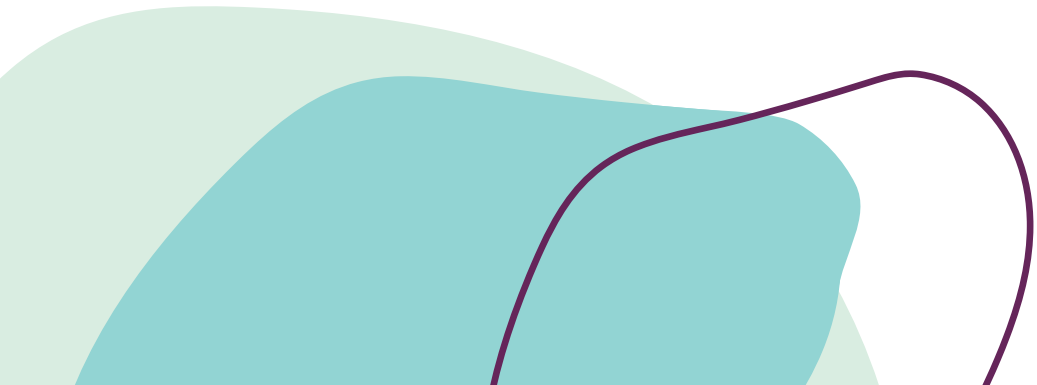
May 2019

Safety by Design Overview

Australian Government | eSafetyCommissioner

Contents

I. Introduction	**1**
II. Background	**3**
III. The foundations of the Safety by Design Principles	**6**
IV. Consultation process	**10**
V. Safety by Design Principles	**19**
Principle 1: Service provider responsibilities	**20**
Principle 2: User empowerment and autonomy	**28**
Principle 3: Transparency and accountability	**32**
Summary	**35**
Appendix A	**36**
Appendix B	**37**



I. Introduction

Safety by Design

Protecting and safeguarding citizens online is a global concern. To reduce risks and counter threats, we need a proactive approach that addresses the complex societal situations and behaviours that manifest in the online world. Online risks and harms are often inter-related and inter-connected and occur across the entire spectrum of online devices, services and platforms.

As the Australian government agency responsible for promoting online safety, as outlined in the *Enhancing Online Safety Act 2015,* the eSafety Commissioner (eSafety) strives to empower and protect all Australians. We do this by providing resources and links to support that focus on building respect, resilience, responsibility and reasoning in the online world. This is achieved through functions including:

- regulatory powers to deal with specific cases of cyberbullying, illegal online content and image-based abuse
- the production of research, best practice guidance, training, outreach programs, educational resources and online content to help raise awareness and prevent harm online.

A combination of prevention through education, awareness, early intervention and harm remediation sits at the heart of the approach taken by eSafety. This holistic, multi-faceted and multi-pronged method enables us to work toward ensuring that all Australians benefit from the richness and opportunities offered by the online world, while enabling us to reduce the negative impact of online risks and harms.

To date, much of our work has focused on promoting online safety to the community. However, we now feel that in order to improve protections for Australians online, a greater degree of attention needs to be placed on service providers themselves. Our goal is to drive-up standards of user safety within the technology community, and to encourage and secure greater consistency and standardisation of user safety considerations.
The technological design and architecture of online services govern how users[1] are able to interact and engage online. The online world is both a facilitator and amplifier of human interactions and while technology may not drive behaviour, 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 user safety.

Advancements in technology, machine-learning and artificial intelligence have the potential to radically transform user experiences and safety online. At eSafety, we recognised the importance of proactively and consciously considering user safety as a standard risk mitigation and development process, rather than retrofitting safety considerations after online harms emerge or the damage has occurred. In June 2018, we stated an intention to develop a Safety by Design Framework ('the SbD Framework') and Safety by Design Principles ('the SbD Principles').

The Safety by Design initiative was developed to provide online and digital interactive services with a universal and consistent set of realistic, actionable and achievable measures to better protect and safeguard citizens online. The SbD Framework is the broad program of resources and support which will help to guide organisations as they embed the rights of users and user safety into the design and functionality of products and services. The framework, and all resources, have SbD Principles at their core.

[1] The term 'user' is used throughout the paper to describe the intended audience or 'consumer' of the service or product

The SbD Principles are intended to act as a model template. They provide a benchmark for industries of all sizes and stages of maturity, and aim to provide guidance in incorporating, enhancing and assessing user safety considerations throughout the design, development and deployment phases of a typical service lifecycle. The principles firmly place user safety as a fundamental design principle that needs to be embedded in the development of technological innovations from the start.

The SbD 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. They place user safety as the third pillar in the developmental process for all online and digital technology, sitting alongside privacy and security.

The result is a framework that provides online and digital services with practical, realistic and achievable guidance for embedding user safety into the design of their product or service. The principles are intended to apply to any technological tool, product, device or service that enables interaction within the general population.

The second phase of the SbD initiative will focus on developing guidance and resources to assist industry partners in actioning the SbD Framework. Like other aspects of the development of the Principles, we will carry out this exercise collaboratively. We will also share the SbD Framework with our international partners, who have expressed a keen interest in its development. Our goal is to develop a shared and consistent global pathway so that greater strides are taken to secure a safer online environment for all.

II. Background

Evidence, research and existing frameworks/models

Theoretical models of online harm and risks

A number of models and research projects were used to categorise the types of risks and harms addressed by the SbD Principles.

Much of the research that focuses on 'online risks' has centred on children and young people, with a number of classification models and theories emerging[2, 3, 4, 5, 6]. One of the most widely accepted is that developed by the EU Kids Online project, which classifies risks across three categories:

1. Content risks — which generally position the child as the recipient of unwelcome and inappropriate content.
2. Contact risks — where a child participates in risky communication, possibly unwittingly or unwillingly.
3. Conduct risks — where a child's behaviour contributes to risky content or contact within a wider peer-to-peer or adult-to-child network.

Researchers have also identified five potential 'harms' relating to online content and conduct as it relates to media consumption[7].

These are:

1. Illegal content — such as hate speech, child exploitation or incitement to terrorism.
2. Age-inappropriate content — such as adult sexual material, disturbing or violent content.
3. Other potentially dangerous content — which poses a significant risk of personal harm, such as videos or images promoting self-harm or violence.
4. Misleading content — including mis-information and mis-leading information.
5. Personal conduct that is illegal or harmful — such as bullying, grooming and harassment.

Research continues to seek to understand why risk translates into harm for certain individuals[8], so that the underlying factors that can place individuals at greater risk or make them more susceptible to risk in the online age, are uncovered[9,10].

Evidence of online risks and harms

Research from eSafety's reporting schemes and outreach programs consistently highlights the magnitude of online harms faced by Australians — especially young Australians.

[2] Livingstone, S., Mascheroni, G. & Staksrud, E. 2015. Developing a framework for researching children's online risks and opportunities in Europe, EU Kids Online.

[3] Youth Protection Roundtable Toolkit. Found at www.kijkwijzer.nl/upload/download_pc/74_final_YPRT_Toolkit.pdf

[4] O'Neill, B. 2018. Research for CULT Committee – Child safety online: definition of the problem, European Parliament, Policy Department for Structural and Cohesion Policies, Brussels. Found at www.bit.ly/2nOHKeN

[5] Teimouri, M, Benrazavi, SR, Griffiths, MD & Hassan, Md Salleh Hassan, 2018. 'A model of Online Protection to Reduce Children's Risk Exposure: Empirical Evidence from Asia'. Sexuality & Culture, Vol 22, Issue 4. pp 1205-1229.

[6] UNICEF, 2017, Children in a Digital World: The State of the World's Children 2017, Report.

[7] Ofcom, 2018, Addressing harmful online content: A perspective from broadcasting and on-demand standards regulation.

[8] The Berkman Center for Internet & Society at Harvard University 2008, Enhancing Child Safety & Online Technologies: Final report of the Internet Safety Technical Task Force to the multi-state working group on social networking of State Attorneys General of the United States.

[9] UNICEF, 2017, Children in a Digital World: The State of the World's Children 2017, Report.

[10] Please refer to Global Kids Online and the DQ Institute research and impact reports for an overview of research on risks and harms faced by children and young people globally.

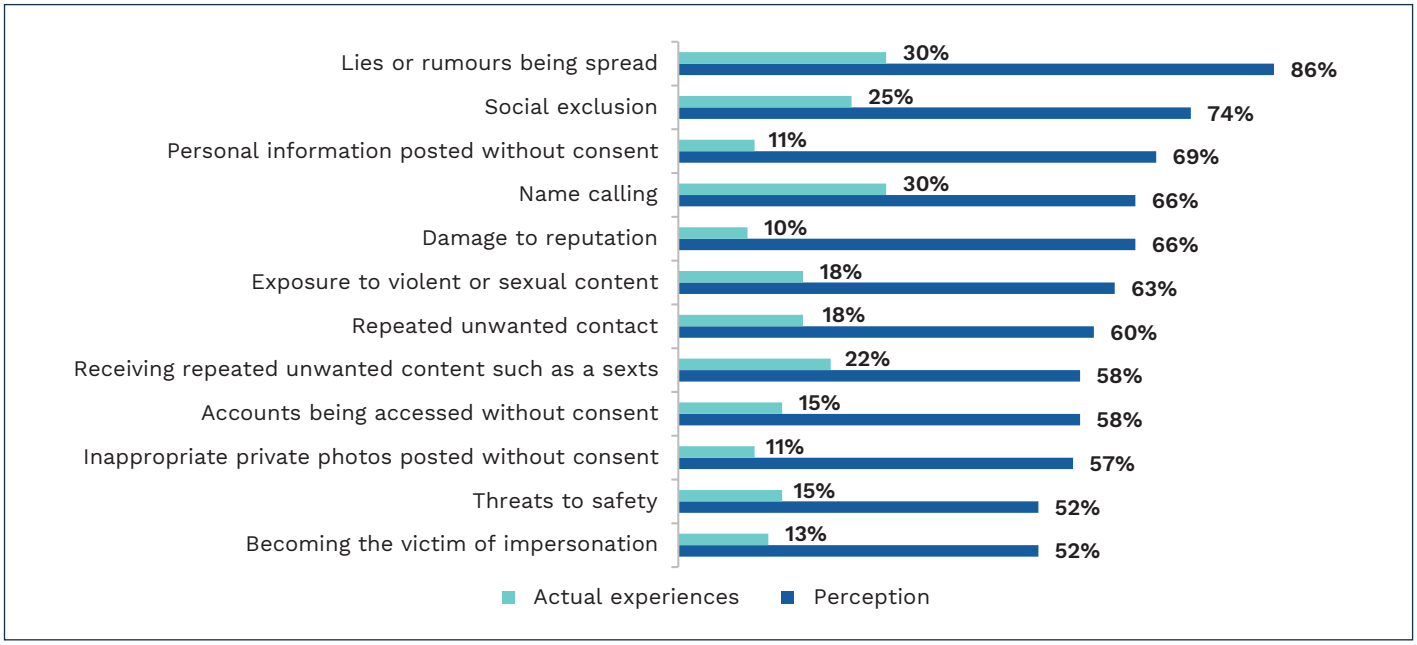


Figure 1: Comparison of perceived online risks and findings from eSafety's 2017 Youth Digital Participation National Survey.

Risks

Two studies conducted by eSafety[11, 12], asked young people to identify the risks that they face online[13] (see Figure 1). Spreading lies or rumours was identified as the top online risk across both studies, followed by social exclusion and name calling. However, it is important to note that the perception of risks was higher than the prevalence rates of actual negative experiences reported between the studies.

Harms

Cyberbullying

eSafety manages a world-first complaints system for the removal of serious cyberbullying material concerning Australian children. Children, parents or another responsible person with the consent of the child, can lodge a complaint and receive timely advice and assistance. In almost three years of operation, the scheme has received over 1,000 complaints about cyberbullying affecting Australian children. The most common complaints include nasty comments and serious name calling (including those that incite suicide and self-harm), impersonation or hacking of social media accounts, unwanted contact, sexting and image-based abuse. Our experience shows that children and young people are predominantly bullied online by those in their own peer group. In many instances, cyberbullying is an extension of bullying or conflict occurring within the school. In reports to eSafety, victims often note that the harassment they experience online broadly mirrors their experience at school. Further, the perpetrators are, in many instances, one and the same.

Similar patterns of behaviour have also been highlighted in reports made to eSafety by adults14, with most complaints relating to complex and longstanding cyberbullying behaviours on social media platforms.

Image-based abuse

In October 2017, eSafety launched an image-based abuse portal. It provides reporting options, support and resources to Australians who have experienced image-based abuse, as well as their families, friends and bystanders.

Image-based abuse affects a significant proportion of the Australian population, with one in ten adults

[11] Please see Part IV: Consultation Process of this report for an overview of the youth consultation exercise.

[12] Office of the eSafety Commissioner, 2018, State of Play – Youth, Kids and Digital Dangers.

[13] In eSafety's National Survey young people were asked to identify actual risks that they had experienced, while the online forum asked for perceived risks that young people faced.

[14] Since 2017, eSafety has helped more than 1,100 adults with cyber abuse enquiries (at date of publication).

experiencing the sharing of nude or sexual images without consent. One in five have been a bystander to this type of crime[15]. Image-based abuse is more prevalent among certain population groups, such as Aboriginal or Torres Strait Islander peoples (25%), younger women (24%) and those who identify as LGBTI (19%). Perpetrators of image-based abuse are typically someone that the victim knows, either an offline friend (29% of reports), an ex-partner (13%), a current partner (12%) or a family member (10%).

Between October 2017 and December 2018, eSafety received over 600 reports of image-based abuse.

Tech-facilitated abuse

The eSafety program eSafetyWomen empowers women to manage technology risks and abuse and take control of their online experiences through three pathways: awareness-raising through targeted social media; training for frontline family and domestic violence workers; and advice and resources for women to help them stay safe online in the face of family and domestic violence.

A recent national study found that 98% of respondents had experienced technology-facilitated stalking and abuse as part of their domestic violence experience[16]. From reports made to eSafety via our eSafetyWomen workshops, we know that current and former partners in a domestic violence situation frequently use fake or impersonator accounts to perpetrate abuse on a woman.

Child sexual exploitation and abuse

eSafety administers the Online Content Scheme, which allows Australian residents and bodies corporate to report illegal and offensive online content to the eSafety Commissioner.

Our Cyber Report team prioritises taking action on child sexual abuse material (CSAM) within two working days with a view to having the material removed. In 2017-18, the Cyber Report team conducted 8,040 reports into online CSAM, a 57% increase on the year before.

The role of social media platforms and online forums as a gateway to online grooming, sexual solicitation, uploading of sexually explicit photos or videos is widely recognised.[17, 18] It is clear that the online world can facilitate the sexual exploitation and abuse of young people online, and evidence from eSafety's regulatory investigations points to the ever-evolving nature of this particular crime.

The lack of information shared by online services and platforms about the scale and nature of risks and harms on their services has hindered a comprehensive assessment of the prevalence of online risks and harms. Regardless, it is undeniable that the risks and harms experienced by users are dynamic and fluid. Continuous and sustained efforts are needed to identify, understand, prevent and intervene effectively to ensure user safety. It is therefore important that risks and harms are seen as 'living', and not static. User safety considerations, and the SbD Principles themselves, will also need to be regularly reassessed.

Abhorrent Violent Material

Events such as the April 2019 live-streamed terrorist attack against two mosques in Christchurch New Zealand, have demonstrated the risk of online services and platforms being used to bring attention to violent and extremist actions. This can both compound the harm experienced by the victims of such actions and contribute to the radicalisation of end-users.

Following the Christchurch attack, new legislation — Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2009 *(Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2009 (Cth)* — was introduced in Australia to reduce the impact and reach of abhorrent violent material that is streamed or made available online by perpetrators and their accomplices. The legislation creates offences to ensure that internet, hosting or content service providers are referring abhorrent violent material to law enforcement when made aware of the material, and that hosting and content services are removing abhorrent violent material that is capable of being accessed within Australia in a timely manner. The legislation also creates a process by which eSafety can provide a written notice to a hosting or content service provider to alert them that abhorrent violent material is accessible through their service, and to make them aware that they may commit an offence if they fail to cease hosting or remove the material expeditiously.

[15] Office of the eSafety Commissioner, 2017, Image-based abuse National survey: Summary report.

[16] ReCharge: Women's Technology Safety, Legal Resources, Research & Training, National study findings 2015. Found at www.smartsafe.org.au/tech-safety-hub/resources/research

[17] Europol, 2018, Internet Organised Crime Threat Assessment.

[18] Child Dignity Alliance, 2018, Child Dignity Alliance: Technology Working Group Report.

III. The foundations of the Safety by Design Principles

Human rights and digital ethics

A broad array of alliances[19], coalitions[20, 21, 22], frameworks[23], guidance[24, 25, 26], codes of practice[27] and principles[28] focused on online safety have been developed globally since the early 2000s. The main objective of these initiatives is to protect young people online, and help parents and guardians protect their children online. The SbD Principles drew on this objective — balanced against privacy and security. Ethical and human rights standards and concepts were also used to underpin and guide the development of the Principles. This is in line with the work being progressed by the Australian Human Rights Commission, as outlined in their issues paper which was published in July 2018[29]. All the SbD Principles are complementary, rather than mutually exclusive.

Human rights

The online world poses new and broad-ranging challenges. One of the most significant is how to meet the human rights responsibilities that apply to children — that is, how online platforms and service providers ensure children's rights by addressing the provision of services, online protections, and children's online participation. A child's right to safety lies at the heart of the Convention on the Rights of the Child[30] (UNCRC). The online world therefore has a duty to ensure that human rights standards are placed at the centre of online design, content and functionality. Consideration and care have been taken to ensure that each of the SbD Principles balance an individual's right to provision, participation and protection.

The National Principles for Child Safe Organisations[31], endorsed by the Council of Australian Governments and further developed by the National Children's Commissioner, underlie that it is the responsibility of all institutions to uphold the rights of the child, and act with the best interests of the child as a primary consideration. Many elements of the National Principles, which specifically address how organisations can create safe online environments for children in their care, are embedded in the SbD Principles[32].

The Council of Europe's Recommendation[33] to respect, protect and fulfil the rights of the child in the online environment sets out clear expectations for businesses in terms of how they can meet their human rights responsibilities to children in the online world. Elements of the expectations raised in the recommendation have been applied to the wording of the SbD Principles.

[19] Alliance to Better Protect Minors.

[20] Better Internet for Kids: CEO Coalition.

[21] The ICT Coalition for Children Online.

[22] Coalition for Digital Intelligence.

[23] European Framework for Safer Mobile Use by Younger Teenagers and Children.

[24] ITU and UNICEF, 2015, Guidelines for Industry on Child Online Protection.

[25] UK Council for Child Internet Safety, 2015, Child Safety Online: A Practical Guide for Providers of Social Media and Interactive Services.

[26] UK Council for Child Internet Safety, 2013. Good practice guidance for the providers of social networking and other user-interactive services.

[27] EU Code of Practice for Disinformation.

[28] Safer Social Networking Principles for the EU.

[29] Australian Human Rights Commission July 2018. Human Rights and Technology: Issues Paper.

[30] In 1989, the United Nations adopted the Convention on the Rights of the Child, the most widely ratified human rights treaty in the history of the UN.

[31] National Principles for Child Safe Organisations. Found at: https://childsafe.humanrights.gov.au/national-principles

[32] See Recommendations 6.4, 6.5, 6.6, 7.8, 7.9, 7.10 and 7.12 in particular.

[33] Recommendation CM/Rec (2018)7 of the Committee of Ministers to member States on Guidelines to respect, protect and fulfil the rights of the child in the digital environment.

Issues of digital ethics, digital governance and digital accountability

The online world continues to reshape our understanding of identity, human experiences and social interactions. We have moved from an entirely analogue existence, to one that is deeply integrated and intertwined with the digital.

While innovations in technology, machine learning and artificial intelligence have the potential to radically transform our lives, there is growing concern over how we are constrained by the online organisations and products with which we interact and engage[34]. We are also increasingly aware of how individuals can be governed, marshalled and influenced by online services. Further, we are starting to question whether digital innovation and technology actually improve human existence. In this environment, digital ethics, governance and regulation play an increasingly important role.

Digital ethics

There are three normative forces that shape and guide the development of the online world: digital ethics, digital governance and digital regulation[35]. Digital ethics play a role in shaping policies, procedures and standards for online governance, as well as the legislation and rules by which the online world is run. There is a clear ethical duty for businesses, governments and international bodies to preserve the rights and dignity of users online and to protect them from abuse and exploitation.

Digital ethicists, governments and, increasingly, those working within the technology sector, are calling for the urgent need to anticipate and steer the ethical development of technological innovation by all those in the digital ecosystem. These efforts have ignited a range of initiatives[36], consortiums[37] and frameworks[38] over the last twelve months, seeking to secure a more ethical, value-focused and human-centred approach to the development of technologies.

Recognising the importance of digital ethics, the SbD Principles have been developed within such an ethical framework. They seek to secure a more ethical and human-centred approach to the development of technology. The Principles also acknowledge that digital technology, human rights and digital ethics need to develop iteratively together in order to create safe user environments. The SbD Principles should therefore be seen as a living-document, and one that will need to be updated to reflect the changing environment.

Digital governance

As a global and distributed phenomena, there is no single stakeholder or group of actors who are responsible for governing the online world. Digital governance is most frequently defined as 'the practice of establishing and implementing policies, procedures, and standards for the proper development, use and management of the infosphere[39]. As such, a consistent narrative has formed that there is a shared responsibility among all stakeholders to ensure that users can navigate the online world safely.

The prerequisite underpinning a shared responsibility model is for all stakeholders to have an equal or, at least, a shared understanding of the issues and processes at stake. For user safety though, there is, at best, an uneven distribution of knowledge. There is little data or analysis to show how aware end-users are about the design, development and management of online services and platforms.

Noting this gap in the evidence, eSafety undertook an extensive consultation with young people, further described in section IV. These findings guided our drafting of the SbD Principles to weight the responsibility for user safety towards vendors, designers, engineers and manufacturers.

[34] Child Dignity Alliance, 2018, Child Dignity Alliance: Technology Working Group Report.

[35] Floridi, L 2018, 'Soft Ethics and the Governance of the Digital'. Philosophy & Technology, Vol 31(1), pp 1-8.

[36] For examples, please see UK Centre for Data Ethics and Innovation, Singapore's Centre for AI and Data Governance, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

[37] For examples, please see: Centre for Humane Technology, Ethical Tech Initiative, Ada Lovelace Institute.

[38] For examples, please see: IBM's Everyday Ethics for Artificial Intelligence; Institute for the Future and Omidyar Network's Ethical OS Guide, UK Government's Data Ethics Framework, Google's AI Ethical Framework. The Ethics Centre, Ethical by Design: Principles for Good Technology

[39] Floridi, L 2018, 'Soft Ethics and the Governance of the Digital'. Philosophy & Technology, Vol 31(1), pp 1-8.

[40] For examples, please see Netzwerkdurchsetzungsgesetz Act (NetzDG), which was passed into German law in June 2017.

Digital regulation

There are real challenges in achieving a consistent approach to the regulation of online and digital technology globally. Individual countries have passed laws to ensure internet providers and services restrict illegal content online[40], while operational measures[41] have been set at the European level to tackle illegal content online. The UK Government has announced the development of Codes of Practice for social media services[42], and Age-Appropriate Design Codes[43] in order to set minimum standards for online safety and data protection. An abundance of inquiries and hearings by different governments have looked at the impact of the online world on: hate crime[44] , children[45], elections[46], filtering practices of social media services[47], fake news[48], intimidation in public life[49], and on competition in media and advertising services markets[50].

There is global recognition of the need to create and develop responsible technology, that takes a holistic view of consumer welfare and considers the broader societal impacts of online and technological products and services. However, a patchwork of legislation, regulation and governance structures contribute to inconsistent and fragmented systems, resulting in gaps in oversight and accountability.

To help overcome some of these regulatory challenges, eSafety's SbD Principles promote and build into the development and design of online and digital technology an informed, transparent and accountable decision-making process that works to minimise harm. From our experience of regulating for specific online harms within Australia, we concluded that the only way to truly get ahead of the multiple safety challenges online is through a combination of cultural and behavioural change, and the development of a safer online infrastructure.

Interconnectedness of other design principles

Privacy by Design

The seven principles of Privacy by Design (PbD) were developed in the early 1990s by Dr. Ann Cavoukian, then the Privacy Commissioner of Ontario, to ensure that privacy assurance was an organisation's default mode of operation. In 2010, a resolution was passed that recognised PbD as an essential component of fundamental privacy protection[51].

'The principles seek to proactively embed privacy into the design specifications of information technologies, organizational practices, and networked system architectures, in order to achieve the strongest protection possible.[52]'

[41] On 1 March 2018 the EU Commission recommended a set of operational measures to be taken by companies and member States to step up on tackling illegal content online, before it determines whether legislation is required.

[42] Please see UK's Internet Safety Strategy and Online Harms White Paper.

[43] The UK's Information Commissioner's Office is seeking to develop an Age-Appropriate Design Code, a requirement under the Data Protection Act 2018, in order to set design standards that are expected from providers of online 'Information Society Services', which process personal data and are likely to be accessed by children.

[44] Please see UK Parliament's Home Affairs Committee Inquiry into 'Hate crime and its violent consequences', and its subsequent report: 'Hate crime: abuse, hate and extremism online'.

[45] Please see UK's House of Lords Select Committee on Communications report 'Growing up with the Internet', and the UK's Information Commissioner's Office call for evidence for an Age-Appropriate Design Code.

[46] Please see Australia's Electoral Integrity Assurance Task Force, the UK's National Security Communications Unit, US Senate Select Committee on Intelligence hearing on 'Foreign Influence Operations' Use of Social Media Platforms'.

[47] Please see US House Committee on The Judiciary hearing on 'Filtering Practices of Social Media Platforms'.

[48] Please refer to the recent law passed in France, in November 2018, the Joint Communication to the European Parliament, the European Council, The Council, The European Economic and Social Committee and the Committee of the Regions Action Plan against Disinformation, the European Political Strategy Centre's High-Level Hearing on 'Preserving Democracy in the Digital Age'.

[49] Please see UK's Committee on Standards in Public Life review on 'Intimidation in Public Life'.

[50] Please see Australian Competition and Consumer Commission (ACCC) 'Digital platforms inquiry', and its preliminary report which was published in December 2018.

[51] Resolution on Privacy by design was passed at the 32nd International Conference of Data Protection and Privacy Commissioners in Israel, October 2010.

[52] Cavoukian, A 'Privacy by Design, The 7 Foundational Principles: Implementation and Mapping of Fair Information Practices'. Found at www.iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf

The General Data Protection Regulation[53], a regulation in European Union law which came into force in May 2018, the Australian Privacy Principles[54], Australian Government Agencies Privacy Code[55] and other recent legislation implemented around the globe has created stronger rules on data protection. Privacy by Design and Privacy by Default concepts have now become legal requirements, and work to ensure that privacy and data protection are considered at the earliest stages of the developmental process.

Privacy and safety are closely interrelated. The fundamental data protection principles of data minimisation, data portability and purpose limitation are as much an imperative for user safety as they are for data protection and privacy. In all cases, it is essential that appropriate technical and organisational measures are incorporated into the design and development of online services and products, and that safeguards are integrated into all processes to protect an individual's rights.

Security by Design

Cybersecurity is generally understood as the protection of internet-connected systems from cyberattacks. In an interconnected, digital and networked world, an overarching security strategy and architecture have become business imperatives.

To guide business risk-mitigation, multiple cybersecurity standards, committees, frameworks and governance arrangements exist at both a national and international level.

Like PbD, Security by Design is proactive, recognising that — by necessity — it is an ongoing process of planning, monitoring and adapting to the changing nature of cyber threats. At their most basic, security principles focus on designing controls that prevent misuse of the application by different types of malicious parties. Again, the relationship between security and user safety, as between security and privacy, is closely intertwined, because cyber attacks target individual users as much as they attack the infrastructure of the product or service itself.

Inclusive design

Inclusive design[56] considers the full range of human diversity in the design and development of products and services. This includes ability, language, culture, gender, age and other forms of human difference. As its name implies, inclusive design is inherently human centred and is much more than just accessibility compliance. It ensures that anyone, in any situation, can use, harness and reap the benefits of online services and products.

Most online services tend to be developed with a general audience in mind, namely adults. However, when considering user safety, an inclusive approach is essential in ensuring that the processes, policies and tools developed to mitigate against risks take into account the permanent, temporary, situational or changing capabilities of the user population. Design can be either enabling or disabling and, as such, it is important to ensure that online environments empower and provide autonomy to all users, particularly when their safety is being, or is at risk of being, compromised. The SbD Principles recognise and advance these aims.

[53] Please see European Commission's General Data Protection Regulation, regulation (EU) 2016/679.

[54] The Australian Privacy Principles are derived from Schedule 1 of the Privacy Amendment (Enhancing Privacy Protection) Act 2012, and which are made up of 13 Privacy Principles. Found at **http://www.comlaw.gov.au/**

[55] The Australian Government Agencies Privacy Code commenced on 1 July 2018, setting out specific requirements and practical steps agencies must take to comply with Australian Privacy Principle 1.2.

[56] Please refer to the Centre for Inclusive Design, found at **http://centreforinclusivedesign.org/**.

IV. Consultation process

In June 2018, eSafety undertook a three-pronged consultation process to guide the development of the SbD Framework. This included consultation with:

1. Industry and other companies, trade bodies or representative organisations that were identified as having synergies or an interest in safeguarding users online.
2. Parents and carers.
3. Young people (aged 14-17).

The purpose of the consultation was to ensure that the SbD Principles encapsulated all aspects of user safety in a meaningful, achievable and practical way. We felt it was imperative that the Principles reflected the needs and expectations of those in need of special protections, such as children and young people and those in vulnerable circumstances, as well as parents and carers.

Industry

Online Safety Consultative Working Group

The Online Safety Consultative Working Group (OSCWG)[57], which provides independent expert advice to eSafety, played a critical role in shaping and developing the SbD Principles. A total of 29 organisations are represented on the OSCWG.

Tiered service providers

One of the core functions of eSafety is the Cyberbullying Complaints Scheme, as outlined in Part 4 of the Act. This scheme provides a complaints mechanism for Australian children who experience cyberbullying. eSafety works closely with social media services to help remove harmful material and provides support for any young person making a report and their family. eSafety consulted with companies participating in the cyberbullying scheme's tier arrangement regarding the SbD Principles. There are currently 14 social media services involved[58] and all were consulted on the SbD Principles in July and August 2018.

Key stakeholders involved in safeguarding users online

eSafety also consulted with organisations that demonstrated best practice in Safety by Design, those with a considerable interest in safeguarding or protecting citizens' rights online, or who represented the SME or start-up community. Consultations took place between July and September 2018. A total of 14 individual organisations[59] were invited to participate.

Parents and guardians

In July and August 2018, eSafety conducted a nationally representative survey (2018 Parent Survey) of 3,520 parents and guardians of children aged 2 to 17 using multiple online panels (n=3,044) and computer assisted telephone interviewing (n=476). The survey included questions about parents' perception of the importance of safety by design features and the role of the tech industry in keeping young people safe online.

The perspective of parents is crucial, as research conducted by eSafety to date suggests that young people approach parents as the first port of call when dealing with online hate, bullying and violent content.

[57] OSCWG members represent industry (Google, Twitter, Facebook, Telstra Foundation, Optus), law enforcement (the Australian Federal Police and Queensland Police Service), academia (Western Sydney University and RMIT), not-for-profits (BraveHearts Foundation, Carly Ryan Foundation, ReachOut, Alannah and Madeline Foundation, and Project Rockit), government departments (Department of Communications and the Arts, Department of Education and Training, Department of Social Services, Office for Women, Department of Home Affairs, National Children's Commissioner), trade or representative organisations (Enex TestLabs, Australian Mobile Telecommunications Association, DIGI Group, IGEA, Communications Alliance) and community based specialists (Australian Seniors Computer Clubs Association, National Association for Prevention of Child Abuse and Neglect, Reality & Risk Community Education Project, an independent psychologist and Cyber Safety Solutions).

[58] Tier 1: airG, Ask.fm, Flickr, musical.ly, Roblox, Snapchat, Twitter, Yahoo!7 Answers, Yahoo!7 Groups, Yubo. Tier 2: Facebook, Google+, Instagram, YouTube.

[59] Organisations included those representing best practice: Lego, Centre for Inclusive Design and individual consultants and child safety advocates; those safeguarding and protecting citizens (including the Australian Human Rights Commission, Infoxchange and THORN) and SME or start-ups such as Change.org.

Young people

Following the findings of the 2018 Parent Survey, eSafety held a five-day, structured online forum with 123 young people aged 14 to 17. The participants were diverse in terms of their gender, ethnicity and socio-economic background. They also came from a range of different schools in both metropolitan and regional settings. To participate, the young person had to be a current internet user who engages in online activities several times per week. This age group was selected to ensure that participants could provide insights based on their experiences as a user of various social media platforms and apps.

The forum explored participants' views about how they felt they could benefit from being online while being adequately protected against risks and harms. The content of the forum focused on:

- what makes young people feel empowered online and whether they are aware of their rights online
- young people's perceptions of online risks, the harms they are exposed to online and what should be done to mediate these
- who they believe is responsible for keeping them safer online
- whether young people think that apps, games and social media platforms provide the right environment and tools to help users navigate critical aspects of user safety e.g. bullying, image-based abuse, harassment, exposure to inappropriate material, unwanted contact
- what tools and features they feel would better assist in navigating the above issues
- the importance of key user safety features and whether they are adequate and/or intrusive
- their expectations of industry in relation to keeping young people safe online.

Findings

Children and young people were a focal point in the development of the principles, as research shows that their rights in the online world are far from being met or addressed[60]. This is despite the fact one in three internet users globally are children[61], children are early adopters of online products and technology, and they are most vulnerable to the potentially harmful impacts.

The important things about being online

Young people were asked to imagine what their lives would be like if they were not allowed to use the internet for leisure. They identified seven online activities as most important:

1. **Communication:** The ability to communicate online through social media is the most important aspect of being online for young people. This was particularly important for reaching out to a wider network domestically and abroad with whom communication would not otherwise be possible. Young people emphasised that social media provides them a voice and a sense of belonging. Females had a higher desire to share their ideas and opinions via social media than males and talked about social media as an important form of self-expression.

2. **Freedom to research:** The ability and freedom to research anything online they like regardless of the time of day. Young people expressed the importance of being able to research topics such as life skills, current affairs and educational topics outside of the curriculum.

3. **Entertainment:** The internet provided entertainment and streaming capability. YouTube was synonymous with the word 'entertainment' regardless of gender. While the type of content was varied, the means of accessing engaging content that matches individual interest is part of their lives and helps them relax, learn new things, avoid boredom and have fun.

[60] Livingstone, S and Third, A 2017, 'Children and young people's rights in the digital age: an emerging agenda'. New Media and Society, 19(5). pp. 657-670.

4. **News:** A source of unbiased news. The majority stated that they do not watch broadcast television. Instead they rely on the internet to stay informed on the latest news and current affairs.

5. **Gaming:** Gaming, regardless of gender, was highlighted as a way to relax, have fun and avoid boredom.

6. **Shopping:** While online shopping was not seen to be as important as communication, most viewed the ability to shop online as vital to their way of life.

7. **Music:** Access to music was also very important for young people and most could not picture living without music streaming services like Spotify.

Young people also highlighted that their internet use has allowed them to form new relationships, share ideas and their creativity, start online businesses, apply for jobs and express themselves.

Online rights and perceived risks

When discussing negative experiences online in the forum, 64% of young people admitted that their awareness of potentially negative online experiences affected the way they interact online. Females were more likely to believe that these negative experiences could happen to young people online than their male counterparts. (Please see section II for a more comprehensive discussion of the risks that young people encounter online.)

Young people were divided (58% - Y and 42% - N) in their views of whether existing safety features on platforms, apps and devices would mitigate online risks. For those who thought that existing features were sufficient, the efforts made by social media platforms and other app developers to i) ensure the privacy of their users and ii) provide users with control over how their information is used, were acknowledged. However, for those who thought that existing safety features were inadequate, they felt that they had very little control, and were at the mercy of the changes that app and platform developers make to their settings. Account hacking, the creation of fake accounts and unwanted exposure to violence or sexual content were the areas where current safety features were seen as inadequate.

A recurrent theme voiced by a significant proportion of young people was that users themselves played a significant role in shaping the online environment. Young people felt that there was often very little that developers could do to prevent negative peer-to-peer interactions related to bullying, the spreading of rumours and targeted discrimination. It is thus not surprising that young people flagged rules, community standards and guidance to the user base on acceptable behaviour, as important.

The three-pronged approach to prevention

Young people were clear about how platform and app developers could target the affected user, bystanders and the 'user collective' in their interventions to ensure a safe online environment. They suggested that interventions which target the affected user should focus on:

i. empowering users to directly address the offending user or content (such as creating awareness and improving blocking, muting and reporting features)

ii. greater punitive measures for users who cyberbully

iii. encouraging users to engage with support services when needed.

Education and anonymous reporting were suggested as key bystander interventions. Young people felt platforms and apps should do more to educate bystanders about steps they could take and should also create awareness of the available support pathways for victims. They believed that anonymous reporting would encourage bystanders to intervene and help victims, given the fear of retaliation and social consequences bystanders face when 'upstanding'.

When targeting interventions broadly to all users, young people focused on:

- Awareness of the terms and conditions of using the platform or app, and the consequences of breaching them. For example, suspension of offending accounts and blocking of IP to avoid the recreation of accounts.

- Awareness of privacy, reporting and blocking features and encouraging the use of these features when required.

[61] Livingstone, S, Carr, J & Byrne, J, 2015, 'One in Three: Internet Governance and Children's Rights. Global Commission on Internet Governance', paper series No 22.

- Improved monitoring to ensure users are safe from harmful behaviour and content.
- Promotion and recognition of positive online behaviour. For example, using celebrities to champion respectful relationships.

Awareness and importance of safety features

The overarching finding from the online forum was that young people prioritised safety features related to control and monitoring. Young people viewed themselves as active agents responsible for their own online safety. They believed that features that offer control allow them to be proactive in keeping safe. They also felt that monitoring was a vital protection, as it allows platforms to identify and proactively remove inappropriate or negative content.

The following quote reflects the attitudes of young people in the online forum:

'I expect the Australian online industry to value the safety of users above anything else so that they can feel safe and comfortable online. To ensure this, I think all online platforms should offer support systems and features to minimise cyberbullying, catfishing and other possible dangers faced by online users. These features should include blocking and reporting for victims and bystanders, punishments for the people engaging in online illegal activity and improved vetting process to stop catfishing and the dangers attached to it. I believe the availability of all these features will ensure users can navigate the online world freely and safely.' (Female, 14, NSW — Sydney)

Young people's awareness of safety features goes hand-in-hand with the importance they place on these features in ensuring safety.

Young people had the greatest awareness of features that provide control (98%) and monitoring (93%), and much lower awareness of features related to breaching of rules (84%) and support (64%).

Similarly, young people viewed settings that enable them to control their online experiences as being the most important in ensuring their safety online, with less importance placed on monitoring, rules and support.

The majority of young people view control features as empowering — enabling them to take charge of their online accounts and user experience (86%) and find these features effective in doing so (74%). However, a noted limitation is that certain social media platforms allow blocked users to view certain parts of the blocker's profile, and this can make users uncomfortable about using this feature.

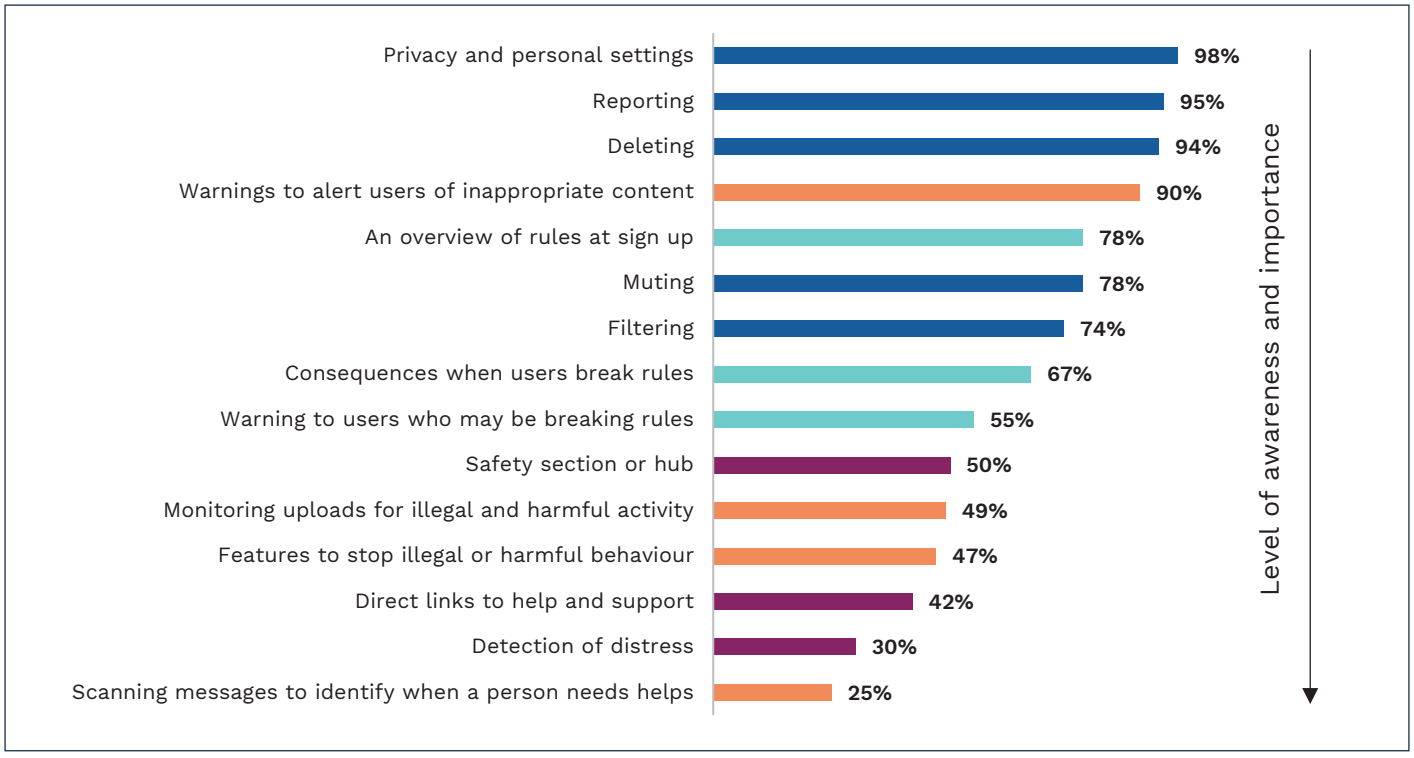


Figure 2: Young people's awareness of individual safety features from 2018 youth consultation for Safety by Design.

When it comes to monitoring, young people were split in their views. While the majority — 71% — believed that monitoring systems scanning messages and content in the background were helpful in preventing negative experiences, over half (57%) were uncomfortable with these features running in the background. A sizeable minority found monitoring features intrusive (42%) and were unsure about their effectiveness in ensuring online safety (43%).

Rules and community standards were viewed as ineffective in shaping how users behave online (53%) and preventing unacceptable user behaviour or actions (70%). There was agreement that rules and community standards were only useful if reinforced by apps and platforms with consequences given to those in breach.

While support information and resources were viewed by the majority, 81%, as useful to those who require assistance, they were mostly seen as a last resort rather than a tool in ensuring safety.

Safety by Design

While an overwhelming majority — 91% — believed that the online industry has a duty to keep their young users safe online, they recognised that safety is a shared responsibility between users, parents, schools, the wider community, government and law enforcement.

Young people were asked to prepare a vision statement, to lay out what they want and expect of the Australian online industry to help users navigate the online world, freely and safely. Their collective vision statement can be found in full in Appendix A. Seventeen themes were drawn out from the discussions, with over a third of respondents highlighting the following as areas they felt were important:

1. Empowering users by giving them greater control of their own safety and experiences online.
2. Providing clear rules and guidance that are easy to read and highly visible.
3. Providing users with safety tools and features, namely ways to make reports and to block both people and content.
4. Imposing sanctions and consequences for violating the rules of the site.
5. Using scanning and filtering technology to ensure user safety is upheld on the site and users are not exposed to inappropriate or sensitive content.

Young people echoed similar sentiments to those in their vision statement, when ranking 12 safety by design concepts that eSafety proposed (see page 15). 'Keeping people safe online is a shared responsibility' ranked the highest among young people.

Young people found all the proposed concepts to be clear and easy to understand. They identified transparency around the management of user information as one area of omission in the principles. They argued that apps and platforms need to increase transparency about how user information is stored and used. More specifically, young people emphasised the importance of clearly naming third party companies that have access to their user information and how their information will be used.

Proposed concepts

1. Keeping people safe online is a shared responsibility.
2. There should be clear guidance about how people should behave on the platform. People who abuse the rules should be penalised fairly. People who are targeted by abuse should be told how the platform will deal with the problem.
3. Safety and privacy settings should be set at the highest level of protection for users from the start.
4. Industry has a role to play in keeping people safe on their platforms.
5. Industry should be aware of what risks and harms exist online and have people and processes in place to deal with those issues on their sites.
6. Users should be given the tools to be able to control how they want to engage and interact online.
7. Industry should take steps to stop users from being targets of, or exposed to, abuse and harm on their platforms.
8. The best interests of the user should be at the heart of the app, game or platform.
9. Support should be provided to users when they need it, especially when they are not feeling safe or are feeling low.
10. People are told what type of risks and harms the platform has to deal with, and how well they are dealing with those problems.
11. Industry should tell the public what steps they are taking to keep their users safe.
12. Users and experts outside of the company should help guide and develop their rules and policies

Parents and carers

Almost all parents and carers, 95%, placed importance on online safety. Parents' level of confidence in dealing with cyberbullying steadily increased with the age of their child — 38% of parents with children aged 2 – 5 years compared with 47% of parents with children aged 6 – 12 and 51% of parents with young people aged 13 – 1 7. Similarly, parents' confidence in dealing with online threats such as contact with strangers increased as their child grew older (40% of parents with children aged 2 – 5 versus 47% of parents with children aged 6 – 17).

Parents of children aged 6 – 17 years were more likely to feel that technology companies needed to do more to build in safety features (see Figure 3). A third of parents struggled to offer a response to a statement which raised questions about their awareness and understanding of built-in features in general. While 52% of parents understood how to use the safety features on social media, apps and games, 28% did not understand how to use the features, and the remainder struggled to choose a response. This, in turn, affected whether they showed their child how to use these features when interacting online, with only 51% stating that they did (45% parents with children aged 2 – 5 versus 53% of parents with children/ young people aged 6 – 17). While the majority of parents stated that parental control was important in how they limited their child's exposure to inappropriate content such as pornography, this decreased with the age of their child — 83% of parents with children aged 2 – 5 years, 80% of parents with children/ young people aged 6 – 12 and 64% of parents with young people aged 13 – 17).

Parents with younger children were more likely to use age guidelines in relation to their child's use of social media, apps and games (75% of parents with children aged 2 – 12 years versus 52% of parents with young people aged 13 – 17). Importantly, a third of parents did not know where to get help in relation to online safety issues affecting their child.

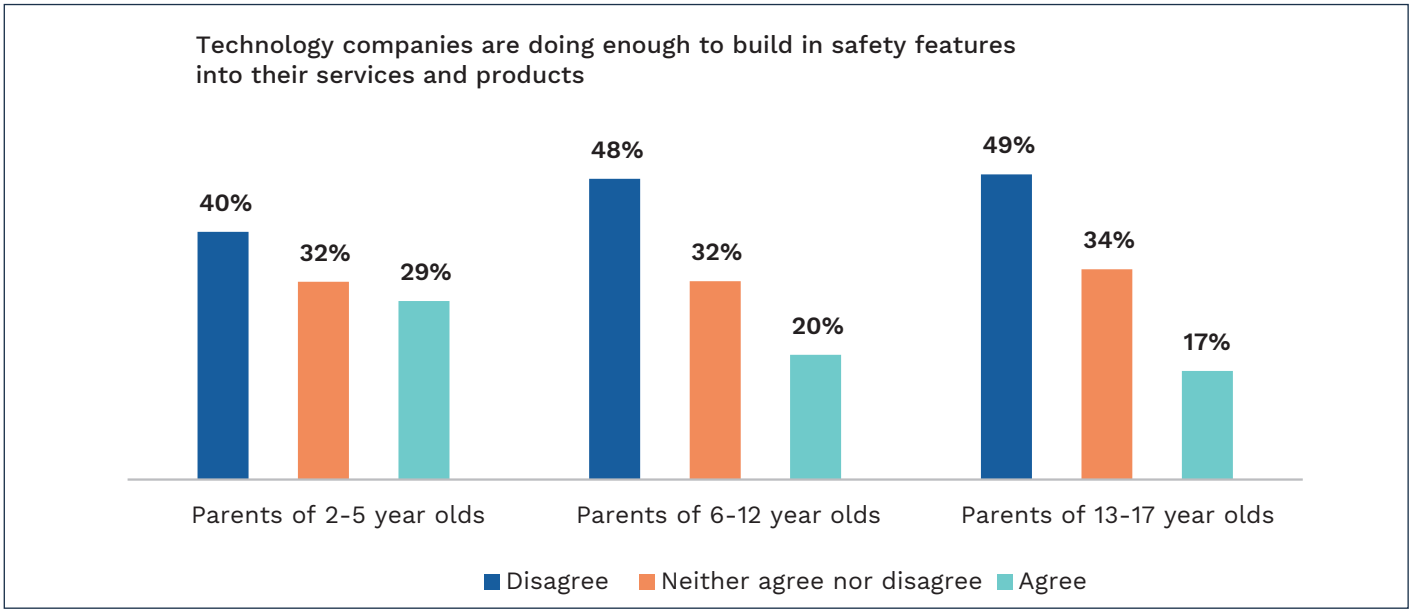


Figure 3: Question from 2018 Parent Survey.

Regardless of their child's age, or their confidence in dealing with negative experiences and online threats or parenting styles, parents were consistent in their ranking of the top five features that are important for technology companies to incorporate to protect children online:

1. Ensuring the highest privacy settings are in place by default.
2. Better measures to restrict access to online content for appropriate ages.
3. Registration processes that prevent users from accessing services that are not targeted at their age range.
4. Automatic flagging of inappropriate language and behaviour to allow users to reflect on what they are about to post.
5. Features that limit who has access to a user's post.

Industry

Industry's current commitment to user safety has been clearly demonstrated in recent consultations conducted by eSafety. This was encouraging and not inconsistent with feedback received from social media services about the collaborative and cooperative regulatory approach taken by eSafety when resolving complaints. There was broad support for ensuring that user safety is adequately addressed, especially at the design stage. Industry representatives indicated that the SbD Framework was a much-needed resource, particularly for smaller-sized ventures and start-ups.

Some industry players stated that much of what the SbD Principles proposed was already being carried out by many online services, and, for some, a safety by design approach is already an established process[62]. For smaller players, a template of user safety considerations that could be presented to the business, design teams and engineers was considered of great benefit. Some policy or trust and safety operators within smaller companies felt that the SbD Framework would enhance their attempts at driving user safety considerations in their businesses and would give user safety greater prominence.

There was strong support for the fact that the SbD Principles were underpinned by human rights responsibilities and set within an ethical framework. The need to address the specific protections for children and young people was acknowledged by almost all those consulted as being of utmost importance, and some felt that not enough attention was placed on this in the SbD Principles. However, other stakeholders expressed the importance of ensuring that the safety needs and vulnerabilities of other targeted groups, and those with distinct and

[62] No documented evidence was provided to eSafety during the consultation process, only anecdotal. However, during the consultation process, Facebook's Global Head of Safety released a public document stating that it adopts a 'safety by design' approach to product development.

diverse characteristics, should be addressed more overtly, and attention should not just be placed on children and young people[63]. In particular, undertaking an age-appropriate design of services was felt to be too prescriptive and not actually in the best interests of the user. The need to adopt an inclusive approach to design was considered more important to many, particularly in relation to safety features and tools. Language, culture, age, gender, ability and all other forms of difference were flagged as important considerations in the development of any safety feature or approach to user safety.

Stakeholders were generally supportive of a principle-based approach to user safety. However, there was some concern that the SbD Principles were taking too much of a one-size fits all approach to tackling user safety, and that the measures raised were too prescriptive and exhaustive. They submitted that a more principle-based approach would ensure that the SbD Principles were not seen as a compliance checklist, and better reflect the ethos of empowerment and rights-recognition upon which they are founded. A number of stakeholders highlighted that online services are not homogenous, in either design, function, culture or usage. Flexibility in the principles was seen as essential, in order for them to address stakeholders' divergent practices, functions, layering of services and products, and multiple business models.

Industry expressed concern that the measure which required services to have reactive, active and proactive safety measures in place was too prescriptive and unrealistic for start-ups. In addition, concerns over legal liability in using proactive tools and features to surface and flag harmful behaviour were raised by a number of companies. However, others highlighted that providing specific examples of tools within the principles would quickly render the principles irrelevant as new technology and emerging trends would arise in relation to user safety and harms. The fast-paced development of new safety features and tools was seen as incredibly positive, but a number of stakeholders stressed that inclusion of features is often dependent on the user-base and experiences that the platform allows. While one set of features may work well within one type of service or community, their inclusion in another context may not translate into a positive impact.

For SMEs and start-ups, some elements of the principles were considered a little overwhelming. Either a graduated approach, or a clearer indication upfront that the principles were a benchmark or 'gold standard', would make the principles more palatable. A number of stakeholders mentioned that it would be useful to scale the safety measures in the principles, to indicate which items are considered essential, and which are less urgent.

When considering feedback about the prescriptive nature of the draft principles, stakeholders were keen to express caution over the principles being too burdensome. Some expressed concerns that the principles appeared to be a prelude to further government regulation of their activities. Many questioned how eSafety would ensure industry compliance. However, these opinions were juxtaposed with other views that saw safety by design as providing companies with a competitive advantage. Consumer awareness and expectations were seen to be shifting, particularly within the youth population, and so stakeholders could see that companies who embed a more value-centred and ethical approach to their products would stand the test of time.

A number of stakeholders emphasised the need for clarity about who the principles specifically target. Terms such as 'online platforms' and 'online services' were considered too broad. In particular, the term 'platform' was considered too indistinct and too restrictive, given that definitions vary across online technologies and sectors. Most felt that the principles were applicable to services that offer users the opportunity to interact or connect, and as such, that the principles should target those that provide interactive functionality in their design and practice.

Transparency and accountability were the areas that raised the most concerns for industry and key stakeholders. Industry representatives repeatedly highlighted that they would be hesitant to publish details of their community standards, technical security measures and statistics on abuse/harm metrics. The main rationale provided was to ensure that individuals could not 'game' the system, or incorrectly compare or misinterpret processes or safety metrics[64]. The publication of statistics was particularly contentious. For some stakeholders,

[63] These include, but are not limited to, Aboriginal and Torres Strait Islander peoples, those from culturally and linguistically diverse backgrounds, those living with disability, those with mental health disorders, those in out-of-home care, those living in regional and remote areas, those identifying as lesbian, gay, bisexual, transgender or intersex, and women.

[64] Of note, during the course of the consultation process a number of companies published transparency reports and copies of their community standards for public consumption.

it was unclear what benefit would come from the publication of safety data, effectiveness of reports and impact of community standards and terms of services, specifically in relation to users themselves. In particular, stakeholders pointed out it would be challenging to transform standardised safety metrics into meaningful comparisons across platforms. That said, there was unanimous agreement about the need for transparency about safety policies, processes and moderation practices in order to better empower and educate users.

Stakeholders also expressed caution about whether some elements of the principles crossed over into other regulatory remits or well-established principles within the domains of data protection, privacy and security. Two measures, in particular, were questioned by a significant number of stakeholders:

1. Mechanisms to protect user identity and personal information made available, adaptive and open to recourse.

2. Providing users with the ability to access, obtain and reuse their personal data for their own purposes across different services, and to request its erasure or modification.

The protection of users' identities and data was considered very important. However, stakeholders felt that a link should be provided to existing security or privacy principles where a nexus occurs rather than duplicating measures that are already incorporated elsewhere. The group acknowledged that the principles had considered the international policy landscape. Many noted the need to ensure a degree of harmonisation across jurisdictions. Given the international reach of most online technologies and services, this is not surprising.

eSafety's phased approach to this work was generally considered a great benefit. Indeed, there was an overwhelming willingness and interest in helping to develop the guidance and resources that will stem from these principles.

In revising the SbD Principles, eSafety took into account feedback from stakeholders. All stakeholders were aligned in their feedback and the revised SbD Principles reflect the findings from our consultations. Appendix B provides an overview of the main themes that were drawn out from the consultations, and indicates what action was taken as a result of the feedback.

V. Safety by Design Principles

While four over-arching principles were initially developed, two of the principles were combined following feedback received during consultations. As a result, three over-arching high-level principles were drafted to provide clear guidance to industry to help them develop services that harness the benefits and potential of the online world, while mitigating against real risks and harms.

The principles set out the measures that eSafety encourages and envisages online and digital services will take to ensure that the safety of users is suitably addressed and considered during the design, development and deployment of their service. While voluntary, adopting the principles will set industries apart in the marketplace, allowing companies to show global leadership in developing responsible, value and safety-centred online products and services.

Good practice notes have also been provided as part of the Principles to guide industry in understanding and implementing the principles and highlighting that the principles are achievable, realistic and actionable. These were highlighted during the consultation process as effective, although there is no robust evidence about the effectiveness and impact of these practices. As such, the good practice notes should be viewed as illustrations only, and act as a guide.

The principles that sit underneath each overarching high-level principle have been ranked according to the sense of urgency with which they need to be addressed — a point which was highlighted by all three stakeholder groups in consultation. While services are encouraged to address each principle, greater attention and focus should be placed on meeting the higher ranked principles.

Principle 1: Service provider responsibilities

This principle is based on the central tenet that while user safety is a shared responsibility, the burden of safety should never fall solely upon the end user. It acknowledges that, even though no platform or service can ensure the complete safety of individuals, every attempt should be made to ensure that known and anticipated harms have been adequately assessed in the design and provision of a service.

Online services can take preventative steps to guard against their service being used to facilitate, inflame or encourage illegal and inappropriate behaviours. To help ensure that known and anticipated harms have been evaluated in the design and provision of an online service, eight preventative steps have been identified under Principle 1.

Principle 1.1: Nominate individuals, or teams — and make them accountable — for user safety policy creation, evaluation, implementation, and operations

Under this sub-principle, nominated representatives need to be accountable for user safety on the service. While staff roles will vary considerably depending on the size and maturity of a service, accountability is critical to ensure that user safety is effectively addressed, incorporated and adhered to throughout the lifecycle of the service.

Managerial leadership will set the strategic direction of the service and determines the values and culture of the service. However, research has shown that clearly defining the roles and expectations of staff will not only ensure that the service goals are met but would also go some way to developing high-performing and committed teams[65]. Assigning one or more individuals to fulfil user safety functions as a core and measurable role objective clearly communicates a commitment to mitigating against risks and harms, enhancing the ability of the firm to act decisively and enforce its terms of service.

Good practice note (Principle 1.1)

All of the major online services have dedicated 'trust and safety' teams. Their 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. Most have a dedicated 'Head' of Safety and, in the case of Microsoft, they also have a dedicated 'Chief Online Safety Officer' position.

[65] Grattan, L & Erickson, TJ, 2007 'Eight ways to build collaborative teams', Harvard Business Review, November issue.

Principle 1.2: Develop community standards, terms of service, and moderation procedures that are fairly and consistently implemented

Human behaviour is complex and nuanced. Modern societies are governed by laws which delimit and determine patterns of acceptable behaviour. In an online context, user behaviour is shaped not only by the laws of the society in which a user lives, but also by the social and technological infrastructures, or institutional arrangements, that have been developed by vendors on the sites they use.[66]

Our youth consultation indicated that the current status quo of providing users with terms of use and community standards at sign up is not seen as effective. Often, users perceive these rules as being rarely or inconsistently applied. Young people indicated that the development of clear guidance about how people should behave on online services is the second most important safety by design principle. If services hope to contribute positively to the structure and stability of user experiences, then proactive steps are required to help create the online environment that users seek.

Services have developed moderation practices to manage behaviour on their sites. Some outsource this workload to specialised moderation services, others use a hybrid model of in-house and outsourced moderators, while others rely on volunteer moderators, upvoting systems or other community-driven approaches to manage the conduct and content on their sites. Our youth consultation exercise and industry stakeholders both indicated that the visible presence of a human moderator affects how people interact on the site, particularly if consequences for violations and breaches are addressed and actioned visibly and promptly.

Good practice note (Principle 1.2)

Some services are taking a proactive approach to ensuring that community standards and rules are clearly understood by their users. For example, French social media service Yubo[67] uses artificial intelligence and machine learning on their livestream function to flag to users when their behaviour is in breach of community standards. Warnings and an opportunity for users to rectify their actions are provided, with penalties and cautions resulting if no action is undertaken.

The youth-focused Australian mental health NGO ReachOut has trained peer moderators to model the community guidelines on its online community service. These moderators set an example of how ReachOut expects the community members to behave. When individuals do not adhere to the guidelines, the peer moderators remind them about expected behaviours. Over time, this had led the community to self-regulate, with young people actively pulling each other up on poor behaviour.

Gaming services, in particular, are beginning to utilise game reward systems, commendation systems and endorsements[68] in an attempt to motivate or change behaviours. Rewards, such as gaming skins or 'prestige' status are provided to users who model good behaviour.

Twitch[69], an online streaming service for gamers, allows streamers to directly appoint content moderators themselves. This provides individuals with an opportunity to shape the norms of their individual streams. Twitch provides streamers with moderation tools and launched a 'Creator Camp' for streamers and their moderators to empower and skill-up its creators to develop the type of culture and community that they want.

[66] Brey, P 2018, 'The strategic role of technology in a good society' Technology in Society 52, pp 39-45.

[67] Please see 'Hidden in plain sight: Safety innovation in live-streamed video' by Anne Collier, published on Medium, 17 October 2018.

[68] Examples include Blizzard's Overwatch endorsements and Riot Games' League of Legends honour system

[69] Please see 'Hidden in plain sight: Safety innovation in live-streamed video' by Anne Collier, published on Medium, 17 October 2018.

Principle 1.3: Put in place infrastructure that supports internal and external triaging, clear escalation paths and reporting on all user safety concerns, alongside readily accessible mechanisms for users to flag and report concerns and violations at the point that they occur

It is a general expectation that businesses provide a means by which consumers can complain or make a report about services and products. In an online context, users expect that they will be able to report issues and concerns about the service, including the behaviour of other users and uploaded content.

For services themselves, distinct and well-defined reporting pathways allow them to build efficient workflows so as to better manage cases and actions. Ensuring clear escalation enhances their ability to prioritise and take action against issues that are illegal, pose a threat to life or relate to users in need of special protections.

For young people, reporting and blocking are considered the most important safety features to address user safety and mitigate against harm. The sense of empowerment and control that these features provide was viewed as overwhelmingly positive during eSafety's consultation with children and young people.

If a service uses algorithms, filters or any other automated system, it is important to ensure that humans are kept 'in the loop'. This provides effective oversight of the governance of these tools. It also ensures that ethical, moral and social expectations are regularly evaluated which helps to develop systems that evolve with community expectations. Lastly, it is important that the use of automated systems is clearly communicated to users.

Good practice note (Principle 1.3)

Most major platforms provide reporting features that target content, conduct and contact. Reporting opportunities are available throughout the in-app experience, for example, at the point of contact — when an individual requests to interact/engage with the user, at the point of downloadt — when a user views content, or at a time convenient for the usert — via reporting forms. Users are able to report across a wide-range of features, from people, profiles, pages, avatars, images, videos, messages, groups, photos, adverts and events.

Users are often requested to categorise reports via drop-down menus, allowing services to triage reports to teams who are trained and skilled at dealing with specific types of abuse and complaints. Most escalate reports depending on the risk to life and illegality of content. Many have dedicated reporting channels for reports involving minors.

Some services, such as Twitter, have recently introduced measures to ensure that users are not confronted by the content that they have reported. Once content has been reported, it is placed behind an interstitial page which states that the user has reported this content.

YouTube has developed a process that provides individuals, government agencies, and non-governmental organisations effective at identifying content that violates community standards with expedited escalation pathways for review by their content moderators. This program is called the 'trusted flagger program'. To participate, individuals and organisations must attend training sessions.

Principle 1.4: Ensure there are clear internal protocols for engaging with law enforcement, support services and illegal content hotlines

In many countries, internet service providers and content hosts are required by statute to notify law enforcement when they become aware of child sexual abuse material (CSAM) on their networks. This generally also applies to content that promotes terrorism, violent terrorist content or abhorrent violent material. In other cases, ISPs and other internet companies have arrangements in place to report CSAM and online extremist material to dedicated hotlines, such as the hotlines that are members of INHOPE[70] national security hotlines[71], or national regulatory bodies or agencies.

Consultations with industry members, including those based in Australia, showed that awareness of the legal obligations relevant to reporting are low. Services need to have clear written policies and procedures stating the reporting requirements that apply in the relevant jurisdiction where CSAM and other prohibited content is concerned. In addition, these documents should establish specific reporting pathways, for example, URLs of web-facing reporting forms, or positions and contact numbers for specialist police personnel.

Finally, services must provide training in these obligations and standard operating procedures to relevant staff on an annual basis and incorporate training into staff induction arrangements.

For those impacted or effected by exposure to online risk and harm, signposting or providing users with links to sources of support or help in-country is important. During our consultation young people stipulated that communicating and raising awareness of support pathways in-app was important, particularly at point of need.

Good practice note (Principle 1.4)

Online service practices of handling reports to, and requests from, law enforcement differ by jurisdiction.

Examples of good practice include those jurisdictions that provide for mandatory reporting of child exploitation and sexual abuse material to law enforcement. In the United States, electronic communication service providers are required to report to the National Centre for Missing and Exploited Children when they obtain actual knowledge of 'facts and circumstances' relating to a variety of child exploitation matters. These include production of child exploitation material, selling or buying of children and distribution of CSAM.

A number of online services provide 'law enforcement only' portals that allow for police to make requests for data about users who violate the platform's terms of service to groom and exploit children. Others provide extensive guides for making such a request, including the requirements that must be met by both domestic and foreign law enforcement, before information will be disclosed.

Most of the major service have a safety centre/hub which houses information about how the platform keeps its users safe. This information usually covers what privacy, security and safety features are available to users, and on some platforms, links to resources and safety partners, including relevant helplines and support services. Some platforms ensure that users are notified of support services when making a report, or when search terms associated with harm (e.g. suicide or self-harm) are made on their platform.

[70] International Association of Internet Hotlines. Found at www.inhope.org.

[71] Hotlines and online reporting tools are available in many countries. In Australia, the Living Safe Together website contains information and a reporting form for individuals to report extremist material.

Principle 1.5: Put processes in place to detect, surface, flag and remove illegal and harmful conduct, contact and content with the aim of preventing harms before they occur

A number of technological tools have been developed to help services review, report and prioritise reporting of illegal content. Indeed, over the last six months proactive technologies have been launched that can detect child nudity and previously unknown child sexual exploitation content as it is being uploaded, with an application programming interface (API) to allow services to utilise these features on their systems[72].

Organisations including Thorn, The Technology Coalition and Two Hat Security have developed guidance to educate and inform services about proven practices, tools and resources that exist to identify, remove, report and prevent CSAM and abusive behaviour[73]. Once a service is made aware of illegal content, either via public reporting or active detection, it must take steps, or take reasonable and proportionate steps, to remove it.

In addition, chat analysis triage tools, behavioural signals, proactive detection and artificial intelligence chips are being used by services to flag risky conduct, content and contact to services, as well as being used to help surface risks to users so that they are able to manage their own safety.

Good practice note (Principle 1.5)

Koko is one example of an automated AI-powered conversational agent that is targeted toward mental health applications. It uses a hybrid human-machine moderation system and is able to identify users who are experiencing acute distress or suicidal thoughts. It either routes them to crisis lifelines or acts as an intermediary to connect those seeking help to a peer support network (based on a cognitive-therapy model). It is currently deployed across a range of messaging platforms, such as Kik, and is a referral partner for Tumblr and Pinterest.

Twitter uses behavioural signals to assess whether an account is adding to, or detracting from, the tenor of conversations. While disruptive tweets are not deleted, they are pushed further down in a list of search results or replies.

Chip manufacturers Nvidia and Intel have developed AI chips to filter live-streamed content to expedite the identification of bad agents, instances of self-harm and suicide and other acts of violence.

Instagram recently launched a bullying comment filter to proactively detect and hide bullying comments from its Feed, Explore, Profile, Photos and Live Videos features.

A number of innovations have been developed by industry to detect and expedite the removal of child sexual abuse material from their platforms. Examples include, but are not limited to, the development of PhotoDNA, webcrawlers such as Project Arachnid, and more recent technologies from both Google and Facebook that help organisations review, report and prioritise this material at speed. This includes proactive detection technologies that detect nudity and previously unknown child exploitation content when it is being uploaded.

[72] Please see Google's artificial intelligence tool, available via their Content Safety API, at www.blog.google/around-the-globe/google-europe/using-ai-help-organizations-detect-and-report-child-sexual-abuse-material-online/ and Facebook's software, found at https://newsroom.fb.com/news/2018/10/fighting-child-exploitation/.

[73] Please see: The Technology Coalition 2015, 'Employee resilience guidebook for handling child sexual abuse images', Thorn's 'Sound Practices Guide' and Two Hat Security's paper on 'How to Moderate Images Efficiently: Save Time, Money, and Resources with these Sample Workflows'.

Principle 1.6: Prepare documented risk management and impact assessments to assess and remediate any potential safety harms that could be enabled or facilitated by the product or service

Determining community standards (Principle 1.2) is vital for developing robust, efficient and evidenced workflows that will effectively moderate and remediate safety risks and harms on services. The risks and harms experienced by users are well catalogued, with three main themes of risk and five overarching categories of harm identified (see page 3). As such, services can use these to develop both the boundaries of acceptable behaviour and the level of risk acceptance for the service[74]. This process enables services to determine what functions it will offer users, what methods of moderation and escalation will be developed, and the restrictions that will be put in place to avert known risks.

Moderation workflows will develop and evolve over time, as the service gains popularity or is used in specific ways by the user community. The more data and information that the service has about how its service is being used and abused, the easier it is for the service to adapt workflows to prioritise and target risks. A host of outsourcing options are available for services that do not have the internal capacity to manage content moderation themselves.

Risk thresholds can be developed for both content and conduct, with users subject to different levels of oversight depending on their risk profile, or how new they are to the service. For example, new users, and users who are repeatedly disruptive or post content outside the acceptable thresholds for the service, can be subjected to higher levels of assessment from the service.

Services that target young users (under the age of 12) tend to be much more risk averse than those which target adult populations. As such, pre-moderation is often a prerequisite to uploading any content to services catering for children. Notwithstanding measures taken on age-specific sites, research shows that children can and frequently do access material on services that are age-inappropriate. This behoves all services to consider adopting measures equivalent to those employed by child-focused services.

As risk profiles can change over time, the service provider or technology vendor may need to perform risk audits to ensure that they are continuing to exercise good practice when it comes to addressing safety risks.

Good practice note (Principle 1.6)

LEGO has developed a risk assessment tool against which every new feature and architectural component is weighed. The legal and child safety teams provide developers with sets of measures and stop-gaps which the development team then assess and address before a product or feature can be brought to market.

Facebook has developed an internal functional expert review through which all products have to pass. These reviews involve experts in privacy, safety and security from both within the organisation and external to it[75].

Cybersecurity company Two Hat Security has produced briefing documents that outline workflows for moderating different types of content for online services.

Crisp, a language analysis firm, has developed AI to profile both users and online spaces, so that moderation attention can focus on areas of the service which attract those with nefarious intentions or where interactions are taking a negative turn. This profiling information allows services to assess existing features or to make the necessary changes to the online environment to curb this activity.

[74] See Two Hat Security paper on 'How to Moderate Images Efficiently: Save Time, Money, and Resources with these Sample Workflows'.

[75] Please see www.facebook.com/notes/facebook-australia-policy/how-can-facebook-do-more-to-reduce-bullying/2178124645848130/

Principle 1.7: 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

User safety is a shared responsibility, and our youth consultation highlighted the important role that users play in shaping a positive and empowering online environment. The safety of users is therefore not solely the responsibility of the service. Users do not relinquish their safety in order to access and utilise services. They also play a central role in developing a strong and safe community.

The development of social contracts[76], that go beyond the mere tick-box acceptance of a service's terms of use and privacy policies, could go some way to engendering a greater understanding, awareness and responsibility among users about the service's responsibilities and practices to keep users safe. It can also help users understand their own role in maintaining a safe community. The focus therefore shifts from services gaining consent, to the mutual role and reciprocal obligations of the service and user to uphold and develop community safety norms.

Good practice note (Principle 1.7)

The LEGO Group has recently introduced a 'Safety Pledge' that is introduced to users (the user base for LEGO Life is primarily children aged between 5–12 years) at registration. Users are taken through an induction process, where they learn the rules of the site via tutorials and top tips. Users have to sign the pledge before they are allowed to use the site after which they can earn badges as they progress. Indeed, the LEGO Group is introducing tighter parental consent flows to the registration and upgrading of accounts, whereby parents will need to provide Government ID or credit card details to allow their child to comment or post in-app.

[76] A 'social contract' is an implicit agreement among the members of a society to cooperate for social benefits. Most commonly, they define and delimit the rights and duties of each members of the society.

Principle 1.8: Balancing security by design, privacy by design and user safety when securing the ongoing confidentiality, integrity and availability of personal data and information

During the consultation phase, industry and other key stakeholders sounded a note of caution about duplicating existing provisions with privacy and principles. Principle 1.8 is premised on the simple tenet that one cannot separate the person from personal data. As such, there are serious safety considerations to take into account when handling this type of information. Data protection is thus not just concerned with the protection of data, but also the protection of the person behind the data.

Principles of safety, security and privacy should be complementary, not mutually exclusive, and must be balanced with one another by necessity. The consequences of imbalance have been keenly felt at the advent of new data protection and privacy legislation in the European Union which relates to industry's ability to use proactive detection of CSAM[77].

Identity theft and hacking were major concerns raised in our youth consultation exercise. Users having greater control over what they do on the and services they use was raised, as was the need for opacity. This is as much about users' identity information as it is about controlling their experiences.

Good practice note (Principle 1.8)

While data breaches are now common topics for media headlines, there are good examples of companies that have handled incidents responsibly and transparently. These firms acknowledge the breach as soon as possible, with clearly written breach notifications that specify the nature of the incident, what is known about its extent, and what steps are being taken to investigate. In addition, these notices provide a clear and reliable avenue for affected customers to learn more about the issue, including whether their personal records were affected. Good practice in this area then requires services to update users as and when more facts become available, with additional notices that act to ensure transparency and accountability.

Quizlet, a mobile and web-based study application, houses data protection, privacy and user safety programs under a unified Trust & Safety department. Product and business initiatives undergo consultative review with this department in relation to a unified set of Trust & Safety principles. This process allows Quizlet to maintain a consistent approach across product areas while also ensuring complementary policy considerations are incorporated. For example, in accordance with Quizlet's Trust & Safety principle to provide a safe and appropriate environment for learning for all users, Quizlet developed automated content moderation systems to detect profanity and other inappropriate language on the platform. These systems were calibrated to reflect varied sensitivity across different products, applying a more stringent filter for Quizlet Live given the increased risk and possibility for harm inherent in a live, interactive service.

[77] Child Dignity Alliance 2018, 'Child Dignity Alliance Technical Working Group report'.

Principle 2: User empowerment and autonomy

This principle is based on two key tenets. First, that the dignity of users is of central importance when designing safe online services, and second, that it is important for services to treat the interests of users as an inalienable consideration.

Human agency and autonomy can be supported, amplified and strengthened in the design of online environments. An inclusive design approach that secures user safety as part of the in-service experience will enable, optimise and support users to have greater control over their interactions. Features and functionality that allow users to fully govern and regulate their own experiences — particularly at times when their safety is being, or is at risk of being, compromised is recommended.

Principle 2.1: Provide 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

In a shared responsibility framework, for users to have control over their own safety they need to be provided with the right tools — and feel empowered to use these tools. In a rights-based framework such as the SbD Framework, individuals should be provided with opportunities to revise their preferences and choices as they see fit. They should also have control over their personal information and how it is portrayed.

Our consultation process with young people, parents and carers highlighted a consumer expectation that privacy and safety settings should be placed at the highest levels by default — that is, at sign-up or registration. This is particularly important when considering that the vast majority of users keep privacy settings at the level they were configured at on registration[78].

For young people, in particular, having greater control and autonomy was of the utmost importance. There was a strong belief that features which offer control, and empower users to directly tackle violations of conduct, foster greater confidence and a sense that the user was safe online. These features included blocking, muting and reporting functions.

Individuals should not have to accept that in order to use a service, they may have to compromise elements of their privacy or safety. Services can develop features and functionality which hand freedom of choice and control back to the individual.

Good practice note (Principle 2.1)

Tumblr has a safe mode filter which is turned on by default for all users. This ensures that sensitive content does not appear on a user's dashboard or in their search results. Users are encouraged to flag content as 'sensitive' or 'explicit' at point of upload. Since December 2018, users have been barred from posting sexually explicit material on the platform.

Most of the major platforms allow users to control their privacy and safety settings, and for some, default settings have been placed on the highest level by default for accounts held by minors. Prompts allow users to review their settings regularly as part of their in-app experience. Pop-ups can also be used to alert users to the possible safety consequences when settings are changed. For example, when a user changes their profile setting to 'public', users are reminded that this means that anyone can view their post.

Google has placed privacy and security setting information and controls in one location. The company reminds users to carry out a 'Privacy Checkup' at intervals decided by the individual themselves.

Facebook has launched features that allow users to hide or delete comments, as well as making it easier for users to search for and block content from their feed or comments that they do not want to be exposed to. In addition, features allow bystanders to report individuals who are being targeted (bullied or harassed). Importantly, these reports are kept anonymous.

Snap has placed 'Ghost mode' on as default for all users in Snap Map.

[78] Spool, J 2011 'Do Users Change Their Settings?' User Interface Engineering, 14 September 2011.

Principle 2.2: Establish clear protocols and consequences for service violations that serve as meaningful deterrents and reflect the values and expectations of the user-base

While community guidelines and terms of use outline the expected and accepted standard of behaviours for the service, genuine consequences and repercussions for those who violate or breach rules can deter toxic behaviour.

Our youth consultation participants wanted to see greater punitive measures for those who misuse platforms to cyberbully. Specifically, they stated that suspending offending accounts and blocking IP addresses were effective interventions that should be more widely used.

Good practice note (Principle 2.2)

Twitter has a number of measures in place for users who violate its standards. These include, requiring users to delete prohibited content before they can continue to engage with the service, temporarily limiting a user's ability to create posts or interact with other Twitter users, asking users to verify account ownership with a phone number or email address and permanently suspending user accounts that violate terms and conditions of the service.

Google have recently updated their warning systems to users who violate their community standards. YouTube has always applied a three-strike and email notification system for violations. In order to better educate and raise awareness of their community guidelines, YouTube are providing first time violators with more time and information to better understand why their posts violate the standards — before their strike system is administered. In addition, they are applying more consistent penalties for violations and have expanded the policy resources available in the help centre to provide more details on what behaviours will result in strikes.

Yubo cautions users who are live streaming when their behaviour is in violation of its policies in real-time. If users do not stop the activity, their live-stream will be stopped and no longer visible to other users until they have taken action to rectify their behaviour.

Microsoft recently made its violations of services more prominent, with users breaching standards being suspended or banned from Xbox services. These include forfeiture of content licenses, gold membership time and Microsoft account balances associated with the offending account.

Games developer Ubisoft launched an automated temporary banning system for Rainbow Six Siege players that use offensive language or behaviour, threats and harassment in the in-game chat. First offences lead to a thirty-minute temporary ban, second and third offences have a two-hour ban, and further offences lead to an official investigation that may result in a permanent ban.

Principle 2.3: Leverage 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

Proactive measures can be taken by services to forewarn or surface risk indicators to users so that they are able to take action against, or mitigate, any potential risks or harms. A proactive approach enables users to act before an incident takes place, rather than waiting until the damage has occurred. Swift action against imminent violations creates opportunities for reflection and evaluation. This can empower users to take steps to safeguard not only themselves, but also to enhance the resilience of the network or system as a whole.

Safety standards and their subsequent protections are essential elements of a value-sensitive design. While users should have the freedom to express themselves and participate fully in the opportunities that the online world offers, they should be required to adhere to the prosocial norms promoted within the environment.

Multiple technical tools and features have been developed that can filter content — not necessarily for removal (except when the content is illegal), but to place it behind age-gates or interstitial warning pages so that users can access the content by exercising informed consent.

Advances in technology have meant that services can use behavioural and content signals to identify risky users, behaviour and content, even at the point of upload or contact. Measures can then be taken to immediately reduce risk of harm to users, as well as providing users with the necessary tools and information so they are able to navigate the service safely. Services are therefore able to prompt or sign-post users towards advice, support and guidance at point of need.

Good practice note (Principle 2.3)

PA Consulting in the UK has developed online risk alert principles[79] for creating safety alerts that educate people on how to spot online dangers. Alerts to users are instant, specific, relevant, private and supportive. They have developed a proof of concept that presents risk-based indicator alerts to users, empowering them to reflect and consider their situation and take appropriate precautionary steps.

Many platforms use age-gates and interstitial warnings on videos, images and messages that have been flagged to the service as being sensitive or age-inappropriate.

Twitter uses behavioural signals to identify users who target others with abuse or harassment. They then limit the visibility of their tweets, as well as allowing users to mute specific individuals themselves. Tweets will only be removed once they have been reviewed as violating Twitter's Rules.

Yubo filters images and live-streaming in order to detect nudity and users posing in their underwear, prompting users to reconsider posting images, or to be fully dressed, before allowing live streams or posts to be uploaded.

Roblox provides users with different safety settings and experiences dependent on the age of the account holder. For those aged 12 and under, posts and chats are filtered for both inappropriate content and behaviour, and to prevent personal information from being posted in both private and public chat.

PopJam uses facial recognition technology and optical character recognition on images uploaded onto its platform. These tools ensure that text embedded onto images does not violate community standards and that users do not post real-life images of their faces onto the platform.

[79] Please see www.paconsulting.com/insights/protecting-children-from-online-sexual-exploitation-and-abuse/.

Principle 2.4: Provide 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

A holistic response to safety concerns is needed: one that not only deals with the issue or problem at hand, but also guides the user to additional sources of support or help. This approach recognises that user wellbeing can be just as important as user safety. It also respects the particular vulnerabilities that might be experienced by users making reports or complaints, especially when the reporting leaves them feeling exposed or visible.

Good practice note (Principle 2.4)

For users to engage with safety mechanisms, they need to trust that the platform will take their concerns and reports seriously. Updating and informing users about their reports will go some way to engendering greater trust in the platform, which may ultimately lead to greater use of existing safety mechanisms.

Facebook has developed proactive detection AI to detect posts and their surrounding context for patterns of suicidal thoughts. Users are immediately shown support options, including resources for help and ways to connect with loved ones. When a Facebook Community Operations team member determines that there may be imminent danger of self-harm, local authorities may be contacted. In addition, Facebook provides users with a list of helplines globally that they can contact for support.

Instagram has taken a holistic approach to supporting users who are looking at or posting content about self-harm and suicide. The platform has a Privacy and Safety section focusing on this topic and they have tightened filters in an attempt to curb content that encourages eating disorders and self-harm. Once this type of content is accessed, users are given the option to access tips and support, talk to a friend or to reach out to support groups.

YouTube has developed a reporting history dashboard which indicates to users whether their reports are active, removed or restricted.

Appeals processes have recently been incorporated into the reporting workflows of Facebook and YouTube. While the processes were initially developed to appeal decisions about content that the service had taken down, Facebook is seeking to expand it further to allow users an opportunity to appeal any decisions made to a report that has been filed. They also plan to appoint an independent body to deal with appeals on content decisions.

Principle 2.5: Evaluate 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

An inclusive design approach to safety features and mechanisms would ensure that the diverse needs and requirements of the general population are met. An evaluation of the design features that takes into account the capabilities of the user population, will ensure that anyone, in any situation, can harness the benefits of the safety considerations that have been incorporated into the service.

Safety features should be enabling for all users, ensuring that online environments empower and provide autonomy to all users, particularly at times when their safety is being, or is at risk of being, compromised.

Good practice note (Principle 2.5)

Microsoft has recently published an inclusive design toolkit[80] that outlines three broad principles:

1. recognise exclusion
2. learn from diversity
3. solve for one, extend to many.

Resources include a range of activities, as well as multiple short films to illustrate inclusive design in practice.

[80] Found at www.download.microsoft.com/download/b/0/d/b0d4bf87-09ce-4417-8f28-d60703d672ed/inclusive_toolkit_manual_final.pdf.

Principle 3: Transparency and accountability

Transparency and accountability are hallmarks of a robust approach to safety. They not only provide assurances that 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.

To enhance users' trust, services can foster awareness and understanding of safety, and its role on their services. Principle 3 sets out some of the steps that can be taken by services to continue developing this trust with their users.

Principle 3.1: Embed user safety considerations, training and practices into the roles, functions and working practices of all individuals who work with, for, or on behalf of the product or service

The first principle in the National Principles for Child Safe Organisations stipulates that that 'child safety is embedded in institutional leadership, governance and culture[81].' This measure extends the ethos of this standard to user safety in online contexts.

Adopting the SbD framework will enable the service to effectively address user safety considerations by acknowledging the importance of policies, procedures and protocols. An awareness of SbD Principles will ensure user safety is embedded as a core business objective and will help to guide working practices throughout the development lifecycle of the service.

Given the extensive development and maintenance roles played by temporary staff and contractors within large technology firms, sub-principle 3.1 should extend to the entire workforce.

Good practice note (Principle 3.1)

The LEGO Group has created a three-pronged approach, where 1) all relevant LEGO employees and affiliates take a mandatory introductory digital child safety eLearning course, 2) targeted in-depth training is delivered either face-to-face or via remote participation, and 3) annual assessments of compliance with the Lego Group's internal Digital Child Safety policy is conducted and results communicated to both leadership and relevant divisions.

Principle 3.2: 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

Raising awareness and educating users, and parents or carers, about the safety features, processes and protocols that exist on services can help inform users about how they can best navigate the service.

During the course of consultations carried out by eSafety, a single, well-publicised repository of safety information was felt to be particularly important for those users who could not, or would not, seek advice or support from those around them. This was especially true for young people.

To assist with user engagement, this information needs to be easy to find, written in plain language that is comprehensible to younger users and ideally designed using short-form notices as standard.

Addressing user safety concerns is an iterative process, and updates to safety policies and processes occur frequently. It is important that users are proactively informed about any changes to allow individuals to review their safety and privacy posture, especially when policy updates relate to secondary uses of data.

[81] National Principles for Child Safe Organisations, see https://childsafe.humanrights.gov.au/national-principles.

Principle 3.2 continued...

Good practice note (Principle 3.2)

All major services have specific sections on their websites that provide users with an overview of their safety features. These include, but are not limited to, tips on how to stay safe, abuse policies, specific parental resources and how-to guides for reporting and blocking.

Twitter has recently started an experiment publicising its rules to users in order to test whether this improves civility on the platform. This experiment stems from research carried out on Reddit which highlighted that making policies visible to users can improve online behaviour[82].

The Lego Group has developed a special Safety character who is a users' first friend on registration. This character communicates all safety related messages to the child, and so is the one to alert, help and support a child with safety issues throughout their in-app experience. In addition, they have developed a safety group inside the app who will be the repository for all guidance that Lego produces. All users are connected to this safety group and receive notifications when new material is uploaded and communicated. The safety character will pop up in contextually relevant places, such as when a child uploads a photo, to remind and offer guidance about safe practices in a visible and easily accessible manner.

Principle 3.3: 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

Risks and harms are inter-related, inter-connected and can manifest in different and sometimes unexpected ways. Services cannot expect to have the expertise or knowledge about how to manage and mitigate against these harms on their own. The global online safety community has built up expertise and skills that can be accessed and leveraged by industry to their advantage. Services can easily seek affiliation with renowned online safety councils or bodies to achieve this aim.

Despite technical features and tools being built with the best of intentions, unintended consequences and negative impacts can occur. Incorporating open channels of communication and feedback with independent experts and the general public can ensure that potential consequences and ramifications are identified, rectified and informed by best practice before the feature launches.

Good practice note (Principle 3.3)

Roblox has a Trust and Safety Advisory Board comprised of four independent members of the digital safety community and have recently launched a Digital Civility initiative. They have joined the Fair Play Alliance, a cross-industry initiative that seeks to share research and best practice to foster fairer and safer play in online games.

Some platforms, such as Google, convene groups of independent experts to provide advice on contentious issues, or to look at how to navigate pieces of legislation in order to balance the rule of law against individual rights. For example, the company assembled an Advisory Council on the 'Right to be Forgotten' comprising eight independent experts. These experts volunteered their time and did not have to sign non-disclosure agreements.

[82] Matias, JN, 2017, PhD Thesis 'Governing human and machine behaviour in an experimenting society'.

Principle 3.4: Publish 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

Offering users and potential users the opportunity to evaluate the services' approach to, and success in, tackling and addressing safety risks and harms is becoming more commonplace. This level of transparency provides the general public with a greater understanding of the steps that services take to address risk and harms on their platform. It also provides assurances, and a degree of accountability to the public, that the safety of users is taken seriously.

While statistics that highlight the scale of reporting are informative, the development of a thorough analysis of metrics on the efficacy of guidelines and rules in curbing harmful behaviour is vital. Attempts are being taken globally to achieve meaningful transparency and accountability, and eSafety will seek to collaborate with stakeholders further in developing these metrics as part of the continuing work of the eSafety SbD Framework program.

Good practice note (Principle 3.4)

Facebook, Google and Twitter have all published transparency reports in the last 12 months in a commitment to a more open exchange of information.

Facebook released its first Community Standards Enforcement report in May 2018, and which was updated in November 2018 to cover eight categories of violations. Facebook have identified five measurement metrics that will be used in future reports, with two metrics still being under development. The five community standards enforcement measures are: prevalence of community standards violations, quantity of content actioned, quantity of content actioned by detection technology and trained teams before items are flagged by users, speed of action based on impact of violating content (under development), and accuracy of actions based on reviews (under development).

Twitter released its 13th biannual transparency report late in 2018, which included a new section 'Twitter Rules enforcement' that provides data and insights into the enforcement of abuse, hateful conduct, private information, child sexual exploitation, sensitive media and violent threats. Twitter have committed to evolving their approach to transparency to ensure it is comprehensive, clear, contextual and meaningful.

The Internet Commission[83], based in the UK, is currently in the process of developing an independent transparency reporting framework, as part of its dialogue on digital responsibility. The transparency framework identifies key indicators to better uncover the impact and effectiveness of content moderation processes and practices within services.

New America recently published 'The Transparency Reporting Toolkit: Content Takedown Reporting[84], which identifies best practice for content reporting takedowns across six categories — including Community Guidelines-based content takedowns.

Principle 3.5: Commit to innovating and investing in safety-enhancing technology, as well as collaborating and sharing safety-enhancing tools, best practices and technology

Advancements in technology, machine-learning and artificial intelligence have the potential to radically transform user safety and online experiences. In order to harness and promote innovation and investment in user safety, a commitment among and within services to collaborating and pushing the boundaries on addressing and combatting online harms should expedite the development of a new wave of safety-enhancing technology.

Good practice note (Principle 3.5)

A number of main industry players are part of global consortiums and initiatives to better protect children online, such as the Alliance to Better Protect Minors Online, WePROTECT Global Alliance, Child Dignity Alliance, and the ICT Coalition for Children Online. Representation on these bodies highlights a commitment from industry to collaborate with others in protecting the most vulnerable groups online.

[83] Please see The Internet Commission website at inetco.org for more information.

[84] Singh, S & Bankston, K 2018, 'The Transparency Reporting Toolkit: Content Takedown Reporting' Open Technology Institute.

Summary

Safety by Design (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.

At its core, SbD is about embedding the rights of users and user safety into the design and functionality of products and services. It places safety as the third pillar in the developmental process, sitting alongside the well-established processes of privacy and security. Safety is a product design imperative.

eSafety's consultation exercise highlighted that SbD is something that Australian citizens — young people, in particular — want and expect. Young people felt that user safety was a shared responsibility, and while the onus is on industry to protect its users, the end-users also have a significant role to play. Most importantly however, young people want greater control over, and more transparency from, the platforms and services they use. Young people stated that they wanted to be made more aware of safety features that exist or are being developed. They also wanted to manage their own safety more confidently. Lastly, they wanted to place their trust in industry.

Our industry and key stakeholder consultation revealed that SbD is not common practice among industry partners. Innovations in user safety are beginning to be developed at pace by some industry players, and there was an overwhelmingly positive affirmation of the importance of user safety among those consulted. All industry players recognised the need to ensure that user safety is addressed adequately in the online world, and that online services have an important role to play in this space.

eSafety believes that if online services can start to adopt key elements of the SbD Framework, they will be taking affirmative steps to make user safety considerations a routine element of their product development cycles. We hope that the SbD Principles will act as a catalyst for further innovation in user safety, while also embedding user empowerment and autonomy as a core business objective for those developing products, platforms or services online.

Though committing to the SbD Framework is voluntary, eSafety firmly believes that adopting these Principles will give online services an edge in the marketplace. The SbD Framework offers a chance for the thriving technology sector in Australia to show global leadership in developing online products and services that are truly centred around users.

Appendix A

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 are priorities 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.

Appendix B

Consultation themes from industry

Overarching feedback on Safety by Design Overview

CONCERN	ACTION TAKEN
Overall	
Lacks framing/context.	The Overview was written to provide the context, theoretical underpinnings and process undertaken by eSafety in developing the SbD Principles. A framing piece will be developed for the SbD principles.
Too focused on children.	An inclusive design approach has been clarified.
Not enough focus on children.	An inclusive design approach has been clarified and children's rights detailed in the Overview.
Too broad (one-size-fits all).	Clarification provided as to who the SbD principles are targeting.
Too prescriptive.	A principle-based approach was reinvigorated into measures, and prescriptive details removed.
Lack of focus on role of parents.	eSafety acknowledges that parents play an important role, but the focus for the principles is user and human-centred. As such, they focus on the user as opposed to carers or guardians for those who have special protections.
Focus too much on industry and not shared responsibility.	The Overview outlines the rationale and justification for focusing on industry as the providers of the platform services.
Terminology and language is complex.	An attempt has been made to simplify language and terminology.
Lack of clarity over who principles are targeting.	Provided in the Overview.
Scope of organisations targeted too broad.	Clarification provided in the Overview.
Overwhelming for SMEs and start-ups.	Clarification that the SbD Principles are a benchmark and template, with measures ranked according to the sense of urgency with which they need to be addressed.

CONCERN	ACTION TAKEN
Overall (continue)	
No information on how platforms can be misused, or what the risks and harms are.	Provided in the Overview. In addition, eSafety has noted the request for explicit guidance on this that could be actioned in the next stage of the work program.
Australia focused, so impact of principles will be limited due to the inability of local industry to influence how overseas-developed products and platforms are designed.	Clarified that Principles are focused on the thriving technology sector in Australia, and that Australia has the opportunity to be global leaders in developing user safety-centred products and services. Reference made in paper to Principles being shared with international partners.
Plaform reponsibility	
a) Comprehensive and documented risk management process and impact assessment	
What expectation is there for industry — Query what expectation?	Carrying out risk and impact assessments is considered key and is, in fact, standard practice in technology development and project management cycles. As such, the expectation is that best practice would dictate that risk and impact assessments should be carried out across every product or service.
Onerous on companies.	eSafety acknowledged that risk and impact assessments carry resource implications. However, the impact — in terms of efficiency and cost effectiveness of taking preventative measures that are clearly outlined and evidenced — was felt to outweigh retrofitted solutions after a major breach or public harm has occurred.
Information needed on what this should/could like.	eSafety has noted the request for explicit guidance on this and believes that this could be actioned in the next stage of the work program.
Documented assessments going too far — and overburdening on SMEs and start-ups.	Documented assessments are important for transparency purposes, but also in protecting companies from liability. eSafety felt that these should be standard practice for mature companies, but acknowledgement has been made that a less rigorous process should be considered for SMEs. This is an area that eSafety would like to consult further on in the next stage of the work program.
b) Designated individuals or teams accountable for safety policy creation, implementation, and operations	
'Designated' is too prescriptive.	This has been altered to 'nominated'. Please see revised SbD Principle 1.1.
Onerous for SMEs and start-ups.	If there is not someone tasked with, and whose performance is measured on safety, there will not be designated accountability within the organisation. For SME's, safety training and awareness could be the benchmarks — supplemented by good practice guidelines. These guidelines could be actioned in the next stage of the work program.

CONCERN	ACTION TAKEN
c) Reactive e.g. blocking and muting functions, active e.g. content filtering, detection algorithms, and proactive e.g. surfacing and flagging harmful behaviour, chat analysis, safety measures that are user-centred and embedded within the platform in a way that is relevant to the product or service experience itself	
Unclear on difference between active and proactive.	eSafety acknowledged that this was a very nuanced distinction relating to acting in time versus acting in advance. These terms have been removed from the Principles.
Not realistic for start-ups.	Our consultation process evidenced that smaller companies and players actively use a range of safety measures, either outsourcing, or in-house. Machine learning, AI, monitoring, moderation and other services are becoming more available all the time.
Legal liability issues for proactive measures.	eSafety is aware of EU legislation (European Privacy Regulations and the General Data Protection Regulations) where this issue has been raised. In relation to online harms, there should be no room for doubt or ambiguity on whether industry is able to use technical tools in a proportionate manner to proactively eradicate the internet of illegal content.
Too prescriptive.	The Principles have been adjusted to be less prescriptive, and do not prescribe specific technologies or actions to achieve stated goals. Please see revised SbD Principle 1.3, 1.5, 2.1 and 2.3.
Provision of examples will become outdated too quickly.	The provision of good practice notes was considered important to illustrate that the Principles are achievable and realistic. Specific reference to technological tools within the Principles themselves have been removed. In addition, the paper clearly states that the Framework needs to be updated, and should be seen as a living-document, which is one of the main reasons that the SbD project was designed as a tiered, continuing work program.
d) Clear escalation and reporting routes with designated single points of contacts for law enforcement, relevant hotlines and regulatory agencies	
Too prescriptive.	Amendments have been made to separate out the different processes raised in this principle, which reflected the concerns from stakeholders on this point. Internal escalation processes with law enforcement is now separate from triaging and escalation paths for safety concerns.
Confusion over what is meant by SPOC.	The wording has been altered to clarify that protocols should be in place to engage with law enforcement, rather than specifying a SPOC.
Definition of hotline needed.	Wording has been altered to ‘illegal content hotlines’, and further guidance will be provided in the next stage of the work program.

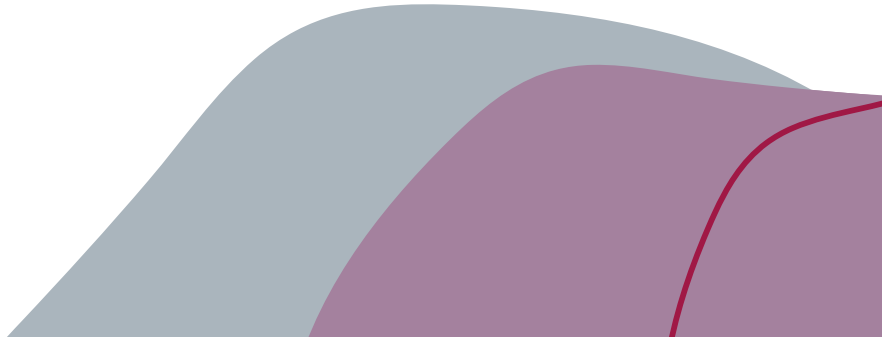
CONCERN	ACTION TAKEN
e) Clear and distinct reporting channels, escalation paths and impact assessments for various report types to manage an array of safety concerns and violations	
Too prescriptive.	Amendments have been made to the Principles, as outlined above. Please see revised SbD Principle 1.3, 1.4 and 1.6.
Simple and transparent rather than 'distinct'.	Wording has been amended to 'put in place infrastructure'. Please see revised SbD Principle 1.3.
Too onerous for SMEs and start-ups.	During the consultation process, a couple of stakeholders highlighted that reporting designations allow services to build flows to manage their load, and, as such, are beneficial to services themselves. Wording has been altered to be more principles based, rather than prescriptive.
Staff training essential.	This point has been included in the revised SbD Principle 3.1.
Visible rather than 'clear'?	The wording for this Principle has now changed, see revised SbD Principle 1.3.
f) A clear outline of a platform's approach to online safety and content moderation made available at account sign-up and at periodic intervals, which also include clear expectations that end users utilise the platform responsibly and lawfully	
Overlap with transparency section?	This Principle referred to the need to develop a social contract with users as well as providing users with a clear understanding of the ground rules for the site. These points have been separated. Please see revised SbD Principle 1.7 and 3.2.
g) Readily discoverable tools, advice, resources and guidance on user safety and digital wellbeing, that are transparent, visible, easy to use and intuitive	
Not enough focus on children's needs.	The Principles have been developed to ensure that the specific protections for children and young people are explicitly met. Care has been taken to ensure that each of the SbD Principles balances and child's right to provision, participation and protection.
How are 'easy to use' and 'intuitive' determined?	Terminology has been altered to 'readily accessible'.
h) Demonstrated commitment to collaborate, innovate and invest in safety-enhancing technologies on an ongoing basis	
Difficult to measure 'commitment'.	This Principle has moved to SbD Principle 3: transparency and accountability. While difficult to measure, publicly committing to innovate, invest and collaborate is considered a step in the right direction. Please see SbD Principle 3.5.

CONCERN	ACTION TAKEN
Recognition and respect for user identity	
i) Age-appropriate design (included in lead-up to measures)	
Not just age — all vulnerabilities and users as a whole.	The paper clearly indicates that the Principles were developed with those who require special protections in mind, and are underpinned by human rights, digital ethics and an inclusive design approach.
Specific reference to age-appropriate design has been removed, but attention has been given to those with distinct characteristics and capabilities. Please see revised SbD Principle 2.5.	eSafety acknowledged that risk and impact assessments carry resource implications. However, the impact — in terms of efficiency and cost effectiveness of taking preventative measures that are clearly outlined and evidenced — was felt to outweigh retrofitted solutions after a major breach or public harm has occurred.
i) Age-appropriate design (included in lead-up to measures)	
Provision of examples will become outdated too quickly.	The provision of good practice notes was considered important to illustrate that the Principles are achievable and realistic. Specific reference to technological tools within the Principles themselves have been removed. In addition, the paper clearly states that the Framework needs to be updated, and should be seen as a living-document, which is one of the main reasons that the SbD project was designed as a tiered, continuing work program.
a) A robust user-centric evaluation of all design and functionality features to be carried out before products or features are released to the public	
What criteria would fall under this measure?	These criteria could be actioned in the next stage of the work program.
b) Technical security measures, disclosures and data retention procedures made transparent, intuitive and available to all users	
Is/should data retention be within scope of these principles?	Reference to data retention have been removed.
Dangers of making some of these measures public.	This Principle has been subsumed within the revised SbD Principle 3.2 and 3.4.
c) Mechanisms to protect user identity and personal information made available, adaptive and open to recourse	
Overlaps with privacy-by-design?	This Principle has been removed in acknowledgment of the overlap with privacy-by-design. However, reference to the need to consider security, privacy and user safety consideration is now contained within revised SbD Principle 1.8.
What is meant by 'open to recourse'.	As above.

CONCERN	ACTION TAKEN
c) Mechanisms to protect user identity and personal information made available, adaptive and open to recourse (continue)	
Unclear of what is expected/ meaning of this measure.	As above.
Is this a threat to user safety?	As above.
Mechanisms will be dependent on function of service — flexibility needed.	As above.
User empowerment	
General	
Similar concepts to recognition and respect.	Principles have been reduced from 4 to 3 to reflect the overlap that existed.
Needs to take into account different cultures and models of business.	A principles-based approach has been taken throughout to avoid prescription.
Term empowerment has connotations for some, also include 'autonomy'.	Both have been included given that there were mixed perceptions to these terms amongst those consulted.
a) On-by-default to the highest privacy and safety levels	
Too high a bar.	Given the resounding expectations from Australian citizens that this be carried out, this is felt to be an important criterion.
Only relevant for personally identifiable information and other sensitive user data and information.	As above.
b) Allow users to control who has access to their posts, location, profile and other personal information	
May need contextualisation for those platforms based on collaboration.	The wording has been changed, please see revised SbD Principle 2.1.
c) 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	
For legal reasons, short-form is not always appropriate.	While this may be the case, this does not absolve services of communicating the essence of these documents in a manner that is easily understood by their users. Please see revised SbD Principle 3.2.
Also need to link to other support and information services.	This point has been inserted in revised SbD Principle 2.4.

CONCERN	ACTION TAKEN
d) Provide clear consequences for violations that serve as a meaningful deterrence for users	
None.	
e) Optimise safe interactions through features and functionality that target, signpost, support, prompt and support user empowerment as part of the in-service experience	
None.	
f) Create holistic processes, policies and procedures underpinned by built-in support functions and feedback loops for users that keep users informed on the status of their reports to the service provider	
None.	
g) Provide users with the ability to access, obtain and reuse their personal data for their own purposes across different services, and to request its erasure or modification	
Reflects GDPR is this the intention?	This Principle has been removed, but the importance of balancing security, privacy and safety considerations is still present in revised SbD Principle 1.8.
Overlap with other principles and regulatory bodies.	As above.
Transparency and accountability	
Overall	
Too onerous for SMEs and start-ups / overburden.	The need for further guidance on what level of transparency is required for mature services versus start-ups has been acknowledged. A tiered or graded approach to transparency reporting will be considered in the next stage of the work program.
Focus should be on empowering and educating users.	This is acknowledged in SbD Principles 3.2 and 3.4.
Needs to include transparency of all policies, processes and protocols.	This has been included in SbD Principle 3.2.
Does not mention expectation on training and wellbeing of moderators.	Moderators have not been pulled out specifically in the Principles, however training of staff has been included in SbD Principle 3.1.
Principles do not really address accountability.	SbD Principle 3.1 and 3.4 have tried to address this.
Publicly available or to a regulatory / independent authority?	SbD Principle 3.4 stipulates that an annual assessment should be published publicly.

CONCERN	ACTION TAKEN
a) Polices and standards enforced through the open publication of meaningful statistics on metrics such as abuse data, the effectiveness of moderation efforts and the manner and impact with which community standards and terms of services are being met	
What is the purpose of providing stats?	Statistics offer users the opportunity to evaluate the services' approach to, and success in, tackling and addressing safety risks and harms. Indeed, the publication of transparency is becoming more common. While the provision of statistics that highlight the scale of reporting is informative, the development of a thorough analysis of metrics on the efficacy of guidelines and rules in curbing harmful behaviour is vital.
Focus should be on 'meaningful'.	Agreed, which is why an emphasis has always been placed on this term.
Metrics would be challenging to transform into meaningful comparisons across platforms.	Agreed, and which is why SbD Principle 3.4 stresses the publication of meaningful metrics — rather than simply statistics. eSafety will seek to collaborate with stakeholders further in developing such metrics as part of the continuing work of the eSafety SbD Framework program.
Too onerous for SMEs and start-ups.	Principles have been reduced from 4 to 3 to reflect the overlap that existed.
Not appropriate for some information to be in public domain.	The publication of transparency reports by major players highlights that the publication of data is perhaps less risky than was initially perceived.
b) A 'safety baseline' created from previously published statistics, followed by periodic updates performed to track progress on this data	
As above.	Reference to a base-line has been removed but should be considered in discussions relating to transparency reporting moving forward.
c) Clear information that demonstrates any steps taken to improve user safety made publicly available	
None.	
Independence is key, and external audits should be included.	The term 'independent' has been added to SbD Principle 3.3.
d) Open engagement with a wide user-base including experts and key stakeholders on the interpretation and application of safety standards and their effectiveness or appropriateness	
Independence is key, and external audits should be included.	The term 'independent' has been added to SbD Principle 3.3.



esafety.gov.au