Australian Government
Department of Communications and the Arts

Online Safety Charter—consultation paper

February 2019

Disclaimer

The material in this paper is of a general nature and should not be regarded as legal advice or relied on for assistance in any particular circumstance or emergency situation. In any important matter, you should seek appropriate independent professional advice in relation to your own circumstances. The Commonwealth accepts no responsibility or liability for any damage, loss or expense incurred as a result of the reliance on information contained in this paper.

This paper has been prepared for consultation purposes only and does not indicate the Commonwealth's commitment to a particular course of action. Additionally, any third party views or recommendations included in this paper do not reflect the views of the Commonwealth, or indicate its commitment to a particular course of action.

Copyright

The material in this paper is licensed under a Creative Commons Attribution—4.0 International licence, with the exception of:

- the Commonwealth Coat of Arms
- this Department's logo
- any third party material
- any material protected by a trademark, and
- any images and/or photographs.

More information on this CC BY licence is set out as follows:

- Creative Commons website—www.creativecommons.org
- Attribution 4.0 international (CC by 4.0)—www.creativecommons.org/licenses/by/4.0.

Enquiries about this licence and any use of this discussion paper can be sent to: copyright@communications.gov.au.

Third party copyright

The Department has made all reasonable efforts to clearly identify material where the copyright is owned by a third party. Permission may need to be obtained from third parties to re-use their material.

Attribution

The CC BY licence is a standard form licence agreement that allows you to copy and redistribute the material in any medium or format, as well as remix, transform, and build upon the material, on the condition that you provide a link to the licence, you indicate if changes were made, and you attribute the material as follows:

> Licensed from the Commonwealth of Australia under a Creative Commons Attribution 4.0 International licence.

Enquiries about the use of any material in this publication can be sent to: copyright@communications.gov.au.

Using the Commonwealth Coat of Arms

Guidelines for using the Commonwealth Coat of Arms are available from the Department of Prime Minister and Cabinet website at www.pmc.gov.au/government/its-honour.

Online Safety Charter—consultation paper

Introduction

Online safety[1] is a shared responsibility, and the work of improving online safety outcomes for our community doesn't and shouldn't rest with either industry or end-users alone. Improving online safety requires genuine effort and commitment from all parties, including Government. Parents and those with caring and teaching responsibilities also have important roles to play in equipping children and other vulnerable members of the community with the tools they need to deal with the potential dangers they might face online.

However, industry does have a unique and important role to play in supporting safe online experiences, particularly for children. Technology companies and online service providers are the conduits for access to the online environment. They control the platforms on which users communicate, interact with each other and consume online content, and provide the vehicles for businesses to market and sell their products and services.

Some industry participants operating in Australia have taken a strong and positive approach to enhancing online safety, recognising that responsibility for tackling harmful behaviours and content goes hand-in-hand with their influential and important position within Australian society. Others have not taken this approach and have exposed users to online harms. It is particularly important that industry participants whose products and services are used by children ensure that they take appropriate action to uphold the safety of their users.

This consultation paper includes a draft Online Safety Charter (the draft Charter), at **Attachment A**. This draft Charter includes a set of proposed expectations or standards for technology firms regarding online safety. The Reader's Guide at **Attachment B** provides an explanation of these proposed online safety standards, and poses a number of questions to help guide responses and comments.

In releasing this draft Charter for public consultation, the Government is seeking to start a dialogue between the community, industry and Government about practical ways to implement the shared obligations for online safety.

Purpose

When it is finalised, the Charter will be an important foundation document to shape the direction of future reform of online safety policy and legislative arrangements in Australia. Although the Charter will not be mandatory and there will be no sanctions for non-compliance, it is intended to articulate a set of community-led minimum standards for industry to protect citizens, especially children and vulnerable members of the community, from harmful online experiences.

[1] Online safety means measures to address the risks and harms that individual users face online, such as exposure to illegal or age-inappropriate content, cyberbullying, hate speech and image-based abuse.

Scope

The draft Charter is directed towards technology firms that offer the opportunity for users in Australia to interact or connect, and technology firms whose services and products enable users to access content and information. This would potentially include social media services, internet service providers, search engine providers, content hosts, app developers, and gaming providers, among others. For the sake of simplicity, the draft Charter uses the term 'technology firms' to encompass these entities.

While the proposed scope of the Charter is broad, it is acknowledged that the digital media landscape is not homogeneous, and that not all technology firms should be expected to demonstrate, or implement, identical measures in relation to online safety.

Social media services, content hosts and app developers have different business models and their activities are quite distinct.

- Businesses that derive value from user-generated content – such as Facebook and YouTube – will potentially raise different online safety concerns and risks than search engine providers (Google Search and Bing) and messaging services (such as WhatsApp).
- In turn, these digital platforms and messaging service providers will raise different potential concerns to Internet Service Providers, such as Telstra and Optus.

This means that not all of the principles included in a Charter can or should be applied to all technology firms, and certainly not on a uniform and undifferentiated basis. The standards stipulated through a Charter will need to be tailored to, and appropriate for: the particular service or product (or class of service or product) in question; the size and impact of the service; and in particular the extent to which it is used by children.

In addition, the digital media landscape is not static. New services and new platforms will continue to emerge over the coming years, while existing apps, services and products may wane in terms of their impact, influence and reach. It is important that policy and legislative settings accommodate this and encourage a diverse and dynamic market, and don't have the perverse and unintended effect of stifling innovation and new business development.

Although the Charter will need to accommodate diversity and change in the digital media environment, it will be important to ensure, where possible, that there is a level of consistency in terms of the minimum safety standards for online safety.

Relationship with Safety by Design

In 2018, the Office of the eSafety Commissioner (eSafety Office) initiated a process to develop a Safety by Design (SbD) Framework. SbD aims to ensure that online safety issues are taken into account at the design stage for new products and services, and are embedded in software and devices. The development of the framework has involved extensive consultation with industry, parents, carers and young people, undertaken through meetings, calls for written submissions on draft principles, a national representative survey of parents and guardians and structured online forums.

The work of the eSafety Office in relation to SbD is consistent with the Charter. The Online Safety Charter seeks to establish standards for online safety by industry at the broadest level, while the SbD Framework takes elements of this work to a more specific level focused on the development and design stage in the product and service life cycle. The Government acknowledges that industry participants have worked collaboratively with the eSafety Office over the course of 2018 on developing the SbD framework, and encourages technology firms to engage in a similarly proactively manner in providing feedback on the standards proposed in this draft Charter.

Broader context

Online safety issues are closely linked with other issues arising in the digital technology space, including privacy, security, disinformation and competition, among others. Although the draft Charter focuses on online safety harms and their mitigation, the Government is conscious of these important linkages and overlaps. There are also other processes underway that will be considering these issues, including the Digital Platforms Inquiry being undertaken by the Australian Competition and Consumer Commission.

Outline of the draft Charter

The draft Charter is underpinned by two fundamental principles:

1. Standards of behaviour online should reflect the standards that apply offline.
2. Content that is harmful to users, particularly children, should be appropriately restricted.

The proposed online safety standards giving effect to these principles in the draft Charter are organised into one of four areas.

1. Control and responsibility
2. User experience
3. Built-in child safety
4. Accountability and transparency.

These areas are not mutually exclusive, and a proposed safety standard might fit into more than one area. Moreover, the draft Charter has a specific focus on children, recognising that children are vulnerable online users and need special protection from inappropriate content and other potential harms.

It is intended that the final Charter take account of and, to the extent possible, be consistent with best international practice to improve online safety. This recognises the fact that many of the larger technology firms are global, and that Australia’s success in improving online safety outcomes will be bolstered if we are consistent with successful overseas precedents. The final Charter will not be set in stone. It will be revisited regularly to ensure it remains relevant and takes account of emerging online safety practices.

Process and next steps

Views are sought on the proposed online safety standards contained in the draft Charter and the translation of these into real-world online safety measures. Some examples are included in the Reader’s Guide at **Attachment B**.

Comments are also sought in relation to the structure of the draft Charter, its application to different classes of products or services, and an indication of likely obstacles to implementation (e.g. technical, financial or legal). Information about actions taken by technology firms that could support the proposed online safety standards, and examples of good and best global practice, would also be valuable.

Submissions will be accepted on a confidential basis. However, the Government cannot guarantee that information provided on an ‘in-confidence’ basis will never be disclosed. Disclosure may be required under Australian law or as directed by a court or relevant tribunal. Information providers will be notified if there is any potential for disclosure.

Comments and views from industry and the public will inform the finalisation of the Charter, which is expected to occur in the second half of 2019.

Comments are sought by 5.00 pm Australian Eastern Daylight Time on Friday, 5 April 2019.

Attachment A—Draft Online Safety Charter

This Charter seeks to outline what the Australian Government, and the Australian community, expect of technology companies and online service providers operating in Australia in terms of protecting the most vulnerable in our community. It is underpinned by two fundamental principles:

1. Standards of behaviour online should reflect the standards that apply offline.
2. Content that is harmful to users, particularly children, should be appropriately restricted.

This Charter is directed towards technology firms that offer the opportunity for users in Australia to interact or connect, and technology firms whose services and products enable Australian users to access content and information. This includes social media services, internet service providers, search engine providers, content hosts, app developers, and gaming providers, among others. For the sake of simplicity, the Charter uses the term 'technology firms'.

1. Control and responsibility

1.1 Content identification

Technological solutions should be fully utilised by technology firms to identify illegal and harmful content, and these solutions should be supported by human resources as appropriate.

There should be a specific point of contact within each technology firm for the referral of complaints about illegal and harmful content or legal notices from Australian authorities. This point of contact should be equipped and trained to manage Australian referrals, with a good understanding of relevant Australian legal requirements.

1.2 Content moderation

The systems employed by technology firms should have the capability and capacity to moderate illegal and harmful content.

Where feasible, this should include a triaging system to ensure high risk content (e.g. content promoting self-harm or criminal activity) is addressed expeditiously and lower risk content is reviewed and actioned within a longer period (for example, within 24 hours).

This triaging system should ensure that complaints made by children, or by adults on behalf of children, are also expedited. Where appropriate, illegal, harmful or inappropriate content targeted towards a child should be removed immediately, and only reinstated once the complaint has been investigated and only if the complaint is not upheld.

The resources devoted to content moderation should be proportionate to the volume of content available to users and relevant to the Australian context. Human content moderators should meet minimum training standards.

Minimum timeframes should apply to the review and moderation of flagged content, whether identified from internal flags, user complaints or regulatory authorities.

1.3 Content removal

Content that is clearly and unambiguously illegal under Australian law should be removed proactively by technology firms.

Content that has been determined to be in breach of terms of use, or identified by regulatory authorities to be illegal or harmful, should be removed within clearly stated minimum timeframes.

Technology firms should take steps to prevent the reappearance of illegal, harmful or offensive content that has been removed.

2. Improving the user experience

2.1 User behaviour

Clear minimum standards for online behaviour should be set and applied consistently across services and service providers.

- Behaviour standards should be visible, easy to find and easy to understand.
- Behaviour standards should be reviewed regularly to ensure they remain fit-for-purpose and user-friendly.
- There should be meaningful and material consequences for breaches of behaviour standards, including account suspension, access restrictions and banning of repeat offenders.
- Banned users should not be able to open a new account in a different name or register a different user name.

2.2 User support

User reporting and complaints systems should be easy to find, understand and complete.

They should include a swift acknowledgement of each complaint and outline expected response timeframes.

They should provide regular updates to complainants and affected users (including the person being complained about), enable decisions to be reviewed, and provide full information to users on how to refer complaints to regulatory authorities in Australia.

Online safety resources should be actively promoted to users, age-appropriate and easy to understand. This should include mental health and other support services, where appropriate.

2.3 Account control

Instructions about how to adjust settings, including privacy settings, should be easy to find, understand and follow.

Users should be able to freeze their account in real time.

Users under 16 years should be required to secure parental or guardian consent to open an account or register as a user. Verifying parental consent should require more than just ticking a box.

Parental control settings should be easy to use and difficult to circumvent.

2.4 Content management

Users should be given full control of content safety options, such as the ability to delete unwanted comments, easily remove content, selectively hide content they no longer want to be visible and impose self-restrictions on uploading content such as time of day lockouts or type of content (for example, videos or images).

3. Built-in Child Safety

3.1 Default settings and age guidance

All products and services (including apps and games), and devices marketed to children, marketed as being appropriate for children, or that are likely to appeal to children, should default to the most restrictive safety and privacy settings at initial use or set up, and should include age guidance.

3.2 Supply chain

App and game supply points should require developers and suppliers to certify that they have considered built-in child safety and any relevant SbD principles before accepting apps and games for distribution.

Information about privacy, online safety and parental control settings should be available at all relevant points in the supply chain, including point-of-purchase (including by download), registration, account creation and first use.

4. Accountability and transparency

4.1 Reporting and compliance

Technology firms should engage broadly with experts and key stakeholders in relation to the development and application of online safety standards.

Technology firms should publish regular reports on:

- content controls, including the type of content is identified, moderated and/or prevented from being uploaded, how it was identified, and the action taken;
- complaints, including the number of complaints received, investigated and resolved, the time taken to resolve complaints, the category of complaint, the action taken and generalised demographic information (including, where known, age and geographic location of complainants); and
- compliance with the standards in this Charter, identifying any gaps and outlining the proposed approach to improving safety outcomes in relation to these gaps.

For firms with a significant presence in Australia, a local version of these reports should be published and the underlying data should be made available to relevant Australian authorities on request.

User safety considerations and practices should be embedded in the leadership structures, operating practices and governance arrangements for technology firms, and appropriate policies and procedures should be core business for all individuals who work within technology firms.

Attachment B—Reader's guide to the Draft Charter and discussion questions

1. Control and responsibility

The burden of safety should not fall wholly on the user. Technology firms have control over the content hosted or made accessible on their sites, apps and platforms, and they can take preventative steps to guard against their services being used to facilitate or encourage illegal content or conduct or inappropriate behaviours. The level of control will vary with the service and activity in question. However, even services that only host user-generated content should be required to meet minimum thresholds for content control and moderation because, ultimately, it is these services' algorithms that determine how and when content appears in users' feeds, and their systems that dictate how easy (or difficult) it is for users to access, upload and spread harmful and illegal content.

1.1 Content identification

A technology-facilitated problem requires, at least in part, a technology-facilitated solution. Technology firms should be required to actively prevent the upload of harmful and illegal content, put in place mechanisms that enable its swift removal and mitigate the potential for the inadvertent removal of legitimate content.

Technology-facilitated approaches

Technology firms should utilise the best technological solutions available to identify and remove illegal or harmful content on their services and to continue to develop these tools.

A non-exhaustive list of examples might include:

- detection algorithms, such as image hashing, to identify potentially illegal or harmful content;
- artificial intelligence (AI) to assist human moderators by prioritising potentially illegal or harmful content for review; or
- machine-learning to improve effectiveness and efficiency of content flagging processes.

Illegal content

Examples of content that is illegal under Australian law include child sexual abuse material (CSAM) and content inciting terrorism.

Designated contact point

Under Australia's cyberbullying scheme, social media companies are required to have a designated contact person to respond to complaints about cyberbullying content in Australia. This approach of providing a nominated single point of contact should apply equally to referrals from other authorities in relation to illegal and harmful online content.

Discussion questions:

1. ***What are the examples of technology-facilitated solutions to enhance online safety, and how effective have these solutions been in addressing harms and mitigating risks?***
2. ***What tools are available and have been deployed to address safety issues for live-streamed content as it occurs?***
3. ***What is the best way to establish a single 24/7 contact point for Australian authorities to ensure there is a timely response?***

1.2 Content moderation

Technology firms with millions of users posting billions of separate pieces of content should invest heavily in resources (both technology and human moderators) to prevent the spread of illegal and harmful content across their services. Resources should be scalable (to manage peaks in reporting and the size of the firm) and there should be low rates of misidentification of content.

Under Australia’s cyberbullying scheme, social media firms can be required to remove cyberbullying content targeting an Australian child within 48 hours. In general, social media firms have removed this material quickly without requiring that a formal social media service notice by given by the eSafety Commissioner. This standard would seek to embed a ‘take-down first’ approach when it comes to responding to complaints relating to children’s online safety. Technology firms should keep a record of material that is taken down, and removed content should be preserved so that it is available if needed as evidence by Australian authorities.

Good practice would see sufficient numbers of human moderators, provided with regular policy and legal training, and who have access to appropriate mental health and wellbeing training and support in order to identify and refer ‘at risk’ users, and to manage their own responses to disturbing content.

Content moderators reviewing content posted by Australians should also have an understanding of the Australian context, culture and community standards. For example, content moderators should be trained to know when to seek further advice about Indigenous Australian culture and protocols.

Backlogs in the review of flagged content should only occur in exceptional circumstance, given that many platforms rely heavily on a flag system to identify content for moderation. Best practice would see a triaging system employed to ensure that high risk content (e.g. CSAM, content promoting self-harm or criminal activity) is assessed and addressed immediately, and less urgent content reviewed and actioned within a specified period (for example, within 24 hours). In Germany, social networks are required by law to delete “manifestly unlawful” posts on their platforms within 24 hours of being notified (by complaints bodies or individuals), or within seven days for more legally ambiguous content.

Discussion questions:

4. ***Are there positive examples of flagging and content moderation? What makes these moderation systems work effectively and are they applicable to other services and applications?***
5. ***Is there an acceptable error rate for inappropriately flagged or misidentified content?***
6. ***What is an appropriate time frame for moderation and removal of content?***
7. ***How should content moderators be trained? What minimum standards should apply?***
8. ***What sort of guidance should be available to moderators about dealing with vulnerable groups, such as children and Indigenous Australians?***

1.3 Content removal

Content that does not need contextualisation to be identified as illegal should be removed expeditiously. Examples of such content would be CSAM, graphically violent images or any content that would be likely to be classified as X18+ or Refused Classification using Australia’s national classification system (e.g. films, games). Under the current Online Content Scheme (set out in Schedules 5 and 7 of the *Broadcasting Services Act 1992*), it is rare for such material to be formally classified before it is removed or reported to authorities.

Technology firms should retain records (including copies of videos, images and text) of any material removed to ensure that evidence is available and can be produced in an investigation by Australian authorities.

Known violating content should be removed from the internet. Although 'permanent removal' is generally not feasible, preventing access by making content difficult to find, such as by ensuring it does not appear in search engine results, is one way to reduce its availability.

Current best practice involves the use of 'hashes' or digital fingerprints to identify and prevent re-uploading of known CSAM or terrorist-related content. This technological approach also helps reduce the potential harm to moderators by ensuring that they are not required to unnecessarily re-review disturbing content.

Ideally, and where feasible, the use of 'hashes' and digital fingerprints to identify information removed from other platforms should be extended beyond illegal content to offensive content, such as cyberbullying material, to prevent it being uploaded on another platform.

Discussion questions:

9. ***Are there positive examples of identification and content removal practices? What makes these practices effective and appropriate?***

10. ***How should records of removed content be kept to ensure that evidence is available if needed by authorities?***

11. ***Are there minimum requirements to uniquely identify content (for example, IP addresses of upload/posting source, geographic identifiers etc)? If so, please provide details.***

12. ***Can content be made invisible on a permanent basis? If so, how?***

13. ***Are there barriers to sharing of information about offensive content removed by an industry participant to prevent it being uploaded to another platform or distributed using another service?***

14. ***What are the potential pitfalls and risks with content removal? How can these risks be mitigated?***

2. Improving the user experience

Technology firms have unique control of users' ability to engage online. The business models for many of these firms are based on user attention and use of this attention to generate revenue (via advertising). These services should ensure that users behave respectfully in terms of the content they upload, access and share, they should help users to have better online experiences, and resolve user complaints. In sum, the interests of users *as users* should be a primary and fundamental consideration for technology firms.

2.1 User behaviour

Terms of use and community standards should be in plain language and easily understood by all users, including children. Technology firms should establish a minimum age for account creation and have user terms that can reasonably be understood by the youngest permitted user group. For example, this might require pre-launch assessment of readability using focus groups.

Consequences for breach of terms of use must be enforced and made a priority. A lack of enforcement should not result in a competitive advantage for firms that turn a blind eye to poor conduct they could prevent or sanction.

The terms and enforcement procedures should be revised if there is evidence of systemic failure (i.e. persistent violations).

Discussion questions:

15. ***What should minimum standards of behaviour be? Should they be higher for products and services directed at children, or that have a substantial number of child users?***
16. ***How frequently should users be required to 'accept' or re-acknowledge terms of use, standards and policies?***
17. ***How should users be required to verify acceptance of terms of use, standards and policies?***
18. ***Are there positive examples of improving user experience currently in use?***

2.2 User support

Reporting and complaints systems should adhere to the fundamental principles of accessibility, fairness, responsiveness, efficiency and integration. Best practice could include:

- In-app reporting functions, for example reporting buttons on the content or conduct the user wants to report (a 'single click' model), and should support the ability to make multiple reports. Reporting functions should, at a minimum, be available on the same screen, page or window as the content.
- Reporting and complaints systems that are appropriate to the age of the users likely to use a site or service. This means they must be easy to use and understand, and avoid use of 'legalese'.
- Flagging or other reporting tools to speed up reporting content or conduct to Australian authorities, and which are aligned to reasonable reporting requirements (for example, to ensure that it is clear what needs to be included in a report).
- An acknowledgement of a report or complaint within 24 hours of receipt which outlines expected timeframes for updates/resolution, a reference code and a contact point for the complainant.
- Explanations of how a complaint/report will be handled and whether the outcome of the investigation will be provided.
- A status report to be given if a complaint is not resolved within minimum timeframes or content is not removed promptly, that includes an explanation for the delay.
- Appropriate training for staff of technology firms.
- Triaging of reports complaints and established escalation processes.
- Providing appropriate referrals to mental health and other support services.
- Review processes, which could include an independent party to arbitrate content moderation decisions on appeal.

Discussion questions:

19. ***Are there positive examples of user support systems and processes currently in use? What are the factors and characteristics of these systems and processes that make them effective?***
20. ***What timeframe is reasonable to respond to complaints and reports?***
21. ***Should reporting and complaint response timeframes vary depending on the complainant (e.g. child or adult), the type of content or other factors?***

2.3 Account and device control

Users should be able to easily understand and manage service and device settings. User profile options that automatically manage settings based on 'typical users' might be helpful, but each setting should be able to be managed independently of other settings.

Illegal or harmful content can be confronting and upsetting for users and the ability to freeze an account (or blank a screen) immediately gives a user an escape button to take time out from an emotionally distressing online experience.

Consent or confirmation that eligibility requirements have been satisfied (such as minimum age of account holders) should be appropriately verified by technology firms. It should not be sufficient for users to proceed with account registration simply by ticking a check box or clicking a button. Users eager to access content, especially younger users, might not be truthful.

Australians expect children to be protected from harmful activities. Australian laws establish minimum ages for potentially harmful behaviours such as drinking alcohol, smoking tobacco and gambling. Australians accept, and expect, that children and young people should be supervised by adults when engaging in potentially risky activities, such as learning to swim or drive a car. Parental controls (over account creation or access to content) provide an option for 'adult supervision' online.

Parental controls could include only permitting those under the age of 16 to access a service as part of a family account, or via an 'associate account' (an account associated with an adult user, that enable the 'responsible adult' to put appropriate protections in place).

There are many apps and software programs that provide 'parental control'. Whilst some are free, many must be purchased or require regular fees to be paid, and choosing the right one can be confusing. Many people find it too hard and don't seek out parental control options, or obtain them but don't understand how to use them. Technology firms are in the best position to provide, or promote, the parental controls that will work best with their products and services (for example, in terms of ease of use, compatibility, cost or other factors).

Discussion questions:

22. ***What options are there for verifying age or ensuring that parental/guardian consent is provided? Is there an optimal method or methods?***

23. ***Are there positive examples of parental settings currently in use?***

24. ***Are there barriers to obtaining or using parental controls? How can these barriers be managed and overcome?***

2.4 Content management

Current practice suggests that technology firms that enable posting, distribution and access to content don't offer a full range of user-controlled content management options. For example, no self-imposed user restrictions such as lockout times of day or bans based on content type have been identified.

Discussion questions:

25. ***Are there positive examples of user content management options currently in use?***

26. ***What user-controlled content management options should be available?***

3. Built-in child safety

Technology firms should embed safety principles and protections into their products and services as key features from the outset. This recognises the importance of getting it right from the start (including through use of SbD). However, even for established services and products, safety features should be introduced where appropriate. The term ‘built-in child safety’ is used in this draft Charter to emphasise that technology firms should focus on improving online safety for children across all stages of the product life cycle. This expectation goes beyond traditional IT industry concepts of SbD.

3.1 Default privacy settings and age guidance

Certain apps that are directly targeted towards children, such as Facebook Messenger Kids and YouTube Kids, have been designed with children’s safety in mind. Many products available in app stores also include ‘age appropriate’ ratings as guidance for consumers.

For products and services that are not explicitly marketed as ‘child friendly’, no examples of default ‘most-restrictive’ safety and privacy settings have been identified, even where the platform or service allows child users (under 16 years). The Government expects child users (and some services allow users from 13 years of age) to be given special protection. This is consistent with community attitudes in Australia.

Discussion questions:

27. ***Are there positive examples of age appropriate products or services currently available?***

28. ***To what extent do any technology firms restrict privacy and control settings as a default for younger users? If so, please provide detail.***

3.2 Supply chain

Some consumer sites enable categorisation of products as ‘child appropriate’. For example, the Google Play Store app categorises books, music and games that are suitable for children and provides advice on relevant age groups. Google Play also provides a Developer Policy Center which provides guidance about restricted content, apps designed for families and children, impersonation, security and deception. The Google Play Developer Distribution Agreement mandates compliance with policies as a condition of distribution.

Discussion questions:

29. ***Are there other positive examples of age guidance in the supply chain currently in use?***

30. ***Do any technology firms have mandatory requirements for products and services to be designed and marketed as suitable for children?***

31. ***Who should be responsible for ensuring built-in child safety?***

4. Accountability and transparency

Transparency and accountability are the hallmarks of a robust approach to online safety. Users should be assured that the services provided by technology firms are operating in accordance with their own published safety frameworks, and should have transparency and clarity with respect to how their complaints and concerns are being addressed. Members of the public should also be able to clearly see how firms are dealing with breaches terms of use and how firms are addressing online safety issues.

4.1 Embedding user safety considerations

The Royal Commission into Institutional Responses to Child Sexual Abuse found that child safety should be embedded into institutional leadership, governance and culture. It is appropriate that this principle extend to technology firms in relation to online safety. These firms should ensure that their policies, procedures and practices effectively address user safety considerations, and that this extends beyond the management level to all individuals working with, for, or on behalf of the technology firm in relation to their consumer-facing services and products.

Knowledge and expertise in relation to online safety should also be leveraged by technology firms, and they should establish open channels of communication with independent experts and bodies to ensure their services and products are well designed and that potential safety concerns are identified and addressed at an early stage.

4.2 Reporting and compliance

Greater transparency about how technology firms manage content and how they handle complaints and breaches of terms of use, standards and policies will build trust among users and enable third parties, including Government, to better monitor online safety efforts and evaluate success.

For example, in 2018 the United Kingdom (UK) introduced a requirement for social media companies to supply annual internet safety transparency reports to the UK Government. The reports are required to contain relevant UK data on what moderation policies each site has in place and how these are reviewed; how many complaints have been received; how they are dealt with; the volume of content removed; and information on how users can get help and access safety centres on their platforms.

Also in 2018, large technology firms (including Google, YouTube, Facebook and Twitter) published transparency reports for the first time and published their content moderation guidelines. While providing this information is an important first step, there should also be greater transparency on content moderation error rates and repeat violations to enable the success of current practices to be evaluated and areas for improvement identified.

Discussion questions:

32. ***Should relationships and engagement with independent experts be formalised, and what are the best mechanisms to achieve timely and productive input?***

33. ***What elements should be reported on and how can consistency of reporting be achieved?***

34. ***How often should reporting take place? The UK requires country-specific reporting. To what extent should a similar arrangement be developed in Australia?***