Amending the Online Safety (Basic Online Safety Expectations) Determination 2022—Consultation paper

November 2023

Purpose

The *Online Safety (Basic Online Safety Expectations) Determination 2022* (BOSE Determination) made under section 45 the *Online Safety Act 2021* (the Act), sets out the Government's minimum safety expectations of online service providers for protecting their Australian users. The BOSE Determination is an important instrument for promoting transparency and sets a benchmark for industry to take responsibility for online safety.

Since the BOSE Determination first commenced in January 2022, the online world has continued to evolve. To address these developments, the Minister for Communications, the Hon Michelle Rowland MP, has asked the Department of Infrastructure, Transport, Regional Development, Communications and the Arts (the Department) to consult on proposed amendments to strengthen the BOSE Determination. These proposed amendments are designed to address emerging online safety issues and gaps in the original BOSE Determination, and to improve its overall operation.

The draft Amendment Determination is available on the Department's website. We are inviting service providers directly affected by the proposed changes, and any other interested stakeholders, to make submissions on the draft Amendment Determination via the Department's website or email.

The Department welcomes feedback on the proposed changes, including whether the changes achieve their intended outcome, whether the changes are feasible, and any risks of implementation.

About the Basic Online Safety Expectations (BOSE)

The key principle underlying Australia's approach to online safety regulation is that industry is primarily responsible for creating safer online spaces. The BOSE Determination promotes this principle through the Government's basic safety expectations of online service providers, which establishes a benchmark for online service providers to take proactive steps to protect the community from abusive conduct and harmful content online. The eSafety Commissioner (the Commissioner) may require services to report on the steps they are taking to meet these expectations, providing transparency on how well industry is performing in its responsibility to Australians online.

Key features of the BOSE Determination

Part 4 of the Act establishes the BOSE regime. Section 45 of the Act authorises the Minister to make a legislative instrument determining the basic online safety expectations for social media services, [1] relevant electronic services, [2] and designated internet services. [3]

This instrument, the BOSE Determination, is composed of the following three elements.

1. **Core expectations** – set out in section 46 of the Act – these *cannot* be amended without an amendment to the Act itself. Core expectations include that the provider of the service will:
    a. take reasonable steps to ensure that end-users are able to use the service in a safe manner (subsection 6(1));
    b. consult the Commissioner in determining reasonable steps to ensure safe use (subsection 7(1));
    c. take reasonable steps to minimise provision of cyber-bullying, adult cyber abuse, non-consensual intimate images, class 1 material, and material that promotes, incites, instructs in, or depicts abhorrent violent conduct (section 11);
    d. take reasonable steps to prevent access by children to class 2 material (subsection 12(1));
    e. ensure the service has clear and readily identifiable mechanisms that enable end-users to report, and make complaints about, certain material provided on the service (subsection 13(1));
    f. ensure the service has clear and readily identifiable mechanisms that enable end-users to report, and make complaints about, breaches of the service's terms of use (subsection 15(1));
    g. on issue of a written notice, provide information to the Commissioner about: the number of complaints about breaches of the service's terms of use (subsection 20(1)); the length of time taken by the provider to comply with a removal notice (subsection 20(2)); and the measures taken to ensure end-users are able to use the service in a safe manner (subsection 20(3)).

[1] A 'social media service' is defined in section 13 of the Act. These are services that have the sole or primary purpose of enabling online social interactions between end-users, where end-users can also link to other end-users and post material on the service (e.g. Facebook, X (formerly Twitter), Instagram, Reddit and TikTok).
[2] A 'relevant electronic service' is defined in section 13A of the Act. These are services that allow end-users to communicate with other end users by means of email (e.g. Gmail), instant messaging (e.g. Whatsapp), SMS, MMS, chat services or online games.
[3] A 'designated internet service' is defined in section 14 of the Act. These are services, other than social media or relevant electronic services, that allow end-users to access material on the internet using an internet carriage service or a service that delivers material to persons by means of an internet carriage service (e.g. websites and other online services).

2. **Additional expectations** – determined by the Minister – these can be amended, and new expectations added in the BOSE Determination without amending the Act. Some examples of additional expectations include that a provider will take reasonable steps:
    a. to proactively minimise the extent to which material or activity on the service is unlawful or harmful (subsection 6(2));
    b. regarding encrypted services (section 8);
    c. to prevent anonymous accounts from being used to deal with material, or for activity, which is unlawful or harmful (subsection 9(1));
    d. to consult and cooperate with providers of other services to promote the ability of end-users to use all of those services in a safe manner (subsection s10(1));

    in addition to having:
    e. guidance on how to make a complaint to the Commissioner (section 16);
    f. accessible terms of use, policies and procedures in relation to end-user safety, reports and complaints, and standards of conduct (sections 14 and 17);
    g. records of reports and complaints about certain material (section 19).

3. **Examples of reasonable steps that services can take in order to meet a core or an additional expectation –** determined by the Minister – these can be amended through the BOSE Determination without amending the Act.

The BOSE Determination does not prescribe how expectations should be met. This is intended to provide flexibility for service providers to determine the most appropriate method for achieving the expectations. Notwithstanding this, the BOSE Determination outlines a number of examples of reasonable steps that could be taken to achieve expectations. Not all reasonable steps have to be taken by all service providers – the examples are guidance only.

The Commissioner has recently published updated Regulatory Guidance to provide service providers with information about the expectations and the functions of the Commissioner in assessing compliance with those expectations.[4]

Scope of the BOSE Determination

Under section 45 of the Act, the BOSE Determination only applies to social media services, relevant electronic services and designated internet services. Classes of services covered by some parts of the Act are not necessarily covered by the BOSE Determination.

As a legislative instrument, the BOSE Determination cannot amend the Act or extend the Commissioner's functions and powers beyond what is contained in the Act. It can only operate within the scope it is empowered to cover under Part 4 of the Act.

BOSE reporting under the Act

Under the Act, the Commissioner may require a provider or class of providers to report on their compliance with one or more basic online safety expectations specified in the BOSE Determination. Subdivision 3-A (periodic reporting about compliance with the basic online safety expectations) and subdivision 3-B (non-periodic reporting about compliance with the basic online safety expectations) provide the Commissioner with the power to issue:

- a periodic reporting notice requiring an individual provider to report to the Commissioner on their compliance with the basic online safety expectations multiple times at regular intervals;
- a periodic reporting determination requiring each provider within a class of providers to report multiple times at regular intervals;
- a non-periodic reporting notice requiring an individual provider to prepare only a single report to be given to the Commissioner;

[4] https://www.esafety.gov.au/about-us/who-we-are/regulatory-schemes#basic-online-safety-expectation

- a non-periodic reporting determination requiring each provider within a class of providers to each prepare a report to be given to the Commissioner.

The Commissioner may issue a reporting notice under Division 3 of Part 4 of the Act seeking information about a provider's compliance with *all,* or *one or more specified,* basic online safety expectations. The provider must prepare the report in the manner and form specified in the Commissioner's notice, and give the report to the Commissioner either within the time period specified in the reporting notice, or such longer period as the Commissioner allows (but not less than 28 days).

A provider who fails to comply with a reporting notice from the Commissioner may be subject to a civil penalty. In addition to a court ordered civil penalty, the Commissioner has access to other enforcement options, including formal warnings, service provider notifications, infringement notices, enforceable undertakings and injunctions.

The Commissioner can issue and publish service provider notifications under sections 48, 55 and 62 of the Act outlining that a service is not meeting the expectations in the BOSE Determination, which may create reputational risks for service providers. There are no civil penalties for failure to comply with the expectations outlined in the BOSE Determination, nor does the BOSE Determination impose a duty that is enforceable by court proceedings.

Proposed amendments to the BOSE Determination

The online environment continues to rapidly evolve which presents new challenges, including developments related to generative artificial intelligence (AI) and recommender systems. The proposals for amendments to the BOSE Determination outlined in this paper and set out in the draft Amendment Determination, encompass new or amended additional expectations, examples of reasonable steps, and explanatory notes to better provide for the safety of end-users. The proposals are discussed below under the following themes:

1. Generative AI, recommender systems and user controls;
2. The best interests of the child and access to age-inappropriate materials online;
3. Safety impacts of business and resourcing decisions;
4. Online hate speech;
5. Transparency;
6. Enforcement of terms of use; and
7. Other clarification and improvements to the BOSE Determination.

Other Government priorities to address online harms

The BOSE Determination is just one mechanism that the Government is using to address a range of emerging online harms. Other relevant Government workstreams include:

- The **independent statutory review of the Act** – the Government has committed to bring forward the statutory review of the Act to ensure it remains fit for purpose in a rapidly evolving online environment. The review will present an opportunity for stakeholders to consider the scope, operation and effectiveness of the Act, including the BOSE regime in Part 4 of the Act. The review will be conducted by Ms Delia Rickard PSM with a period of public consultation commencing in early 2024. The review is anticipated to be completed by the second half of 2024.

- Response to **group hate speech online** – the proposal outlined in this Consultation Paper provides one mechanism for addressing the growing concern of group hate speech online. The Government will continue to consider what more can be done to address group hate speech online, including through the independent statutory review of the Act.

- The **voluntary code of practice for online dating services** – the Government has requested that online dating services develop a code of practice to better protect Australians using their services. The proposals outlined in this Paper are intended to complement measures that online dating services may implement through the voluntary code.
- **Discussion Paper on AI** – On 1 June 2023, the Government released the *Safe and responsible AI in Australia Discussion Paper*. The paper sought feedback on governance mechanisms to ensure AI is used safely and responsibly. In particular, the discussion paper sought stakeholder views on any gaps in existing laws and the domestic governance landscape for AI, and further regulatory and governance mechanisms to mitigate emerging risks. Consultation on the paper closed on 4 August 2023. Feedback will inform consideration across Government on appropriate regulatory and policy responses.
- **Algorithms commitment** – coming out of the *Australian Government response to the House of Representatives Select Committee on Social Media and Online Safety report*, the Department and the Department of Home Affairs are progressing work to understand algorithms on digital platform services. This work is seeking to understand how algorithms operate on digital platform services, identify potential harms attributed to algorithms on digital platforms, and report to Government on possible regulatory reform options for algorithms. The Departments are due to report back to Government in Quarter 1, 2024.
- Measures to address the **complaints and dispute resolution processes** of digital platforms – the Australian Competition and Consumer Commission's September 2022 Digital Platforms Services Inquiry interim report found that digital platforms' complaints and dispute resolution processes are ineffective. It recommended the Government establish mandatory minimum standards for internal dispute resolution processes and an external ombudsman scheme for this industry. The Government is considering the report and is expected to publish a response. The proposals outlined in this Paper signal the Government's desire for online services to improve their complaints processes for the matters dealt with in the Act, ahead of any proposed regulatory action more broadly.
- **Reforms to the *Privacy Act 1988* (Privacy Act)** – on 28 September 2023, the Government released its response to the Privacy Act Review Report. Reforms will ensure Australia's privacy framework is fit for purpose in the digital age, and provide Australians with greater transparency and control over their personal information.

1. Generative AI, recommender systems and user controls

Additional expectations – Generative AI capabilities

The BOSE Determination empowers the Commissioner to request reports on how service providers are deploying technologies to enhance user safety. A technology that is becoming increasingly prevalent, both as a means of providing new functionalities for end-users and improving user experience, is generative AI. However, the accessibility and usability of generative AI capabilities also creates potential for the production of harmful material and activity on a scale and speed not previously possible. While the BOSE Determination already captures generative AI capabilities, in light of this potential for harm, new additional expectations will emphasise and clearly signal the Australian public's interest in how service providers are designing and deploying these capabilities.

Proposal 1: Proposed section 8A will create new additional expectations which provide:

Section 8A – Additional expectations – provider will take reasonable steps regarding generative artificial intelligence capabilities

(1) If the service uses or enables the use of generative artificial intelligence capabilities, the provider of the service will take reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of artificial intelligence capabilities on the service.

(2) If the service uses or enables the use of generative artificial intelligence capabilities, the provider of the service will take reasonable steps to proactively minimise the extent to which generative artificial intelligence capabilities may be used to produce material or facilitate activity that is unlawful or harmful.

(3) Without limiting subsection (1) and (2), reasonable steps for the purpose of this section could include the following:

(a) ensuring that assessments of safety risks and impacts are undertaken, identified risks are appropriately mitigated, and safety review processes are implemented throughout the design, development, deployment and post-deployment stages of generative artificial intelligence capabilities;

(b) providing educational or explanatory tools (including when new features are integrated) to end-users that promote understanding of generative artificial intelligence capabilities on the service and any risks associated with the capabilities;

(c) ensuring that training materials for generative artificial intelligence capabilities and models do not contain unlawful or harmful material;

(d) ensuring that generative artificial intelligence capabilities can detect and prevent prompts that generate unlawful or harmful material.

The new additional expectation at subsection 8A(1) provides that at all stages during the life cycle of generative AI capabilities (from development to implementation and maintenance), service providers must consider user safety and incorporate safety features to minimise any risks that have been identified. The new additional expectation at subsection 8A(2) provides that service providers must take reasonable steps to proactively minimise the extent to which generative AI capabilities produce material or facilitate activity that is unlawful or harmful. This would cover, for example, the production of 'deepfake' intimate images or videos, class 1 material such as child sexual exploitation or abuse material, or the generation of images, video, audio or text to facilitate cyber abuse or hate speech. These additional expectations are central to ensuring that users can enjoy the benefits of generative AI capabilities, while being protected against some of the most significant harms that can be facilitated through its use.

New paragraph 8A(3)(a) proposes an example of a reasonable step to achieve these additional expectations. This draws from the example of a reasonable step contained in paragraph 6(3)(e), but applies specifically to the design, deployment and post-deployment of generative AI capabilities. This means that service providers should take steps to identify how its generative AI capabilities are working

in practice to minimise unlawful and harmful material and activity. Assessments could identify gaps in generative AI capabilities so that providers can then take appropriate steps to mitigate them. Given the speed with which generative AI capabilities are developing, regular assessments about the operation and impact of these technologies will help address emerging risks more effectively and efficiently.

New paragraph 8A(3)(b) outlines another example of a reasonable step. It provides that services should increase transparency by providing explanatory or educational material to end-users outlining how generative AI capabilities are deployed on a service, and the risks those capabilities may pose. These explanatory or educational materials should also be updated as new features are integrated into the generative AI capability, with visibility about these new features and updated explanatory or educational materials being promoted through user prompts and nudges. These measures will enable end-users to make informed decisions about engaging with generative AI capabilities on a service and have greater awareness about any risks associated with material that is created, presented, or distributed on the service using generative AI capabilities.

New paragraph 8A(3)(c) provides that service providers should ensure that training materials for generative AI capabilities and models do not contain unlawful or harmful material. Service providers must make important decisions when developing generative AI capabilities about the content and quality of the input data that is used to train the generative AI model. If this process is not managed appropriately, there is a risk that training data could include unlawful and harmful content, such as child sexual exploitation and abuse material, image-based abuse, hate speech and other online harms.[5] This can lead to the generation of unlawful or harmful material by the generative AI capability. To the extent possible, service providers should proactively remove unlawful or harmful content from training material so as to mitigate the risk of online safety risks emerging later in the lifecycle of the capability. Improving training data quality, including by removing unlawful and harmful material, is an ongoing process that must be managed throughout the lifecycle of the generative AI capability.

New paragraph 8A(3)(d) provides that service providers should ensure that generative AI capabilities can detect and prevent prompts that generate unlawful or harmful content. This is an important mechanism to limit the production of unlawful or harmful material in the first instance. Measures such as educative prompts or nudges can also provide a valuable opportunity for end-users to reconsider their engagement with a generative AI capability and curb misuse.[6] As with proposed paragraph 8A(3)(c), these preventive efforts must be ongoing throughout the lifecycle of the generative AI capability.

Generative AI capabilities will continue to evolve. In parallel, governments are increasingly looking to proactive forms of safety testing and evaluation of foundation, or frontier, models underpinning generative AI capabilities prior to their deployment. In November 2023, Australia was one of 29 countries to sign the Bletchley Declaration on AI Safety, committing countries to collaborating internationally with industry and academia on safety standards and testing mechanisms, incorporating a breadth of safety-related issues including: accuracy and reliability; cyber security vulnerabilities; and online user harms. The Government has appointed CSIRO Chief Scientist Professor Bronwyn Fox as Australia's representative on the Expert Advisory Panel on the annual State of the Science report, which summarises the latest international research on AI safety.

The types and volume of harm that may be produced or facilitated through the use of generative AI will continue to change. It is incumbent on service providers to both proactively and reactively respond to these challenges by ensuring that the systems, tools and processes they employ to prevent online harms continue to evolve so that they can provide effective protections for end-users.

[5]Tech Trends Position Statement – Generative AI (August 2023), eSafety Commissioner, p.6. The Position Statement is available here: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai.
[6] Ibid, p. 8.

Additional expectations – Recommender systems

Recommender systems are 'systems that prioritise content or make personalised content suggestions to users of online services. A key element of the system is the recommender algorithm, a set of computing instructions that determines what a user will be served based on [a range of] factors.'[7]

Australians engage with recommender systems routinely, particularly through social media platforms like Facebook, Instagram, TikTok, X (formerly Twitter) and YouTube. Recommender systems can help end-users by exposing them to information, ideas, products, artists and friends which can benefit a user's online and offline experiences, and help businesses reach new audiences.[8] However, they can also cause significant online harm. As noted by eSafety, there is the potential for recommender systems to amplify harmful and extreme content, and increase the risk of virality of harmful or hateful content.[9] At a societal level, the amplification of harmful content can increase the likelihood of discrimination, such as racism, sexism and homophobia and normalise such prejudice or hatred. It can also contribute to radicalisation towards terrorism or violent extremism.[10]

As noted with respect to generative AI capabilities, the BOSE Determination already provides that services develop and deploy new technologies in a manner that promotes user safety. However, given the prevalence of recommender systems and the extent of user engagement with this technology, a stronger emphasis is warranted to ensure there is a clear signal that service providers are responsible for designing and deploying this technology with user safety in mind.

Proposal 2: Proposed section 8B will create new additional expectations which provide:

Section 8B – Additional expectations – provider will take reasonable steps regarding recommender systems

(1) If the service uses recommender systems, the provider of the service will take reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of recommender systems on the service.

(2) If the service uses recommender systems, the provider of the service will take reasonable steps to ensure that recommender systems are designed to minimise the amplification of material or activity on the service that is unlawful or harmful.

(3) Without limiting subsection (1) and (2), reasonable steps for the purpose of this section could include the following:

- *(a) ensuring that assessments of safety risks and impacts are undertaken, identified risks are appropriately mitigated, and safety review processes are implemented throughout the design, development, deployment and post-deployment stages of recommender systems;*
- *(b) providing educational or explanatory tools (including when new features are integrated) to end-users that promote understanding of recommender systems on the service, their objectives, and any risks associated with such systems;*
- *(c) enabling end-users to make complaints or enquiries about the role recommender systems may play in presenting material or activity on the service that is unlawful or harmful.*

[7] Recommender systems and algorithms – Position Statement (December 2022), eSafety Commission, p. 1. The Position Statement is available here: https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms.

[8] https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms

[9] Recommender systems and algorithms – Position Statement (December 2022), eSafety Commission, p. 3-5.

[10] Ibid, p. 3-5.

The new additional expectation at subsection 8B(1) makes clear that service providers are expected to consider how the design and implementation of recommender systems on their service might impact end-users and incorporate safety measures to minimise any risks. This assessment should cover the lifecycle of the recommender system, including its implementation and maintenance phases.

The new additional expectation at subsection 8B(2) will provide that services take reasonable steps to minimise unlawful or harmful material or activity that can be caused by recommender systems deployed on their service. While services may not always be able to prevent what a recommender system serves up, services are responsible for developing and adopting the algorithms that recommender systems are founded on (i.e. the inputs). It is reasonable to expect that services make design choices in the development of algorithms so as to minimise the risk of unlawful or harmful material or activity being served to end-users. Such measures are not unusual or unprecedented. For instance, the European Union's *Digital Services Act,*[11] which came into effect for large platforms on 25 August 2023, has instituted regulations governing recommender systems, such as mandatory transparency requirements, in order to address algorithmically amplified harms.

Consistent with the examples of reasonable steps in relation to generative AI capabilities, paragraphs 8B(3)(a) and (b) are intended to ensure service providers identify gaps in the safety associated with recommender systems so as to appropriately mitigate them, and to increase transparency by providing explanatory or educational material to end-users outlining how recommender systems are deployed on the service, and the risks those systems may pose.

Paragraph 8B(3)(c) will provide that services should establish mechanisms (or amend existing mechanisms) so that end-users can make complaints or raise queries in relation to the operation of recommender systems and their role in presenting unlawful or harmful material or activity on the service. It is important that end-users not only flag unlawful or harmful material or activity on the service that may relate to the content outlined in section 13 or breaches of terms of use, policies, procedures and standards of conduct, but also raise more systematic concerns that a services' recommender system is presenting unlawful or harmful material or activity. This may allow service providers to identify underlying risks in the operation of their recommender systems and take more effective steps to prevent the provision of such material or activity on the service in the future.

Additional expectation – user empowerment controls

Services should design features and functionality that enable users to protect their own best interests and give them autonomy in how they engage with a service. A key tenet of eSafety's *Safety by Design* is that user empowerment can be supported by services providing 'technical measures and tools that adequately allow users to manage their own safety, and that are set to the most secure privacy and safety levels by default'.[12]

Proposal 3: Proposed subsection 6(5) will create a new additional expectation which provides:

(5) The provider of the service will take reasonable steps to make available controls that give end-users the choice and autonomy to support safe online interactions.

(6) Without limiting subsection (5), reasonable steps for the purposes of that subsection include the following:

(a) making available blocking and muting controls for end-users;

(b) making available opt-in and opt-out measures regarding the types of content that end-users can receive;

(c) enabling end-users to make changes to their privacy and safety settings.

[11] Article 27

[12] https://www.esafety.gov.au/industry/safety-by-design/principles-and-background

This new additional expectation will set the standard that service providers are expected to make user empowerment controls available to their end-users to support safe online interactions and experiences. These tools will enable end-users to control how they engage with the service provider, the content on the service, and with other end-users. Paragraphs (6)(a), (6)(b) and (6)(c) provide some examples of user empowerment controls that service providers should be employing on their service.

Paragraph (6)(a) provides that an example of a reasonable step for making available user empowerment controls is providing blocking and muting controls. Blocking allows an end-user to prevent other users from following and interacting with their account, or from viewing their posts and activity. Muting allows an end-user to hide other accounts and their activity, content, content tags etc. from visibility on their feed. These controls, blocking especially, are often accompanied with an option to report the account or content being blocked or muted for breaches of terms of use, policies or procedures, or standards of conduct.

Paragraph (6)(b) provides for user content controls allowing users to opt-in or opt-out from receiving or viewing certain content, such as class 2 material or content which may be distressing or upsetting to them even if it is does not infringe a service's terms of use, policies or procedures, or standards of conduct. Opt-in controls provide users with the ability to decide whether they wish to view certain kinds of content before it is made visible. Such controls include having default opt-in controls for adult (class 2) and (where possible) other potentially distressing images and video, such as warning prompts and content blurring. Opt-in controls can also include the option to hide potentially distressing text with content warning tags. Opt-out controls allow users to flag or filter content they do not want to see, and can be particularly helpful in the context of recommender systems.

Paragraph (6)(c) provides that end-users should be able to make changes to their privacy and safety settings. Providing users with autonomy over their privacy and safety settings is critical to ensure that end-users can engage with a service in a manner that suits their particular preferences, context and concerns. However, this should not prevent services from setting particularly robust and restrictive default settings in particular contexts, such as for services or components of services that are targeted at, or used by, children (see the example of a reasonable step in paragraph 6(3)(b)).

The examples mentioned above are only some of the user empowerment controls that service providers could employ. Other controls should also be considered such as tools to flag content, providing links to external reporting options (such as eSafety), feedback on reports or complaints made, and providing education and information on how users can keep themselves safe (such as through pop-up warnings when an end-user is about to engage in activity such as sharing personal information).

2. The best interests of the child and preventing access to age-inappropriate materials online

Safety of children, continually implement technologies to prevent access to class 2 material, and appropriate age assurance mechanisms

The BOSE Determination contains certain protections for children, including providing for privacy and safety settings to be set to the most restrictive level for services used by children (paragraph 6(3)(b)), and ensuring that service providers take reasonable steps to ensure that technological and other measures are in place to prevent access by children to class 2 material provided on the service (subsection 12(1)).[13]

[13] The Government responded to eSafety's Age Verification Road (Roadmap) in August 2023. The Roadmap does not recommend the Government legislate age assurance technology for access to pornography and notes that the technology is immature but developing. The Government Response to the Roadmap noted that the Government will consider a pilot of age assurance technology following the development of the next phase of industry codes under Part 9 of the Act, which will require services to do more to protect children from exposure to online pornography. The Roadmap and Government Response can be found here:

However, the BOSE Determination does not currently contain any provisions that encourage services to consider the best interests of the child throughout the development and implementation phases of an online service. Further, the BOSE Determination does not explicitly encourage services to focus on continually improving technologies to better prevent children from accessing class 2 material or to ensure that age assurance mechanisms are appropriate to the level of risk.

Proposal 1: Proposed subsection 6(2A) will create a new additional expectation which provides:

> *The provider of the service will take reasonable steps to ensure that the best interests of the child are a primary consideration in the design and operation of any service that is used by, or accessible to, children.*

Proposal 2: Amended paragraph 12(2)(a) will provide that an example of a reasonable step to ensure that technological and other measures are in effect to prevent access by children to class 2 material is:

> *implementing **appropriate** age assurance mechanisms;*

Proposal 3: Proposed paragraph 12(2)(c) will provide a new example of a reasonable step to ensure that technological and other measures are in effect to prevent access by children to class 2 material:

> *continually seeking to develop, support or source, and implement improved technologies and processes for preventing access by children to class 2 material.*

New subsection 6(2A) provides that service providers will design and implement any service that is used by, or accessible to children, in a manner consistent with the objectives underlying Article 3 of the *Convention of the Rights of the Child* that "[i]n all actions concerning children … the best interests of the child shall be the primary consideration".

Noting the unique vulnerabilities of children, and their particular susceptibility to online harms, it is important that the best interests of children are treated as a priority throughout the lifecycle of a service and in relation to all aspects of a service, including when designing and implementing new features. While the best interests of the child are traditionally understood as an analysis based on the specific circumstances of a child,[14] such an approach would not be feasible in the context of service providers that have a large cohort of users (including children) with varying circumstances and needs. Rather, services will be expected to consider the best interests of the child generally, including having regard to the physical, psychological and emotional wellbeing of children on a service. It is likely that different services, functions and features will pose different safety risks to children so it is helpful for services to consider and identify these risks early, and take appropriate steps to ensure that children can use the service, functions and features safely. Subsection 6(3) has also been amended to include reference to new subsection 6(2A) to highlight that the examples of reasonable steps that have been listed provide useful guidance and can be considered in the context of children (noting paragraph 6(3)(b) already applies directly to children).

The proposed reasonable step at paragraph 12(2)(a) will signal to services that age assurance mechanisms need to be appropriate. Age assurance is an umbrella term which can include both age verification and age estimation mechanisms. The inclusion of the word 'appropriate' signals that age assurance mechanisms to prevent children's access to class 2 material should be calibrated to the level of risk and harm of the material. This means that in some instances, asking users to self-report their age or date of birth may provide an effective signal or barrier to unintentional access by children, while in other instances, services will be expected to establish a user's age with a greater level of certainty that is appropriate for the level of risk of the material they may access on the service.

https://www.infrastructure.gov.au/department/media/publications/australian-government-response-roadmap-age-verification.

[14] See General Comment No. 14 (2013) on the right of the child to have his or her best interests taken as a primary consideration (Article 3, Paragraph 1), para 32, and Privacy Act Review Report 2022, p. 152.

The proposed new reasonable step at paragraph 12(2)(c) will recognise that technologies and systems for age assurance and age appropriate design are continuously improving. This new example of a reasonable step will encourage service providers to keep up to date and to seek ways to improve upon or refine their existing approaches.

3. Safety impacts of business and resourcing decisions

Reasonable step regarding business decisions affecting user safety

Proposal 1: Proposed paragraph 6(3)(f) is a new example of a reasonable step to ensure that end-users are able to use the service in a safe manner:

> *assessing whether business decisions will have a significant adverse impact on the ability of end-users to use the service in a safe manner and in such circumstances, appropriately mitigating the impact;*

The new example of a reasonable step in paragraph 6(3)(f) will clarify that providers not only need to consider safety impacts on end-users in Australia in the development and implementation of their products and services, but also when making business decisions that are likely to have a significant adverse impact on the ability of end-users to use their service in a safe manner. If business decisions are likely to have a significant adverse impact, service providers should take appropriate steps to minimise these impacts.

Service providers are also encouraged to document these important decisions and the mitigations put in place. In developing potential mitigations, service providers should engage with relevant stakeholders, including trust and safety staff and external experts, and monitor the ongoing impacts of the business decision and regularly assess the effectiveness of the mitigations put in place.

It is intended that the Explanatory Statement to the Amendment Determination will provide guidance, through the use of non-exhaustive examples, on the types of business decisions that may have a significant adverse impact on the ability of end-users to use the service in a safe manner. Examples may include significant changes to a service's terms of use, policies or procedures and standards of conduct that diminish the protections available to end-users (such as permitting abusive conduct that was previously prohibited); decisions relating to the creation of subscription tiers with different safety features for each tier; major staffing cuts or removing staff from certain countries, regions; or service functions that reduce a service's ability to identify and address (whether proactively or reactively) unlawful or harmful content on the service effectively and efficiently. The broad framing of 'significant adverse impact' means that service providers have flexibility in how they meet this expectation depending on their specific circumstances and the specific circumstances of the business decisions they are making. This broader framing also provides flexibility to the Commissioner in issuing reporting notices relating to a range of organisational decisions that ultimately impact user safety.

Reasonable steps regarding resourcing and investment

Appropriate resourcing and investment in human and technological interventions is critical to a service provider's ability to safeguard end-users from material and activity that is unlawful or harmful. While the BOSE Determination contains some measures encouraging services to uplift technology and practices, it does not expressly encourage investment of resources for the purpose of achieving better safety outcomes for end-users.

Proposal 2: Proposed paragraph 6(3)(g) is a new example of a reasonable step to support services to respond to reports and complaints within a reasonable time (as required by new subsection 14(3)):

> *(g) having staff, systems, tools and processes to action reports and complaints within a reasonable time in accordance with subsection 14(3);*

Proposal 3: Proposed paragraph 6(3)(h) is a new example of a reasonable step to ensure that end-users are able to use the service in a safe manner:

> *(h) investing in systems, tools and processes to improve the prevention and detection of material or activity on the service that is unlawful or harmful;*

While the new examples of a reasonable step outlined above may be broadly covered by the existing examples of reasonable steps in subsection 6(3), there is benefit in explicitly drawing out these steps as they go towards key elements of a service provider's ability to protect end-users. Rather than mandating a particular level of resourcing or investment, the new examples of reasonable steps would be outcomes focused (i.e. responding to reports and complaints promptly and minimising material or activity on the service that is unlawful or harmful).

The new example of a reasonable step at paragraph 6(3)(g) will provide that services should have adequate staff and systems, tools and processes to respond to reports and complaints by end-users in a timely manner. This is linked to the new additional expectation is subsection 14(3) which will provide that services should respond to reports and complaints within a 'reasonable period of time'. The ability to meet this new additional expectation is contingent on services having an appropriate level of resourcing and effective systems, tools and processes to review and respond to reports and complaints efficiently.

The new example of a reasonable step at paragraph 6(3)(h) will encourage service providers to ensure there is appropriate investment in systems, tools and processes that prevent and detect unlawful and harmful content or activity. This example of a reasonable step builds on paragraph 6(3)(d) which provides that services should continually improve technology and practices relating to the safety of end-users. A focus of this new example of a reasonable step is to support proactive minimisation of unlawful or harmful material or activity on a service. This means service providers should be particularly interested in investment targeted at improving the prevention and detection of such material or activity on their service. This could be through investment in safety technologies such as hash matching, machine learning and artificial intelligence, and investment in personnel to enhance their ability to detect unlawful or harmful material or activity, and take appropriate action.

Noting that safety solutions will continue to develop and improve over time, ongoing investment in systems, tools and processes is needed to ensure that providers continually seek out proactive and preventive safety solutions and apply these to their services. The extent of investment will vary from service to service, depending on the service's objectives, demographics, and risk profile. It is acknowledged that for some providers, certain systems, tools and processes may not be reasonable to implement because of a lack of widely available technology.

Designated contact point

Proposal 4: For the avoidance of doubt, a new explanatory note will be included at the end of subsection 21(1) stating that:

> *The provider of the service is expected to have a designated contact point regardless of whether the service has staff physically located in Australia.*

The new explanatory note will confirm that service providers are expected to comply with the expectation to provide the Commissioner a designated contact point for the purposes of the Act, irrespective of whether the service provider has staff physically located in Australia. This is intended to ensure that where a service provider, who previously had staff physically located in Australia, decides to cut staff or move staff overseas, they will still be expected to provide the Commissioner with a designated contact point.

4. Hate Speech

Reasonable step regarding detecting and addressing hate speech

According to eSafety research, approximately 1 in 7 Australian adults (aged 18-65) were targeted by online hate speech in the 12 months to August 2019 (approximately 2 million people).[15] Online hate speech disproportionately impacts many groups, including people identifying as Aboriginal or Torres Strait Islander or as LGBTQI+, who experience online hate speech at more than double the national average.[16] Online hate speech can have significant negative impacts, silencing individuals and forcing their withdrawal from the online world. The current BOSE Determination does not include any specific expectations in relation to online hate speech. However:

- Section 6 of the Determination includes a reasonable step that providers will minimise the provision of unlawful and harmful material on a service. Harmful material can include material that should fall under a service provider's terms of use, policies and procedures in relation to user safety, and standards of conduct for end-users.
- Section 14 of the Determination includes that a service provider will ensure it has terms of use, policies and procedures in relation to user safety, and policies and procedures for dealing with reports and complaints. Providers are expected to take reasonable steps to ensure penalties for breaches of the terms of use are enforced against offending end-users. [17]

Proposal 1: Proposed paragraph 6(3)(i) will provide that a reasonable step in ensuring end-users are able to use a service in a safe manner is by:

> *having processes for detecting and addressing hate speech which breaches a service's terms of use and, where applicable, breaches a service's policies and procedures and standards of conduct mentioned in section 14.*

New subsection 6(4) will provide a non-exhaustive definition of 'hate speech':

> *For the purposes of paragraph 6(3)(i),* ***hate speech*** *is a communication by an end-user that breaches a service's terms of use and, where applicable, breaches a service's policies and procedures or standards of conduct mentioned in section 14, and can include communication which expresses hate against a person or group of people on the basis of race, ethnicity, disability, religious affiliation, caste, sexual orientation, sex, gender identity, disease, immigrant status, asylum seeker or refugee status, or age.*

As outlined above, hate speech is covered by the BOSE Determination insofar as services are expected to have and enforce terms of use, which may prohibit the posting of hate speech (see additional expectations at subsections 14(1) and (2)). It is generally the case that major service providers' terms of use prohibit hate speech or hateful conduct. The new example of a reasonable step in paragraph 6(3)(i) will build on these existing expectations by encouraging services to have processes for detecting and addressing hate speech if it violates their terms of use, policies and procedures or standards of conduct.

New subsection 6(4) will provide a non-exhaustive definition of 'hate speech' to clarify the intended meaning of the term and the types of communication that may constitute hate speech. The proposed definition is not intended to override the definition of 'hate speech' or 'hateful conduct' that services may use in their terms of use, policies and procedures or standards of conduct. For the avoidance of doubt,

[15] eSafety Commissioner (2020), *Online hate speech – Findings from Australia, New Zealand and Europe,* https://www.esafety.gov.au/research/online-hate-speech.
[16] Ibid.
[17] Note, it is proposed that subsection 14(2) will also cover enforcement against breaches of a service's policies and procedures in relation to user safety and standards of conduct for end-users (see *Theme 6 – Enforcement of terms of use,* Proposal 2).

the proposed non-exhaustive definition of hate speech in new subsection 6(4) is indicative only and will only apply for the purposes of the BOSE Determination.

5. Transparency

Additional expectation – Transparency reporting

The BOSE Determination does not currently contain any expectations for service providers to publish information about the safety measures deployed on their service, the effectiveness of those measures, or how their service is enforcing its own terms of use, policies and procedures and standards of conduct. Transparency reporting would provide users with relevant information so that they can make informed choices about the services on offer and the safety measures available.[18]

Proposal 1: Proposed section 18A will create a new additional expectation which provides:

Section 18A – Additional expectation – provider will publish transparency reports

(1) The provider of the service will publish regular transparency reports, at regular intervals of no less than 1 month and no more than 12 months, with information regarding:
- *(a) the service's enforcement of its terms of use, policies and procedures and standards of conduct mentioned in section 14;*
- *(b) the safety tools and processes deployed by the service (including in relation to a service's key features), and their effectiveness;*
- *(c) metrics on the prevalence of harms, reports and complaints, and the service's responsiveness; and*
- *(d) the number of active end-users of the service in Australia (including children) each month during the relevant reporting period.*

(2) For the purpose of paragraphs 1(a) to (d), the information and data contained in a transparency report must be specific to Australia, unless to do so is not reasonably practicable.

(3) For the purposes of this section, a transparency report must:
- *(a) identify each relevant service a provider provides (where applicable);*
- *(b) set out the information regarding the matters in paragraph (1)(a) – (d) separately, in respect of each service provided by a provider; and*
- *(c) be published within a reasonable time of the end of the relevant reporting period to which the report relates.*

This additional expectation provides that services will prepare regular transparency reports with information and data specific to Australia (where reasonably practicable). Where a provider provides multiple services, transparency reports must include information and data specific to each service. The specific information and data covered in paragraphs 18A(1)(a) to (d) will provide valuable information to the Commissioner and the public in understanding what safety measures a service has in place and how effective those measures are, how services are enforcing their terms of use, policies and procedures and standards of conduct, and the prevalence of harms on a service.

It is also intended that service providers will provide information on how they are ensuring key features, such as generative AI and recommender systems, are being kept safe and preventing the generation, posting, distribution and amplification of unlawful or harmful activity or material. The information and data contained in a transparency report should not include personal information and should be de-identified to protect users' privacy. From an end-user perspective, transparency reports paint a useful

[18] Note, it is proposed that subsection 14(2) will also cover enforcement against breaches of a service's policies and procedures in relation to user safety and standards of conduct for end-users (see *Theme 6 – Enforcement of terms of use*, Proposal 2).

picture of how service providers are promoting safety on their platforms, allowing end-users to make informed choices about how they engage with a service.

It is proposed that service providers determine the frequency with which they publish transparency reports, provided they do so at least annually. It is also expected that transparency reports be published within a reasonable time of the end of the relevant reporting period (i.e. the relevant quarter or the end of the financial year) to ensure the reports are timely and retain currency.

The new additional expectation provides that unless it is not reasonably practicable, the information and data contained in the transparency report must be specific to Australia. This is to ensure that transparency reports are relevant to Australia and provide sufficient detail to enable the Commissioner and the public to make assessments about the extent to which a service is protecting Australian end-users. Where providing Australia-specific information and data is not reasonably practicable, services should consider providing information and data from the closest regional equivalent (such as Australia and New Zealand).

Consideration will be given to whether this additional expectation should only be triggered where service providers meet specific conditions, such as having a certain number of end-users in Australia, or where the service providers meet certain criteria that suggest they may pose a material risk to online safety in Australia.

Additional expectation – Reporting on Australian end-usage of services

Currently the BOSE Determination enables the Commissioner to issue written notices to service providers under section 20 requesting a variety of information relating to breaches of terms of use, removal timeframes and measures to ensure user safety. However, there is no ability for the Commissioner to request information about the number of end-users a service has in Australia.

Proposal 2: Proposed subsection 20(5) is a new additional expectation which will provide:

> *Notwithstanding section 18A, if the Commissioner, by written notice given to a provider of the service, requests the provider to give the Commissioner a report on the number of active end-users of the service in Australia (including children) during a specified period, the provider will comply with the request within 30 days after the notice of request is given.*

While proposed section 18A provides that service providers publish the number of monthly active end-users of their service in Australia as part of their transparency reporting, there may be circumstances where the Commissioner may require information on the number of active end-users in Australia (including children) in advance of the publication of a service provider's transparency report.

Information about the number of Australian end-users of particular services will assist the Commissioner in assessing the reach and prevalence of the service within Australia, and consequently the level of risk a service poses to Australians. This will improve the Commissioner's capacity to support Australians by identifying where Australians are most likely to need support, and enable more efficient deployment of resources.

6. Enforcement of terms of use

Detecting breaches of terms of use, policies and procedures and standards of conduct

Proposal 1: Proposed subsection 14(1A) is a new additional expectation which will provide:

> *The provider of the service will take reasonable steps (including proactive steps) to detect breaches of its terms of use and, where applicable, breaches of policies and procedures in relation to the safety of end-users, and standards of conduct for end-users.*

It is important that service providers do not rely just on end-users bringing to their attention breaches of their terms of use, policies and procedures and standards of conduct for end-users, in order to prevent unlawful or harmful material or activity on their platforms. Platforms are encouraged to take reasonable steps, including proactive measures, to identify breaches, and then take reasonable steps to enforce penalties for breaches in accordance with subsection 14(2).

Reasonable steps that service providers could take include technological interventions that detect material and activity that may breach their terms of use, policies and procedures and standards of conduct, either before it is created, uploaded or shared on a service, or immediately after it is provided on the service. Examples of such measures include hash matching technology to detect known videos or images of unlawful material such as child sexual exploitation and abuse material and terrorism material, AI classifiers to identify new material that could be unlawful or harmful which then get prioritised for human review, and technologies such as language or text analysis.[19]

Service providers could also use proactive nudges and prompts to notify end-users that the material they are about to upload, send, or otherwise share may be in breach of a service's terms of use, policies and procedures and standards of conduct.[20] Additionally, services should remain alert and detect to ongoing patterns of unlawful and harmful behaviour that breaches its terms of use, policies and procedures and standards of conduct once such conduct has been reported by others.[21]

These measures, combined with services responding in a timely manner to user reports and complaints, would help minimise the extent and duration of unlawful and harmful material or activity on a service that breaches the service's term of use, policies and procedures and standards of conduct.

Enforcement of policies and procedures and standards of conduct, and readily identifiable reporting mechanisms

Currently, subsection 14(2) of the BOSE Determination applies to breaches of a service's terms of use, but does not apply to breaches of a service's policies and procedures in relation to the safety of end-users (referenced in paragraph 14(1)(b)), and standards of conduct for end-users (referenced in paragraph 14(1)(d)). This a potential gap as a service's policies, procedures and standards of conduct may contain important requirements relating to what material end-users can create and share on the service (such as prohibitions on creating and sharing class 1 material, hate speech, cyber abuse and cyberbullying) and what kind of activities are permitted (and not permitted). It is proposed that, in addition to breaches of a service's terms of use, subsection 14(2) also applies to breaches of a service's policies, procedures and standards of conduct.

Proposal 2: It is proposed to amend subsection 14(2) to provide:

> *The provider of the service will take reasonable steps **(including proactive steps)** to ensure that any penalties specified for breaches of its terms of use, **policies and procedures in relation to the safety of end-users, and standards of conduct for end-users,** are enforced against all accounts held or created by the end-user who breached the terms of use **and, where applicable, breached the policies and procedures, and standards of conduct, of the service**.*

[19] Basic Online Safety Expectations – Regulatory Guidance (September 2023), eSafety Commission, p. 44 – 45. eSafety's Regulatory Guidance can be found here: https://www.esafety.gov.au/about-us/who-we-are/regulatory-schemes#basic-online-safety-expectations.
[20] Ibid, p. 45.
[21] Ibid.

Relatedly, it is also proposed to amend subsection 15(2) to provide:

> *The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable any person ordinarily resident in Australia to report, and make complaints about, breaches of the service's terms of use* ***and, where applicable, breaches of the service's policies and procedures and standards of conduct mentioned in section 14****.*

These amendments will remove the existing gap in the BOSE Determination by ensuring that additional expectations which currently apply in respect of terms of use, also apply in respect of policies, procedures, standards of conduct for end-users, where it is possible for end-users to breach those policies, procedures and standards of conduct, and where penalties apply for such breaches.

Amended paragraph 14(2) will also provide that service providers should take proactive steps to enforce any penalties that have been specified in relation to breaches of its terms of use, policies, procedures and standards of conduct. This should be read with together with new subsection 14(1A) which encourages service providers to take reasonable steps on their own to detect breaches of terms of use, policies and procedures and standards of conduct. Once breaches have been identified, service providers should take proactive steps to enforce any applicable penalties for those breaches against all accounts held or created by the responsible end-user.

Amended subsection 15(2) will build on amended subsection 14(2) by requiring providers to make clear and readily identifiable mechanisms that enable any person ordinarily resident in Australia to make reports and complaints about breaches of a service's policies and procedures in relation to the safety of end-users (referenced in paragraph 14(1)(b)), and standards of conduct referenced in paragraph 14(1)(d)). This is to reflect, consistent with amended subsection 14(2), that terms of use may not contain or link to a service's policies and procedures and standards of conduct relating to unlawful or harmful material. In such cases, it is not sufficient to only have clear and readily identifiable mechanisms to make reports and complaints about breaches of terms of use. There must also be clear and readily identifiable mechanisms to make reports and complaints about breaches of a service's policies and procedures and standards of conduct.

Additional expectation – Timely resolution of complaints and reports

The BOSE Determination does not currently contain any expectations about the speed with which service providers should review and respond to reports and complaints about unlawful and harmful material or activity, or the need to keep those who make reports and complaints informed of what action has been taken in response.

Proposal 3: Proposed subsection 14(3) (accompanied by new subsections 14(4) and 14(5)) will create a new additional expectation which provides:

> *(3) The provider of the service will, within a reasonable period of time:*
> *(a) review and respond to reports and complaints mentioned in sections 13 and 15; and*
> *(b) provide feedback on the action taken.*
> *(4) For the purposes of subsection (3), in determining 'a reasonable period of time', the provider must have regard to:*
> *(a) the nature and impact of the harm that is the subject of the report or complaint;*
> *(b) the complexity of investigating the report or complaint; and*
> *(c) any other relevant matters.*
> *(5) For the purposes of paragraph (3)(a):*
> *(a)* ***review*** *means considering a report or complaint when it is first made; and*
> *(b)* ***respond*** *means taking and implementing a decision to have content removed and reported, have an end-user banned, or other content moderation decisions, or a decision to take no action.*

Proposed new subsections 14(3), 14(4) and 14(5) will address these gaps. While it is important that service providers have complaint and reporting mechanisms for unlawful and harmful material on the service, it is equally important that services address reports and complaints within a reasonable period of time so as to minimise harm. It is also important that those who make reports and complaints receive feedback on how their reports and complaints are being addressed.

New subsection 14(4) provides guidance about the factors that service providers should consider when determining what is a reasonable period of time to review, report and provide feedback in relation to the complaint or report. The nature and impact of the harm that is the subject of the complaint or report will generally be the primary consideration that a service provider should have regard to. However, the complexity of the complaint or report may also affect the time taken to review and respond. For example, it is reasonable to expect that reports and complaints relating to class 1 material (such as child exploitation material or terrorism material) be actioned immediately to prevent ongoing harm, whereas other reports and complaints which may require a more nuanced analysis of the context and circumstances of the relevant conduct, such as allegations of hate speech, may take longer. Other relevant matters a service must have regard to could include referring to any guidance material that is made available by the Commissioner. The timely resolution of reports and complaints can often be aided by having systems and processes that allow service providers to triage and prioritise the most severe and harmful reports and complaints for review.

Proactive prevention of recidivism through use of anonymous accounts

Currently, paragraph 9(2)(a) provides that one example of a reasonable step to prevent anonymous accounts from being used to deal with unlawful or harmful material is having processes that prevent the same person from repeatedly using anonymous accounts to engage in such conduct. However, this example of a reasonable step does not provide that services should take *proactive* measures to prevent anonymous accounts being used to engage in this conduct.

Proposal 4: Amended paragraph 9(2)(a) provides that a reasonable step to prevent anonymous accounts from being used to deal with material or activity that is unlawful or harmful is:

> *having processes, **including proactive processes**, that prevent the same person from repeatedly using anonymous accounts to post material, or to engage in activity, that is unlawful or harmful;*

This amendment emphasises that services should take proactive steps to prevent individuals from using anonymous accounts to circumvent enforcement action (e.g. bans or suspensions) taken by a service against them. ‘Anonymous accounts’ includes ‘pseudonymous accounts’ where an individual has registered with a service without using their real name. It is important that services do not simply rely on responding to user reports and complaints in identifying individuals who may have previously posted material, or engaged in activity, that is unlawful or harmful.

This proposal is also linked closely to the enforcement of penalties for breaches of terms of use, policies and procedures and standards of conduct in subsection 14(2), as effective enforcement should include measures that prevent circumvention of enforcement actions. It is appropriate to encourage service providers to take reasonable steps to proactively prevent anonymous accounts from being used to repeatedly engage in conduct that is unlawful or harmful through the deployment of tools and detection of signals (for example email, phone number, IP address).

For completeness, it is noted that paragraph 9(2)(a) must be read in accordance with Australian Privacy Principles (APP) 2 – Anonymity and pseudonymity. APP 2 provides that individuals must have the option of dealing anonymously or by pseudonym with APP entities, unless the entity is required or authorised by

law or a court or tribunal order to deal with identified individuals, or it is impractical for the entity to deal with individuals who have not identified themselves.[22]

7. Other clarifications and improvements

Assessments of safety risks and mitigations

Currently, paragraph 6(3)(e) provides that a reasonable step for meeting the expectations in subsections 6(1) and 6(2) is ensuring that assessments of safety risks and impacts are undertaken, and safety review processes are implemented, throughout the design, development, deployment and post-deployment stages of the service. This example of a reasonable step does not provide any guidance on what should be done if safety risks have been identified through this process.

Proposal 1: Amended paragraph 6(3)(e) will provide that an example of a reasonable step to ensure that end-users are able to use the service in a safe manner by:

> *ensuring that assessments of safety risks and impacts are undertaken,* ***identified risks are appropriately mitigated,*** and safety review processes are implemented, throughout the design, development, deployment and post-deployment stages for the service;

Amended paragraph 6(3)(e) will encourage service providers to appropriately mitigate the risks identified. Service providers should document the risks they have identified after undertaking a safety risk and impact assessment, and consider what appropriate mitigations could be instituted to address the risks. Mitigation measures should be commensurate to identified risks and be subject to consultation with relevant stakeholders, such as trust and safety staff and external experts.

While it may not be possible to eliminate all safety risks in all instances, service providers should ensure the mitigation measures proposed minimise safety risks as much as possible, and regularly monitor the impacts of the mitigations to determine whether they are operating as intended and are effective.

Cooperation across service providers to promote safe use

Currently, subsection 10(1) encourages service providers to take reasonable steps to consult and cooperate with providers of *other services* to promote the ability of end-users to use all those services in a safe manner. Although implicit in this expectation, it does not explicitly account for the fact that some providers operate multiple services.

Proposal 2: Amended subsection 10(1) will provide:

> *The provider of the service will take reasonable steps to consult and cooperate with providers of other services, and to take reasonable steps to ensure consultation and cooperation occurs between all relevant services provided by that provider in order to promote the ability of end-users to use all of those services in a safe manner.*

This amendment clarifies that the intention of the additional expectation is that consultation and cooperation occur between all relevant services operated by the same provider, [23] as well as services operated by different providers. The amendment refers to 'relevant' services, as not all services provided by a provider may be covered by the BOSE Determination. Minor consequential amendments are proposed to paragraphs 10(2)(a) and (b) to reflect this clarification.

[22] For further guidance on APP 2, refer to the Australian Privacy Principles (APP) Guidelines on APP 2 here: https://www.oaic.gov.au/privacy/australian-privacy-principles-guidelines/chapter-2-app-2-anonymity-and-pseudonymity.

[23] The services covered by the BOSE Determination are social media services, relevant electronic services and designated internet services.

Consultation and next steps

This consultation is being conducted in compliance with the requirements in section 47 of the Act for varying the BOSE Determination. This Consultation Paper is intended to assist interested parties to understand the existing expectations in the BOSE Determination, and why changes are being proposed. The Department welcomes feedback on the proposed changes, whether the changes achieve the intended outcome, the feasibility of the changes and any risks in implementation.

The draft Amendment Determination is available on the Department's website.[24] The Department invites submissions on the draft Amendment Determination from individuals and interested stakeholders by **5pm on Friday 16 February 2024**.[25] Submissions may be lodged in the following ways:

Website https://www.infrastructure.gov.au/have-your-say/online-safety-basic-online-safety-expectations-amendment-determination-2023

Email BOSEreform@communications.gov.au

Submissions should include your name, organisation (if relevant) and contact details. The Department will not consider submissions without verifiable contact details.

To promote transparency, submissions may be made publicly available on the Department's website **unless you specifically request that your submission, or part of a submission, be kept confidential**, or if we consider (for any reason) that it should not be made public. If you would like personal information contained in your submission to be redacted prior to publication (such as names), please provide us with the relevant details to make those redactions.

The submissions made will inform the Department's development of recommendations to the Minister in respect of proposed changes to the BOSE Determination. In varying the BOSE Determination, the Minister will have due regard to the comments provided during the consultation process.

Questions about the consultation or submission process can be directed to BOSEreform@communications.gov.au.

[24] https://www.infrastructure.gov.au/have-your-say/online-safety-basic-online-safety-expectations-amendment-determination-2023

[25] The consultation period is greater than the minimum 30-day consultation requirement under section 47 of the Act.