Australian Government | eSafetyCommissioner

Safety by Design – Implementation and Impact

Proposal for Open Forum session at the Fifteenth Annual Meeting of the Internet Governance Forum 2020: Internet United

22 April 2020

P: +61 800 880 176 E: SafetybyDesign@esafety.gov.au

Overview of eSafety

The Australian eSafety Commissioner was established in July 2015 under the *Enhancing Online Safety Act 2015* (Cth) (the Act). At the time, the Act gave eSafety a remit of enhancing online safety for children which was expanded in July 2017 to include promoting online safety for *all* Australians. The establishment of eSafety reflects the Australian Government's significant commitment to protecting Australians online, so they can safely access the opportunities that the digital environment provides.

Under the Act, eSafety has a powerful combination of functions, which range from prevention through education, research and awareness raising, to early intervention and harm minimisation. eSafety's strong suite of regulatory powers facilitate the rapid removal of harmful content online, which operate under three regulatory schemes which cover Cyberbullying, Image-Based Abuse and Offensive and Illegal Online Content. The Commissioner has also been empowered to issue notices to content and hosting services being used for abhorrent violent material. Further the Commissioner can formally direct Australian Internet Service Providers to block harmful content that promotes, incites or instructs in terrorists acts or violent crimes.

The eSafety Commissioner's prevention work focuses on education, research and awareness raising. This includes leading and coordinating online safety efforts across Commonwealth agencies to help safeguard Australians at risk from online harms and to promote safer, more positive online experiences.

eSafety supports, conducts and evaluates educational programs and research about online safety. This research and evaluation work programme underpins all of eSafety's policies and programs, ensuring that all of our work is based on robust evidence of relevant risk and protective factors.

eSafety also plays a key role in educating Australians to develop critical digital skills. We achieve this through outreach programs in schools and the community; online information for children, young people and parents; virtual classrooms; peer-led 'digital leaders' programs; lesson plans; face-to-face training; expert guidance; and supporting NGOs and experts to deliver their own online safety programs and presentations to schools. Details of all our programs can be found on our website.

Reporting and Complaints Mechanisms

A significant component of eSafety's mandate is to administer three reporting and complaints schemes:

- The **Cyberbullying Scheme** came into force in 2015. It allows Australian children under 18 years who are experiencing cyberbullying—or their parents, carers or authorised persons—to make a complaint to eSafety if social media sites fail to remove cyberbullying material from their sites within 48 hours of being alerted to it. The scheme serves as a safety net for children who have been cyberbullied on a relevant service and have not been able to resolve the issue via social media services' reporting functions.

  Since its commencement, eSafety has assisted more than 1,100 children and families with cyberbullying issues and has had success in working collaboratively with social media services to have material removed very quickly. Schools have reported to us that when eSafety acts to address a complaint from a student, bullying dissipates in the school community.

- The **Image-Based Abuse Scheme** came into force in 2018. It allows Australians of all ages to report to eSafety if their intimate image has been posted, or a threat has been made to post, on a relevant service without their consent. The

intent of the legislation is to send a clear message to the community that the non-consensual sharing, or threatened sharing, of intimate images is unacceptable.

There are two aspects to the scheme: first, a legal framework for the removal of image-based abuse content through enforceable removal notices, and second, options for taking action against the person responsible for the image-based abuse. We take a graduated approach to enforcement action and have a wide range of informal and formal options available. We may give removal notices to hosts or individuals responsible for sharing (or threatening to share) the material, for which failure to comply may attract civil penalties of up to $105,000 AUD for individuals and up to $525,000 AUD for corporations. These penalties serve as a deterrent whilst also providing a significant incentive to individuals who have engaged in image-based abuse and relevant services to remove material quickly, minimising harm to victims.

To date, we have received 784 reports concerning over 1,700 URLs. Approximately 30% of our image-based abuse reports involve a person who was under 18 years of age in the intimate image. eSafety takes a holistic approach to ensure victims are well-supported and to minimise risk of further harm. Our reporting portal helps victims regain control by providing reporting options, referrals to support services and helpful resources. We have strong collaborative arrangements in place with social media services to facilitate rapid removal of intimate images, providing quick relief to victims. eSafety has taken action against persons responsible for image-based abuse on five occasions and has successfully removed images reported to us in over 80% of cases. When the individual responsible for image-based abuse is a child, we generally take a remedial and educative approach, combined with removal action. If material is indicative of child sexual abuse, it is addressed through our Online Content Scheme.

- The **Online Content Scheme** came into force in 1999 and, before the creation of eSafety, was administered by the Australian Communications and Media Authority. It allows Australians to report (including anonymously) certain illegal and offensive online content, including child sexual abuse material. eSafety deals with thousands of reports each year, with almost 75% of these concerning child sexual abuse material. eSafety is able to notify material to the Australian Federal Police for criminal investigation and onward referral to Interpol, and to facilitate the rapid removal of the vast majority of this material through our international networks.

  Our experience in administering the Online Content Scheme, in particular, points to the value of international cooperation in regulating the online world. The vast majority of the offensive and illegal online content reported to eSafety is hosted overseas. In these instances, we do not have jurisdiction to issue a takedown notice. However, our membership of the International Association of Internet Hotlines (INHOPE) enables us to refer content directly to a relevant INHOPE member and have content removed expeditiously so as to prevent the spread of child sexual abuse material and the re-victimisation of the children who are the subject of these images. These types of partnerships and collaborative approaches are key to effective online regulation.

These three schemes offer Australians practical help in managing the impact of these types of abuses, but their real uniqueness lies in the fact that eSafety can formally direct certain online service providers to remove content from their services, providing and empowering victims of online abuse to take control and help reduce harm and re-victimisation. While the schemes are largely operating as a cooperative model between government and industry, the powers available to eSafety to compel the removal of material provide a critical safety net and drive industry to be proactive in addressing online harms.

Safety by Design

eSafety recognises the need to drive-up standards of user safety within the technology community, and to encourage and secure greater consistency and standardisation of user safety considerations. To reduce risks and counter threats online, a proactive approach is critical. We recognise the importance of proactively and consciously considering user safety as a standard risk mitigation and development process.

In June 2018, eSafety laid out our intention to develop a Safety by Design Framework and set of Safety by Design Principles. At its core, Safety by Design (SbD) is about embedding the rights of users and user safety into the design, development and deployment of online and digital products and services. Consideration and care was taken to ensure that the SbD Principles and Framework balance an individual's right to provision, participation and protection. In addition, the clear expectations on businesses to meet human rights responsibilities in the online world have been reflected in the SbD Principles.

SbD places user safety considerations at the centre of product development. It recognises and responds to the intersectionality of risk and harm in the online world and acknowledges the potential of advancements in technology, machine-learning and artificial intelligence to radically transform user safety and our online experiences. While personal information privacy and cyber security fall outside of eSafety's remit, SbD looks to the well-established processes surrounding privacy and security, and endeavours to elevate safety as the third design pillar in the developmental process.

Following an eight-month consultation process with industry, parents and carers and children, eSafety developed a set of SbD Principles that provide online and digital interactive services with a universal and consistent set of realistic, actionable and achievable measures to better protect and safeguard citizens' safety online. Three overarching principles were developed:

1) **Service Provider responsibilities**: This is premised on the fact that the burden of safety should never fall solely upon the end user. Preventative steps can be taken to help ensure that known and anticipated harms have been fully evaluated in the design and provisions of an online service, and steps taken to make services less likely to facilitate, inflame or encourage illegal and inappropriate behaviours.
2) **User empowerment and autonomy**: The dignity of users is of central importance, with users' best interests a primary consideration. Human agency and autonomy can be supported, amplified and strengthened when designing services – allowing users greater control, governance and regulation of their own experiences, particularly at times when their safety is being, or is at risk of being, compromised.
3) **Transparency and accountability**: These are hallmarks of a robust approach to safety, that provide assurances that services are operating according to their published safety objectives, as well as educating and empowering the public about steps that can be taken to address safety concerns.

Our work on SbD is ongoing. With the core SbD Principles established, we are currently in the process of developing a Framework of resources and guidance to assist industry in implementing the SbD Principles.

As the concept of SbD gains momentum internationally, eSafety's has been sharing its work with international partners to ensure that a consistent global approach to 'safety by design' is undertaken. At eSafety we know that meaningful change can require a pivot in approaches, and that change can take time. We are committed to working with all our partners to reduce, and counter, the risks and threats that citizens face online. We believe that Safety by Design is a catalyst for this change to occur.

IGF Forum

Australia's eSafety Commissioner has developed a world-first Safety by Design (SbD) initiative – which seeks to help encourage online platforms to build in user safety from design to deployment. The SbD Principles were established following an extensive consultation process with industry, trade bodies and organisations with responsibility for safeguarding users, as well as parents, carers and young people. They have received broad support from a wide range of industry members who vary in size, composition, location and market share – and there is increasing interest in the initiative from Governments across the world, as well as multi-lateral and international organisations.

The IGF Forum would provide eSafety with an opportunity to raise awareness of its work to a wider audience - encouraging and securing greater consistency and standardisation of user safety considerations across the globe. Long-term, sustained social and cultural change to promote and protect the rights of citizens online requires the coordinated efforts of the global community. The significance of national and international collaboration, multi-stakeholder engagement, and investment in Safety by Design has never been more important.

Open Forum - Submission

eSafety is making a submission for an Open Forum slot at the 15th Annual Meeting of the Internet Governance Forum: Internet United.

Its submission falls under the 'Trust' thematic track - as eSafety believes that Safety by Design is a fundamental prerequisite for enabling healthy and empowering digital environments. eSafety is seeking to raise awareness of its work and to engage with the IGF community to achieve greater impact and change for user safety.

The technological design and architecture of online services governs how users are able to interact and engage online. These aspects act as both a facilitator and amplifier for how humans interact, engage and behave. While technology may not drive behaviours, it is a medium through which these behaviours can manifest. As such, developers, engineers and vendors of online services play an incredibly important role in shaping online environments and users' safety therein.

The session will discuss the development of eSafety's Safety by Design principles - and its consultation process, particularly its youth consultation and the development of a youth vision statement (provided in full at appendix A). It will touch on how SbD is not common practice among industry partners, but that innovations in user safety are beginning to be developed at pace by some industry players. It will explore the impact that embedding user empowerment and autonomy as a core business objective for those developing products, platforms or services online could have. Finally, it will discuss the development of eSafety's SbD Framework of guidance and resources - seeking audience feedback and suggestions on how to ensure greater implementation and impact.

Session timings

- Overview of the Australian eSafety Commissioner
- Overview of Safety by Design
- Panel session to discuss implementation and impact of SbD
    - Panel to include representation from industry, civil society organisation and young people: facilitated by eSafety
        - Youth participation may be in the form of online videos - rather than in-person attendance
- Audience questions and participation (including examples of best practice)

Appendix A: Safety by Design Youth Vision Statement

Our Vision: Young People

Our vision is that the Australian online industry:

- Enables users to **control** their online experiences and safety through the provision of tools and features that provide them with choices.
- Develops a strong set of easy-to-understand, highly visible, **ground rules** that have user safety at their core.
- Ensures users can easily **block and report** both people and content, placing control in the hands of the individual. This allows users to manage their online experiences and to help shape a more positive environment.
- Implements impactful **consequences and sanctions** for those who violate rules. This will reassure other users that their safety and security is a priority and sets clear expectations about how users should behave.
- Uses **developments in technology** to identify and minimise exposure to threats, risks, problems or content that is triggering, harmful or inappropriate. These precautions will help prevent harm or abuse, while also ensuring help is provided to those at risk.
- Provides users with **information and awareness** about safety features because knowledge leads to greater understanding, confidence, trust and, ultimately, use.
- Uses **human moderators**, alongside algorithms, to create a safe but not restrictive environment. Abuse and hatred should not be tolerated, and moderation would help prevent these from spreading.
- Provides users with **support**, and support networks, when they need it—especially when they are feeling low or do not feel safe. This will make users feel that they are not alone, that there are people and systems to help.
- Ensures **privacy settings** are comprehensive and set at the highest levels of protection by default. Also ensures that users know how to maintain and control their privacy, safety and security.
- Enforces some means of **verification** to make sure that people are real, are who they say they are and are accountable for their actions.
- Is aware of, and responsible for, the safety of users by valuing them above all else, **understanding the issues** they face and protecting their privacy and safety.
- **Empowers** users to interact freely online and to enjoy the benefits that the online world offers—without fear and without their rights or safety being put at risk.