eSafety response to Ofcom Call for Evidence

Preliminary questions

Q1. Please provide a description introducing your organisation, service or interest in Online Safety.

The eSafety Commissioner (eSafety) is Australia's independent regulator and educator for online safety. eSafety represents the Australian Government's commitment to protecting citizens from serious online harms.

Online harms are activities that take place wholly or partially online that can damage an individual's social, emotional, psychological, financial or even physical safety. These harms occur as a result of content, conduct, or contact and can include online activity or material that:

- depicts sexual exploitation or sexual abuse of children
- promotes, instructs, or incites terrorism, violent extremism or other criminal activity, such as rape or murder
- encourages or promotes suicide or self-harm
- bullies, abuses, threatens, harasses, intimidates, or humiliates another person
- involves non-consensual sharing of intimate images or videos
- is inappropriate and potentially damaging for children to see.

The Online Safety Act 2021 (the Act) governs the functions of eSafety, which includes administering complaints-based schemes, and enabling industry to develop mandatory codes to regulate the most seriously harmful online content, such as child sexual abuse material and pro-terror content.

The Act also provides for Basic Online Safety Expectations (the Expectations), which outlines the Australian Government's expectations that social media, messaging and gaming service providers and other apps and websites will take reasonable steps to keep Australians safe.

eSafety can now require online service providers 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 mechanisms. eSafety can also publish statements about the extent to which services are meeting the Expectations.

The first mandatory reporting notices focussed on understanding the steps being taken by services to prevent, detect and remove child sexual exploitation and abuse material. Services have 28 days to respond, or longer as agreed with eSafety. eSafety may publish information received from industry, where this meets the objectives of the Act.

Risk assessment and management

Q2. Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services?

eSafety's complaints schemes provide insight into the harms and illegal content Australians are exposed to online. In recent years, eSafety investigative staff, Australian and international law enforcement, and child safety organisations have all identified significant increases in the scale and volume of online child sexual exploitation material (CSEM). Experts are particularly concerned about rises in cases of offenders using social media and messaging services to groom and extort children into self-producing child sexual exploitation material, and the livestreaming of child sexual exploitation and abuse (CSEA) as an increasingly common crime type.

eSafety has handled more than 61,000 complaints about illegal and restricted content since 2015, with the majority involving CSEA, with numbers surging since the start of the COVID-19 pandemic. This increase has been sustained in the last 12 months, with sexual extortion (of both minors and young adults) our biggest issue at present. Most recent data across all four regulatory schemes illustrate the surging demand:

- Image Based Abuse – 186% increase in complaints in August 2022 compared to August 2021
- Illegal and Restricted Online Content – 31% increase in complaints (August 2021-22)
- Cyber Bullying - 78% increase in complaints (August 2021-22)
- Adult Cyber Abuse – 1687 complaints since the scheme commenced under the OSA on 23 January 2022.

The WeProtect Global Alliance 'Global Threat Assessment' reports that in 2020, more than 1 million individual media files were exchanged via INHOPE's child sexual abuse material collection and classification platform, and the UK Internet Watch Foundation reported 77% increase in child self-generated sexual material from 2019 to 2020. The US National Centre for Missing and Exploited Children reported a 100% increase in reports from the public of online sexual exploitation from 2019 to 2020, and a 35% increase in 2021.

Despite these systemic issues and harms being well-known, the tools and resources that companies are using to prevent and detect CSEA are opaque. Key areas like the lack of user-reporting features on some major platforms, and insufficient action on repeat offenders (recidivism), are also issues relevant to online harms.

In recent years, many of the largest companies started to develop transparency centres, and report on the types and volumes of online harms on their services. eSafety welcomes these efforts. They have also formed industry groups on certain issues, such as the Technology Coalition, and Tech against Terrorism and the Global Internet Forum to Counter Terrorism (GIFCT), and have engaged in processes including the OECD Voluntary Transparency Reporting Framework. However, eSafety is disappointed by the transparency shown by industry through multi-stakeholder and industry groups to date. Furthermore, engagement with multistakeholder bodies tends to result in obfuscation and removal of active industry participation into industry-only constructs. Industry-only groups tend to provide broad generalisations about industry data rather than valuable and meaningful data.

eSafety 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. See response to Q27 for further information.

Terms of service and policy statements

Q5. What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?

eSafety's Safety by Design initiative, outlined below, provides guidance to online service providers on a range of safety considerations, including the following recommendations for online services regarding terms of service and community standards:

- provide easily discoverable and transparent community standards, terms of service and related protocols, which are in plain language and use short-form notices as a standard.
- implement social contracts at the point of registration. These outline the duties and responsibilities of the service, user and third parties for the safety of all users (SbD Principle 1.7).
- ensure that user safety policies, terms and conditions, community standards and processes about user safety are visible, easy to find, regularly updated and easy to understand. Users should be periodically reminded of these policies and given the option of being proactively notified of changes or updates through targeted in service communications (SbD Principle 3.2).
- carry out open engagement with a wide user-base, including independent experts and key stakeholders, on the development, interpretation and application of safety standards and their effectiveness or appropriateness (SbD Principle 3.3).

Safety by Design also highlights the importance of implementing social contracts on platforms and services – in which members of a community agree to cooperate – to engender support among users for a service's community guidelines. Social contracts:

- Encourage users to understand their own role in upholding and maintaining community safety norms, as well as the service's obligations.
- Create a sense of shared responsibility between the service and its user base to foster positive experiences for everyone.

Further guidance can be found in the Safety by Design assessment tools and the resources provided in the assessment tool end-reports.

Overview of Safety by Design

Safety by Design is an initiative focuses on the ways technology companies can minimise online threats by anticipating, detecting and eliminating online harms before they occur. This proactive and preventative approach focuses on embedding safety into the culture and leadership of an organisation. It emphasises accountability and aims to foster more positive, civil and rewarding online experiences for everyone.

Research and consultation on Safety by Design began in 2018. The Safety by Design Principles have been developed from information collected through eSafety's research and reporting schemes, outreach programs, industry and key stakeholder consultations, a youth consultation exercise and a parent and guardian survey. They are anchored in earlier work focusing on the safety of users online, along with well-established theoretical models and human rights instruments.

Further information on eSafety's Safety by Design initiative can be found at www.esafety.gov.au/industry/safety-by-design

Reporting and complaints

Q7. What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms?

eSafety consulted over 180 organisations over 18 months to develop the Safety by Design assessment tools. Throughout the consultation process, eSafety heard that there are a range of measures that online services can implement to ensure users can understand and access reporting and complaints mechanisms. These include:

- consultation with a broad user base to ensure user needs, including those most at-risk of harms, are understood.
- user empowerment, through the use of nudges, prompts and educative updates
- establishing appropriate reporting mechanisms with regular updates to users, links to third party support services and feedback loops, and providing easily accessible reporting mechanisms in platform, in real time.
- publishing an annual assessment of reported abuses on the service, accompanied by the open publication of meaningful analysis of metrics such as abuse data and reports, the effectiveness of moderation efforts and the extent to which community standards and terms of service are being satisfied through enforcement metrics (SbD Principle 3.4).
- ensuring that information on terms of use, policies and complaints mechanisms is:
    - readily accessible to end-users
    - accessible at all points in the end-user experience (for online safety settings, parental controls, and eSafety resources)
    - regularly reviewed and updated
    - written in plain language. (BOSE S17 Regulatory Guidance).

The Canadian Centre for Child Protection has previously analysed[1] different services reporting options for child sexual exploitation and abuse and highlighted the importance of dedicated reporting categories in app.

Moderation

Q11. Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?

Measures that could be employed to improve content moderation include:

- providing technical measures and tools that allow users to manage their own safety, and that are set to the most secure privacy and safety levels by default (SbD Principle 2.1)
- leveraging the use of technical features to mitigate against risks and harms, which can be flagged to users at point of relevance, and which prompt and optimise safer interactions (SbD Principle 2.3).
- using human moderators, alongside algorithms, to create a safer but not restrictive environment.
- documenting moderation procedures that are fairly and consistently implemented.

- publishing an annual assessment of reported abuses on the service, accompanied by the open publication of meaningful analysis of metrics such as abuse data and reports, the effectiveness of moderation efforts and the extent to which community standards and terms of service are being satisfied through enforcement metrics (SbD Principle 3.4).
- providing built-in support functions and feedback loops for users that inform users on the status of their reports, what outcomes have been taken and offer an opportunity for appeal (SbD Principle 2.4).

While eSafety has graduated measures to act against online harm and operates on principles of harm minimisation, there are times when removal, blocking and proactive scanning of content is warranted due to the circumstances and severity of harm.

For example, proactive scanning and removal of child sexual exploitation material. The industry codes under the *Online Safety Act 2021* may require some services to take measures to detect and prevent, or limit such material. Examples of such measures could include:

- ongoing investment in, and development and use of, tools to detect, moderate and report material (for example, through the use of hashing, machine learning, artificial intelligence or other safety technologies).
- development and use of effective moderation practices and procedures (for example, automatic pre-moderation, proactive machine monitoring, human monitoring, hybrid moderation, appointed community moderators and community moderation) to take action against harmful content and activity, including through warning account-holders, suspending or removing accounts, removing content and deindexing of search results.
- default settings for services marketed to children which are set to the highest possible privacy and safety level at registration or sign up (for example, access to device hardware such as cameras and microphones is limited and photos, location, friends lists, profile information and chat functions are only accessible to approved contacts. This might also include safety settings such as safe search mode on by default and measures which would prevent comingling).
- standard operating procedures which include clearly specified channels for escalating and/or reporting unlawful and harmful material, including to law enforcement, child protection or relevant authorities.

Design and operation of the service, including functionalities and algorithms

Q18. Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?

Reducing the likelihood of online harms through prevention is critical. eSafety believes it is always better to solve the chronic problem rather than continually grapple with the acute symptoms. Adopting Safety by Design can help minimise the threat surface before harms occur, through measures such as:

- evaluating all design and function features to ensure that risk factors for all users—particularly for those with distinct characteristics and capabilities—have been mitigated before products or features are released to the public (SbD Principle 2.5).

- employing technical solutions to support proactive detection and removal of illegal and restricted content, such as CSAM and pro-terror content.
- ensuring that safety settings such as safe search mode are turned on by default.
- requesting internal policies and procedures that include safety risk and impact assessments, which are subject to ongoing consultation, review and evaluation.
- providing clear, easily accessible and effective complaints mechanisms and reporting tools.

Child protection

Q22. What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?

In June 2021, the Australian Government requested that eSafety develop an implementation roadmap for a mandatory age verification (AV) regime relating to online pornography, which aims to mitigate harms associated with children and young people's access to online pornography. In developing the Age Verification Roadmap, eSafety is examining a range of age assurance measures.

While eSafety is careful not to pre-empt the findings and recommendations of the Roadmap, its analysis and consultation to date indicate that factors such as privacy, security, data minimisation, equity and inclusion, choice, trust, accuracy and the impact on competition will be key in assessing the efficacy, proportionality and feasibility of any age assurance regulation.

It is important to note that the Age Verification Roadmap is considering age assurance and age verification technology as one intervention among a holistic regime of tech and complementary non-tech measures, including educational measures.

The Roadmap will propose a series of suggestions and considerations for the Australian Government, including:

- options for designing a regulatory and legislative framework or leveraging existing frameworks to apply to age assurance.
- what types of safety technologies could be appropriate and what standards would be needed to support them.
- what complementary and public awareness measures would be required to support regulatory measures.

Through consultation on the Age Verification Roadmap, eSafety found that stakeholders:

- held diverse views on the cost and impact of age verification technology, particularly whether the responsibility was placed on large technology providers or on small content producers.
- encouraged policymakers to look up and down the stack, including websites, device manufacturers, operating system developers and internet service providers (ISPs), and consider different levels of intervention at each layer.
- indicated support for device level filtering and age assurance measures. Stakeholders suggested that measures at the device level can reduce privacy concerns and the number of actors required to implement compliance measures.

- reflected significant concerns about the privacy implications and data bias risks of both age assurance and verification measures.
- noted their concern about the impact of inconsistent or incompatible international regulatory regimes and approaches.

The implementation of age verification technology needs to consider the broader domestic and international regulatory context, including the development of international standards. This ensures consistency and interoperability across jurisdictions, but also to ensure the regulatory and technological environment can support any age assurance measures or recommendations.

eSafety has engaged with Ofcom on the AV Roadmap to date and looks forward to continuing our engagement as this work develops.

For more information on eSafety's Age Verification Roadmap, including timeframes and consultation summaries, please see our website.

Q23. Can you identify factors which might indicate that a service is likely to attract child users?

A range of features may increase the likelihood of children using a service, including social features, messaging, livestreaming, gaming and education. It is important to note that there is nothing intrinsically wrong with services attracting users – children and young people should be able to access the benefits of online participation. However, where services are designed for, and marketed to children, or likely to be used by children, companies have a particular responsibility to ensure they are safe by design.

Where services are specifically designed for adults, it will be important to ensure that appropriate measures are used to prevent young people from accessing adult content. This may include age assurance or verification, privacy by default measures, and use of proactive tools to detect likely harm, such as grooming, clear reporting and blocking options, and protection from content that may be harmful to children like pornography.

Transparency

Q27. For purposes of transparency, what type of information is useful/not useful? Why?

As outlined in Q2, eSafety has welcomed the increased global attention on the types and prevalence of online harms on platforms and services. However, eSafety is disappointed by the transparency shown by industry through multi-stakeholder and industry groups to date, as previously outlined. More than two years ago, some of the largest companies committed to improve transparency through an industry group on child sexual exploitation and abuse, the Technology Coalition. What we have seen since then has amounted to anonymised and aggregated information and high-level commitments to improve in the future.

On terrorist content, despite the involvement of companies and a wide range of countries, only two companies have published reports under the OECD's Voluntary Transparency Reporting Framework.

We also note that 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 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. Australia's Basic Online Safety Expectations are designed to improve providers' safety standards, and improve transparency and accountability. eSafety intends for notices under the Basic Online Safety Expectations (the Expectations) to ask a combination of:

- Qualitative information on safety tools, processes and policies, and why these are reasonable steps to implement the Expectations. These may be phrased as yes/no questions, multiple choice questions or worded to seek descriptive information.
- Quantitative information on the operation of safety tools, processes and policies. This may consist of metrics to determine the impact of interventions or information about resources allocated.

eSafety has published a summary of the high level areas, and corresponding expectations, that the first notices on the Expectations focussed on, here. The objectives of the Expectations are to improve transparency and accountability, and eSafety has therefore said that we will publish information received from industry, where appropriate and that meets the objectives of the Online Safety Act 2021. In doing so, eSafety considers that there is significant potential to incentivise companies to improve their safety processes and protect users.

eSafety will not publish information that aids bad actors in misusing services, and it will consider any claims from industry that information is commercial in confidence.

We also 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, and look forward to working closely alongside Ofcom to promote harmonisation of approaches to promote coordination and positive outcomes for users globally.