systems, including proactive safety measures for industry and advice for users. Along with Australia's other digital platform regulators, we also contributed to a literature summary on the harms and risks of algorithms and their relevance to each regulator's remit.

Benefits and risks

Recommender systems could have beneficial impacts for mental health. For example, they could suggest content to a user that enhances their happiness and wellbeing, reaffirms their identity and sense of belonging or provides help-seeking information.

However, they also have the potential to amplify harmful and extreme content, which can have significant individual impacts on those exposed to such material. On a broader societal level, the amplification of discriminatory content can promote and reinforce systemic inequality. They can do this by amplifying some voices at the expense of others, which diminishes inclusion within online spaces and the diversity of voices in public discourse.

People's experience of algorithms and their harms and benefits varies depending on intersectional factors, including the individual user's circumstances. Whether certain content is harmful typically depends on the specific context, such as the identity or other relevant factors pertaining to the person who encounters it. For example, content that promotes self-harm is likely to present a greater risk and have deeper impact for someone already experiencing or contemplating self-harm. This can make it challenging to prevent specific instances of harm or to assess the severity of harm caused by a recommender system.

It is therefore important that user safety is addressed from an intersectional perspective at all stages of the development and deployment of algorithms.

Online services may optimise their recommender algorithms for different purposes, such as:

- maximising user engagement through likes and comments or further queries;
- delivering recommendations that best meet its users' needs;
- maximising the time users spend on its platform; or
- a combination of all of these.

Recommender systems that optimise for user engagement require particular care. If adequate safety steps are not taken, these systems may facilitate access to increasingly extreme content. In some instances, this may involve exploiting or exacerbating psychological harm or mental ill-health.

There are also concerns about the potential of recommender systems designed to maximise engagement to exploit and exacerbate **excessive use** of online services. Certain groups, such as children, older people, people with learning disabilities, and people with addictions may be especially vulnerable. Excessive internet use refers to a state where individuals "lose control" and continue using the internet despite experiencing negative outcomes. Some studies suggest recommender systems drive excessive usage of video websites.

Educational response

Recognising the importance of enhancing digital literacy and giving children and young people the skills and confidence to manage their online experiences, eSafety is developing education and professional learning programs to help raise awareness of the potential risks of algorithms and the tools to manage them.

Critical digital literacy is important for a number of reasons, including:

- knowing algorithms exist and how they work;

- being able to identify instances of when they are being targeted;
- understanding how an algorithm may be shaping an individual's experience, beliefs, feelings, or reasoning;
- helping give people greater influence over their recommendations; and
- equipping someone with skills to have a role in their choices around media.

eSafety's approach to critical digital literacy includes incorporating relevant guidance into our universal prevention programs. This includes our professional learning, Best Practice Framework for Online Safety Education, Best Practice Framework, Toolkit for Schools, classroom resources, webinars for parents and carers and outreach through our Trusted eSafety Provider program.

We also have special resources for at-risk groups, especially children.

eSafety's research team is currently developing questions on algorithmic literacy to include in our 2024 youth survey. This research will inform our online safety programs and contribute to the international evidence base on children and young people's digital literacy.

Regulating algorithmic harms

eSafety has a number of tools at our disposal to promote platforms' transparency and accountability in relation to how algorithms may mitigate or contribute to risks to users' online safety.

In February 2023, we issued Basic Online Safety Expectations reporting notices to Google, TikTok, Discord, Twitch and Twitter (subsequently X). The notices focused on child sexual exploitation and abuse and included questions in relation to the safety of recommender systems.

We published a summary of the findings, including detailing industry's responses regarding the design objectives of their recommender systems and the signals used to prioritise content for users. These findings provided important insights into services' priorities and design of their recommender systems.

On 21 June 2023, we also issued a separate reporting notice to Twitter (subsequently X), requiring it to explain what it is doing to minimise online hate, including how it is enforcing its terms of use and hateful conduct policy. A summary of the response was published on 11 January 2024. Notably, the notice included questions related to the potential of recommender systems to amplify harmful content. The response showed that as of May 2023, no tests were conducted on Twitter (subsequently X) recommender systems to reduce the risk they could amplify hateful conduct.

The amended Basic Online Safety Expectations Determination includes new expectations that if a service uses recommender systems, the provider will take reasonable steps to:

- consider end-user safety and incorporate safety measures in the design, implementation and maintenance of recommender systems on the service (s 8A(1)); and
- proactively minimise the extent to which recommender systems amplify material or activity on the service that is unlawful or harmful (s 8A(2)).

Reasonable steps could include risk assessments, educational or explanatory tools, enabling complaints about the way recommender systems present material, and enabling end-users to opt out of receiving recommended content.

The Internet Search Engine Services Online Safety Code (Class 1A and Class 1B Material) requires search engine services to review and improve the effectiveness of their algorithms, or other models that recommend indexed links or images, to prevent the discoverability of illegal and restricted material in search results. It also contains measures aimed at reducing the safety risks to end-users concerning synthetic materials generated by AI that may be accessible via internet search engine.

Transparency of algorithms and Safety by Design

There are significant practical, technical and regulatory challenges in explaining the functionality of complex algorithmic decision-making systems and their rationale in specific cases. Algorithms will also change over time to respond to user needs, economic opportunities, regulation and public sentiment.

The most sustainable approach to enabling oversight and empowering people to make informed choices is for creators to have meaningful transparency in relation to what they can do and are intended to do.

Similarly, the complexity of algorithms, and the intersectional and individualised nature of the risks and harms they present, mean it is crucial that industry reduces risks from the earliest stages and throughout the algorithm's whole lifecycle, by implementing Safety by Design measures and interventions.

A Safety by Design approach to minimising the risk of harm from algorithms could include:

- adjusting algorithms to focus on more quality-focused metrics instead of, or in addition to, engagement that favours shocking and extreme content, informed by appropriate consultation, public scrutiny and testing;
- offering users alternative curation models for their news feeds, such as a reverse chronological news feed for greater visibility and context of news developments. These should be accessible and simple to use;
- providing greater choice, control and clear feedback loops for users. Platforms can empower users to explicitly shape their algorithms. For example, by flagging types of content that they do not want to see in suggested posts;
- establishing and enforcing content policies, and actively moderating harmful content to ensure that a platform's pool of available content meets a baseline threshold;
- introducing human reviews as a circuit breaker for content on the path of being amplified, before potentially harmful material can go viral;
- introducing additional friction through design features, such as prompts and educative nudges around whether content has been read and restricting how often content can be shared;
- labelling content as potentially harmful or hazardous;
- introducing behavioural cues and prompts that can help users to establish positive patterns of behaviour, such as those that help users reconsider posting harmful content, or to manage their time spent online;
- engaging with at-risk community groups through consultation to better understand what labels and keywords users may be using to exploit and cause harm.

The complex, evolving, and dynamic nature of algorithms in the online environment means there is no single, fixed regulatory approach to address their potential benefits and harms. eSafety sees our work as forming part of a broader, multi-faceted response by government, industry and the public, which will require ongoing attention.

Algorithmic transparency and regulation are being considered internationally, including in the European Union, where the European Centre for Algorithmic Transparency (ECAT) contributes scientific and technical expertise to assist in enforcing systemic obligations under the Digital Services Act. Work is also being undertaken by transparency working groups convened through cross-sector initiatives. such as the Christchurch Call and the Global Internet Forum to Counter Terrorism. eSafety is keeping a watching brief on these developments and engaging in regular cross-jurisdictional and multi-disciplinary dialogue with our international counterparts and colleagues.

The current review of the Online Safety Act, discussed further below, is considering whether additional arrangements are warranted to address potential online safety harms raised by technologies like recommender systems. If it is determined that the Online Safety Act needs to have stronger protections in relation to algorithms, consideration should be given to the requisite powers and technical resources required to enforce them.

Corporate decision making

Online services have an obligation to make sure their services are safe. It is imperative that they prioritise user safety when making business decisions.

To better serve those who use their platforms, products and services, tech companies need to ensure that creating, delivering and capturing value includes assessing the impacts their products and services have on the safety of individuals and society.

For example, online services with advertising-based revenue models are incentivised to increase engagement, as the more time users spend online, the more advertising revenue is generated.

However, the systems designed to recommend content with the purpose of increasing engagement run the risk of exploiting human cognitive biases, inherently drawing people to shocking and extreme content. These types of harms need to be anticipated, detected and eliminated before they occur through investment in risk mitigation at the front end, and a commitment to embedding user protections from the get-go.

There are important inflection points in the technology ecosystem that need to be leveraged to enable real change and embed safety into the culture and leadership of organisations. Our Safety by Design initiative explicitly recognises these linkages. As well as providing principles and resources for the online industry, we have worked with investors, venture capitalists and start-up and incubation communities to develop a suite of resources for investors and financial entities, who can play a pivotal role in nurturing tech ventures.

Our Safety by Design initiative can be applied to platforms and services of all size and structure. The first Safety by Design principle is service provider responsibility: the burden of safety should never fall solely upon the user.

This recognises the inherent power imbalance that exists between digital platforms and the individual user.

Platforms and services can, and must, take preventative steps to help ensure that known and anticipated harms have been evaluated in the design and provision of an online service. This needs to occur alongside steps to make services less likely to facilitate, inflame or encourage illegal and inappropriate behaviours. Every attempt must be made to ensure that online harms are understood, assessed and addressed in the design and provision of online platforms and services.

Our best practice guidance on this principle emphasises the role of dedicated 'trust and safety' teams, whose responsibilities include, but are not limited to:

- responding to and investigating safety situations;
- developing, implementing and enforcing product and content policies;
- managing connections with law enforcement to resolve incidents and handle legal requests;
- influencing product decisions to ensure user safety and user experience is fully considered; and
- driving awareness of safety features and functions to the user base.

These teams need to be adequately resourced and empowered to carry out their responsibilities effectively.

Regulation and decision making

In May 2024, the Government amended the Basic Online Safety Expectations Determination to highlight that reasonable steps to ensure safe use could include assessing and mitigating the adverse impact of business decisions on the ability of end-users to use the service in a safe manner. The Explanatory Statement to the amended Determination highlights decisions that should be assessed for their safety implications.

eSafety has used our reporting powers to understand the resourcing of online service providers and the impact this may have on the safety of their services.

Notices given in 2023 found that the number of languages covered by services' content moderators varies considerably. Google informed eSafety that it has content moderators in 71 languages, with TikTok stating that it covers 73. On the other hand, we learnt that Discord covers 29 languages, Twitch 24 and Twitter (subsequently X) covers 12.

Notably, we found that some of the most widely spoken languages in Australia were not covered by default by Discord, Twitch and Twitter. For harms like grooming and online hate, context and linguistic understanding is essential for accurately identifying harms at scale.

eSafety also issued a reporting notice to Twitter (subsequently X) on the steps it was taking to enforce its hateful conduct policy, as well as the impact of recent reductions in staffing. We published the findings in a report in January 2024. This found that Twitter had:

- an 80% reduction in engineers focussed on trust and safety issues globally, following Twitter's acquisition in October 2022 (279 to 55); and
- the reduction coincided with a 75% increase in Twitter's median time to respond to user reports of hateful conduct in Direct Messages, and 20% for Tweets.

Industry Codes and standards also require providers to make certain business decisions to enable safety. For example, the SMS Code requires certain services to:

- adequately resource the trust and safety functions to oversee the safety of the service, operationalise the requirements of the SMS Code and ensure training for staff handling reports of illegal and restricted material;
- invest in continuous improvement in capacity to detect child sexual abuse material and pro-terror material and enforce their policies concerning class 1A and 1B material; and
- annually review the adequacy of its processes for responding to reports of and removing illegal material.

From our work, we know enhanced regulatory tools that enable accountability and transparency are necessary levers to better understand industry activity and enhance safeguards for users.

Illegal and restricted content

Harmful online content is an umbrella term that captures a wide variety of harmful content. At the more serious and extreme end, is illegal and restricted content. eSafety has a number of measures to deal with this content, which is divided into the following categories:

- **Class 1 material** includes material that would be likely to be refused classification under the National Classification Scheme. This includes online content that depicts child sexual abuse and exploitation material, advocates terrorist acts, or promotes, incites, or instructs in matters of crime or violence.
- **Class 2 material** includes material that would likely to be classified R 18+ or X 18+ under the National Classification Scheme. This includes adult pornography and high-impact material that may be inappropriate for children under 18 years old, such as material featuring violence, crime, suicide, death, and racist themes.

To demonstrate how eSafety's regulatory schemes work systematically together to address online harm, we'll outline how three of our regulatory levers:

- the Online Content Scheme;
- industry codes and standards; and
- the Basic Online Safety Expectations

apply to three forms of illegal and restricted content:

1. child sexual exploitation and abuse;
2. violent extremist material; and
3. age-restricted content.

Child sexual exploitation and abuse

Online Content Scheme

Under the Online Safety Act eSafety can issue removal notices in relation to Class 1 material to the online service where the content is available and to the hosting service provider that hosts the content for the service.

For Class 2 material that is (or is likely to be) X 18+ content and provided from Australia, we can give a removal notice to the online service or the hosting service provider of the service, requiring the recipient to remove the material. Class 2 material that is (or is likely to be) R 18+ content and provided from Australia is required to be placed behind a restricted access system (RAS) to prevent children under the age of 18 from accessing it. If a RAS is not in place, we can give the online service or its hosting service provider a remedial notice to remove or restrict access to the material. Complaints relating to this type of content are very rare and to date eSafety has not issued a Class 2 removal or remedial notice.

eSafety usually approaches online service providers informally in the first instance to ask them to remove Class 1 or Class 2 material. We know from our experience that informal requests often lead to faster removal of the material compared to formal action, resulting in fewer Australians being exposed to harmful online content.

eSafety's powers under the Online Safety Act provide further enforcement options, if needed. However, there are limitations. For example, while we can issue removal notices to providers of online services, we cannot issue Class 1 removal notices to end-users. Further, while our powers are effective in facilitating removal of specific items of content such as web pages, threads, posts, images and videos, our powers do not extend to removing whole websites, domains or user accounts.

The vast majority of reports to eSafety's Online Content Scheme relate to child sexual exploitation and abuse material. Very few reports relate to content that advocates terrorist acts or material that incites or instructs in matters of crime or violence.

For the financial year to date (1 July 2023 to 31 May 2024) we have received 12,525 complaints about illegal and restricted content at 30,577 URLs. 82% of these reports related to child sexual exploitation and abuse material.

Child sexual exploitation and abuse material hosted within Australia is notified to the Australian Federal Police (AFP) and removed by eSafety via removal requests or notices. Online child sexual exploitation and abuse matters are notified to the Australian Centre to Counter Child Exploitation (ACCCE), where perpetrator and/or victim information is identified. We also send intelligence and information reports through the ACCCE to assist law enforcement in Australia and overseas.

The child sexual exploitation and abuse material we investigate is nearly always hosted overseas. In order to enable takedown of this content, we refer it to the International Association of Internet Hotlines (INHOPE), a global network comprising more than 50 organisations around the world which facilitates the rapid removal of child sexual exploitation and abuse material across jurisdictions. 74% of all child sexual exploitation and abuse reports shared between INHOPE members result in takedown within three days.

eSafety is the Australian member of INHOPE. Our membership of INHOPE and strong relationships with law enforcement allow us to make a significant contribution to global efforts to combat the online availability of child sexual exploitation and abuse material.

Industry codes and standards

For industry codes, examples of measures that certain social media services must now take under the SMS Code to prevent and address child sexual exploitation material include:

- Proactively detecting known child sexual abuse material on the service, including employing systems, processes and/or technology which automatically detect known child sexual abuse material, prevent the posting or sharing of material or known links to material and identifying phrases commonly linked to child sexual abuse material.
- Removing identified child sexual abuse material on the service and taking action against accounts sharing such material, including termination and taking steps to prevent the person from making a new account.

- Investing in systems processes and/or technologies that aim to detect, disrupt and/or deter users from using the service to create, post or disseminate child sexual abuse material.
- Providing reporting facilities to allow users to report illegal content and investigating all such reports.

Basic Online Safety Expectations

Under the Basic Online Safety Expectations, eSafety has focused on child sexual exploitation and abuse as well as associated harms like sexual extortion in 12 of the 19 notices issued to date.

These notices have provided visibility on what industry is actually doing on child sexual exploitation and abuse. For example, eSafety found:

- If a user is banned on Facebook for child sexual exploitation and abuse, information is not always shared with other Meta platforms such as Instagram, and vice versa, in order to prevent the account from operating on the other service. Meta reported that WhatsApp information on child sexual exploitation and abuse is not shared with either Facebook or Instagram.
- Apple does not use hash matching tools (digital fingerprinting technology) to detect known child sexual exploitation and abuse images or video on iMessage or iCloud.
- Apple does not have an option for users to report child sexual exploitation and abuse in-service on iMessage, FaceTime or iCloud. Similarly, Google does not for Gmail and Google Messages.
- Microsoft Teams, Skype and OneDrive take a median time of 2 days to respond to user reports of child sexual exploitation and abuse, and up to 19 days for cases requiring re-review. This is the longest of any services covered by eSafety's notices.
- Snapchat does not use language analysis technology to detect likely grooming in Snaps or direct chat.
- Google does not block URLs to known child sexual exploitation and abuse on YouTube, Drive, Meet, Chat, Google Photos, Google Messages, Gmail or Blogger.
- Twitter (subsequently renamed X) does not use language analysis technology to detect likely grooming in tweets or direct messages, nor to detect other child sexual exploitation and abuse activity such as sexual extortion or the trading and sale of child sexual exploitation and abuse material on direct messages.
- Professional trust and safety staff at Discord are not automatically notified when volunteer moderators or administrators take action against child sexual exploitation and abuse material. This increases the risk of offenders continuing to abuse and re-victimise children on other parts of the service.

Since eSafety's notices, associated public scrutiny and our follow-up engagement, some providers have made specific improvements to their services to rectify gaps in their safety measures. This demonstrates the impact that transparency and accountability can have.

We will continue to use our reporting powers on child sexual exploitation and abuse and other harms, including through the future use of 'periodic' notices, which track key safety issues over time.

We have also used our enforcement powers in relation to a number of instances of non-compliance with reporting notices on child sexual exploitation and abuse. This includes an infringement notice given to X Corp (in relation to the Twitter service, subsequently renamed X) for its failure to comply with a reporting notice on child sexual exploitation and abuse in 2023. X Corp decided not to pay the infringement notice and applied for judicial review of the notice and the infringement notice. eSafety commenced civil penalty proceedings in December 2023. Separately, Google was given a formal warning for failing to adequately answer specific questions in a notice on child sexual exploitation and abuse.

We will continue to apply the compliance and enforcement tools we have in a proportionate way to deter providers from failing to comply with notices. We are also considering where additional powers may be needed as part of the Online Safety Act review.

Violent extremist conduct and material

Illegal and restricted content

The removal powers outlined above for class 1 material also apply in relation to violent and extremist material that is class 1 material.

Abhorrent violent conduct

Material that depicts abhorrent violent conduct includes material that depicts, promotes, incites, or instructs in 'abhorrent violent conduct', such as terrorist acts, murder, attempted murder, torture, rape, or violent kidnapping.

Terrorist and violent extremist material that is abhorrent violent conduct is addressed in the whole-of-government Online Content Incident Arrangement (OCIA) framework administered by the Department of Home Affairs.

The OCIA includes provisions relating to an Online Crisis Event (OCE), which gives eSafety:

- the power to declare an OCE when material that depicts abhorrent violent conduct is shared or spread online in a manner likely to cause significant harm to the Australian community, in circumstances warranting a rapid, coordinated, and decisive response by industry and government.
- when an OCE has been declared, blocking powers to help prevent the spread of this material. This allows us to leverage internet service providers to prevent access to the relevant material for a limited time by taking steps like blocking domain names, URLs, and IP addresses. This is intended to prevent the exposure to, and rapid distribution of, online material closely connected to the OCE, as occurred, for example, after the 2019 Christchurch mosque attacks.

To date, we have not declared an OCE nor implemented the Internet Service Provider blocking powers. The use of blocking powers is at the eSafety Commissioner's discretion and not all situations involving the spread of abhorrent violent conduct material during on OCE will require eSafety to take action. eSafety may choose to exercise its other powers under the Online Safety Act to require the removal of the material.

Abhorrent violent material

Under the *Criminal Code 1995* (Cth), eSafety may also give notices relating to abhorrent violent material (AVM), which is perpetrator-produced material that depicts abhorrent violent conduct. The notices do not require the AVM to be removed. However, if a service is later prosecuted for failing to remove or to cease hosting the AVM, the notice can be used in legal proceedings to show recklessness regarding the AVM. The AVM regime also gives powers to the Australian Federal Police to enforce certain offences.

Industry codes and standards

For industry codes, examples of measures that certain social media services must now take under the SMS Code to prevent and address pro-terror material include:

- Using systems, processes and/or technologies to detect and remove known (pre-verified) pro-terror material. This can include, for example, key word searches, machine learning detection or analysing against databases of indexed material (Hash matching).
- Implementing systems, processes, and technologies that enable the provider to take appropriate enforcement action against end-users who breach policies prohibiting pro-terror material and preventing recidivism.
- Providing tools which enable Australian end-users to report, flag, and/or make a complaint about pro-terror material accessible on the service.

Basic Online Safety Expectations

On 18 March 2024, eSafety gave reporting notices to Google, Meta, WhatsApp, Reddit, Telegram, and Twitter (subsequently renamed X) requiring them to report on the steps they are taking to tackle the risk of terrorist and violent extremist material and activity on their services. The notices require answers to questions about the tools, processes and resources they use to ensure safety.

eSafety will publish appropriate information on the findings to improve transparency and accountability.

Age-restricted content

Illegal and restricted content

For Class 2 material that is or is likely to be X 18+ content and provided from Australia, eSafety can give a removal notice to the online service or the hosting service provider of the service, requiring the recipient to remove the material.

Class 2 material that is or is likely to be R 18+ content and provided from Australia is required to be placed behind a restricted access system (RAS) to prevent children under the age of 18 from accessing it. If a RAS is not in place, we can give the online service or its hosting service provider a remedial notice to remove or restrict access to the material.

Complaints relating to this type of content are very rare and to date eSafety has not issued a Class 2 removal or remedial notice.

Industry codes and standards

eSafety has been engaging with industry about the Phase 2 Codes since November 2023 and we anticipate issuing formal notices for the production of Codes shortly after the conclusion of the Phase 1 Standards process. In addition to content like online pornography, Class 2 material also includes material containing high-impact themes.

Basic Online Safety Expectations

The Basic Online Safety Expectations expect services to take reasonable steps to ensure that technological or other measures prevent access by children to Class 2 material.

The Basic Online Safety Expectations Determination was amended in May 2024 to highlight that reasonable steps to implement this expectation could include the use of 'appropriate' age assurance mechanisms, and continually developing, supporting or sourcing, and implementing improved technologies and processes for preventing children's access.

Related matters

Responding to conduct

eSafety categorises risks of online harms across three categories:

1. Content, which are risks associated with harmful material that is shared online.
2. Contact, which are risks associated with unsafe exposure to people online.
3. Conduct, which are risks associated with exposure to people engaging in harmful behaviour online.

While our complaints schemes arguably focus primarily on content risks, eSafety is concerned with mitigating all three types of risks to better protect Australians online.

Contact and conduct risks of particular concern in the context of this inquiry include online grooming and sexual extortion.

Online grooming

Online grooming is a term used to describe the tactics used by perpetrators to sexually exploit children online. It involves harmful online conduct and is associated with unsafe online contact. An example of online grooming is when an older person tricks someone under 18 into thinking they are in a close relationship in order to abuse them. They may do this, for example, by participating in sexual conversations, sharing images or video, or even progressing to offline sexual activity.

eSafety research shows one in four young people have been contacted online by someone they don't know, and that in the past year, one in ten (11%) of teens aged 14-17 had been asked on the internet for a photo or video showing their private parts when they didn't want to. Further, the research indicates that 10% of Australian teens share personal information and 21% share photos and videos with people they only know online, which can be associated with a higher risk of grooming.

We are particularly concerned about the relationship between online grooming and 'self-generated' child sexual exploitation and abuse material. Self-generated child sexual exploitation and abuse material refers to explicit content that appears to have been taken by the child in the image or video. Research has shown that the creation of self-generated child sexual exploitation and abuse material can result from both consensual and coercive or manipulative experiences, such as sextortion, grooming and screen capturing.

Our investigators analysed a sample of the child sexual exploitation and abuse material they dealt with in the 2022-23 financial year and it indicated that 12% of URLs in the sample involved 'self-generated' material. Globally, the Internet Watch Foundation (IWF) has reported exponential increases in self-generated child sexual exploitation and abuse material in recent years. Most webpages actioned by IWF across all age groups now include self-generated child sexual exploitation and abuse material.

Given the increasing urgency of this issue, eSafety's Research Team is currently developing questions on self-generated child sexual exploitation and abuse material, sextortion and grooming to include in our 2024 youth survey.

The Basic Online Safety Expectations expect services to take proactive measures to minimise the extent to which material or activity on a service is unlawful or harmful. Harms such as grooming and sexual extortion are key harms covered. eSafety has prioritised these issues for its transparency notices over the last two years.

While the Social Media Services Code is focused on Class 1A and 1B content, it does contain a number of protections for children when using a social media service that are aimed at reducing the risk of grooming. Certain SMS providers have mandatory obligations to:

- have default settings to prevent young children from unwanted contact from unknown users and location sharing;
- take reasonable steps to prevent an Australian child that is known to be under the minimum age permitted on the service from holding an account on the service; and
- provide easy to use and accessible tools and functionality to help safeguard a child on the service and provide safety information on such tools for parents and children.

There are limitations in the capability of a service to deploy detection technology in relation to self-generated child sexual exploitation and abuse material that may result from grooming. The SMS Code requires that higher risk services take meaningful actions to disrupt or deter users from creating, posting or disseminating images that have not been pre-verified as child sexual exploitation and abuse material.

In addition, the eSafety Guide includes information on app features, such as chat functions and location sharing, that can increase the risk of exposure to online abusers. eSafety also provides general resources on child sexual abuse online, unwanted contact, as well as targeted resources on a range of relevant topics for parents, educators, young people and children.

Sexual extortion

Sexual extortion, often referred to as 'sextortion', is a form of blackmail that involves threatening to share an individual's intimate image or video online unless they comply with certain demands. Depending on the situation, these demands are typically for money or cryptocurrency, additional intimate images, or meeting for sex or other sexual acts. Perpetrators often target people through social media, dating apps or emails. Individuals may also be coerced by current or former partners in domestic, family or sexual violence situations.

Sexual extortion has increased substantially in recent years and remains a significant – and growing – issue.

Between 1 July 2023 and 31 March 2024, sexual extortion reports comprised 59% of total reports to eSafety's image-based abuse scheme, meaning sextortion was the most-reported form of image-based abuse. Boys and young men account for most of these reports. The recent increase in reports of sexual extortion, particularly amongst young men aged 18 to 24, was also noted in the Issues Paper for the Statutory Review of the Online Safety Act 2021.

eSafety's image-based abuse scheme enables us to respond to threats to share intimate images, as well as remove images that have already been shared. However, due to the volume of

complaints we receive, we prioritise complaints that require the quickest compliance and enforcement action. As part of this process, we take into account a number of factors, including:

- the urgency of the situation;
- the extent and nature of the abuse;
- whether the image is currently online and the accessibility of the image;
- any vulnerability or risk factors experienced by the person being targeted; and
- whether the intimate image was created for a commercial purpose.

If the intimate image has been shared online, we may be able to act to have it removed. We also refer reports of sextortion or image-based abuse targeting children to the AFP-led Australian Centre to Counter Child Exploitation (ACCCE), as it is the national coordination mechanism for online child sexual exploitation and abuse.

Where an unknown end-user is threatening to share an intimate image of an adult unless their demands – often financial – are met, eSafety provides information about how to manage the situation, as well as online safety tips and avenues for support. This approach assists us to prioritise actionable reports.

We also share intelligence with law enforcement and other government departments and use information from reports to inform potential regulatory action against high-risk platforms. Additionally, through our partnership with the Joint Policing Cybercrime Coordination Centre (JPC3), we engage in discussions with the Joint Cybercrime Prevention Network around strategies to reduce the prevalence of sexual extortion.

In addition to regulatory and enforcement action, we continue to actively promote our extensive online safety resources. This includes both preventative and general advice for people experiencing sextortion as well as tailored advice. We also provide advice and support to education sectors through professional learning and to the National Online Safety Education Council (NOSEC).

However, more needs to be done to protect Australians against this growing and significant harm. Industry must take greater steps to build-in safety features and systems from the outset, especially given the potential for new technology such as generative AI to be weaponised against children and young people.

Review of the Online Safety Act 2021

The Australian Government announced an independent review of the *Online Safety Act 2021* to be conducted during 2024. As the terms of reference indicate, the review is broad ranging. It will include consideration of eSafety's existing statutory schemes, including those outlined above. It will also consider whether additional arrangements are warranted to address online harms not explicitly captured under the existing statutory schemes, such as online hate, or potential online safety harms raised by technologies like generative AI and recommender systems.

Public consultation commenced with the release of an Issues Paper on 29 April 2024, and submissions close on 21 June 2024. eSafety will provide a public submission touching on our goals for the review. We will also continue to work closely with the Australian Government to ensure the Online Safety Act and related enabling legislation remains fit for purpose and adequately reflects Australians' needs and expectations.

The Final Report of the Review is expected to be provided to the Minister for Communications by 31 October 2024.

Conclusion

As Australia's leader in online safety and foremost among international online safety experts, eSafety does not underestimate the challenge – or importance – of online safety.

We reiterate our fundamental objective of keeping Australians safe online.

We also reiterate the need for online safety measures to be evidence based, reasonable, proportionate, balanced and practical. They should also focus on accountability and transparency, while seeking to promote the rights, empowerment and capacity of all Australians, including children.