Online Safety (Basic Online Safety Expectations) Amendment Determination 2024

I, Michelle Rowland, Minister for Communications, make the following determination.

Dated

Michelle Rowland
Minister for Communications

Contents

1 Name

This instrument is the *Online Safety (Basic Online Safety Expectations) Amendment Determination 2024*.

2 Commencement

This instrument commences the day after this instrument is registered.

3 Authority

This instrument is made under section 45 of the *Online Safety Act 2021*.

4 Schedules

Each instrument that is specified in a Schedule to this instrument is amended or repealed as set out in the applicable items in the Schedule concerned, and any other item in a Schedule to this instrument has effect according to its terms.

Schedule 1—Amendments

Online Safety (Basic Online Safety Expectations) Determination 2022

1 After subsection 6(2)

Insert:

Additional expectation

(2A) 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 likely to be accessed by children.

2 Subsection 6(3)

Omit "without limiting subsection (1) or (2), reasonable steps for the purposes of this section" and substitute "without limiting subsection (1), (2) or (2A), reasonable steps for the purposes of those subsections".

3 Paragraph 6(3)(b)

Repeal the paragraph, substitute:

(b) if a service or a component of a service (such as an online app or game) is likely to be accessed by children (the ***children's service***) – ensuring that the default privacy and safety settings of the children's service are robust and set to the most restrictive level;

4 Paragraph 6(3)(e)

Repeal the paragraph, substitute:

(e) ensuring that assessments of safety risks and impacts are undertaken (including child safety risk assessments), identified risks are appropriately mitigated, and safety review processes are implemented, throughout the design, development, deployment and post-deployment stages for the service;

5 After paragraph 6(3)(e)

Insert:

(f) 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;

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

(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;

(i) 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;

(j) preparing and publishing regular transparency reports that outline the steps the service is taking to ensure that end-users are able to use the service in a safe manner, including:

(i) the use of online safety tools and processes;

(ii) providing metrics on the prevalence of material or activity on the service that is harmful;

(iii) the service's responsiveness to reports and complaints; and

(iv) how the service is enforcing its terms of use, policies and procedures and standards of conduct mentioned in section 14.

Additional expectation

(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.

Examples of reasonable steps that could be taken

(6) Without limiting subsection (5), reasonable steps for the purposes of that subsection could 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.

6 Paragraph 8(2)(a)

Repeal the paragraph, substitute:

(a) implement or build a systemic weakness, or a systemic vulnerability, into a form of encrypted service;

7 After section 8

Insert:

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 generative 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.

Examples of reasonable steps that could be taken

(3) Without limiting subsection (1) or (2), reasonable steps for the purposes 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, to the extent reasonably practicable, that training material for generative artificial intelligence capabilities and models do not contain unlawful or harmful material;

(d) ensuring, to the extent reasonably practicable, that generative artificial intelligence capabilities can detect and prevent the execution of prompts that generate unlawful or harmful material.

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 proactively minimise the extent to which recommender systems amplify material or activity on the service that is unlawful or harmful.

Examples of reasonable steps that could be taken

(3) Without limiting subsection (1) or (2), reasonable steps for the purposes 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;

(d) where technically feasible, enabling end-users to opt-out of receiving recommended content, or providing alternative curation options.

8 Paragraph 9(2)(a)

Repeal the paragraph, substitute:

(a) 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;

9 Subsection 10(1)

Repeal the subsection, substitute:

(1) The provider of the service will take reasonable steps to:

(a) consult and cooperate with providers of other services; and

(b) 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.

10 Paragraphs 10(2)(a) and (b)

Repeal the paragraphs, substitute:

(a) working with other service providers and between all relevant services provided by a service provider to detect high volume, cross-platform attacks (also known as volumetric or 'pile-on' attacks);

(b) sharing information with other service providers and between all relevant services provided by a service provider on material or activity on the service that is unlawful or harmful, for the purpose of preventing and dealing with such material or activity.

11 Paragraph 12(2)(a)

Repeal the paragraph, substitute:

(a) implementing appropriate age assurance mechanisms;

12 After paragraph 12(2)(b)

Insert:

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

13 After subsection 14(1)

Insert:

(1A) 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.

14 Subsection 14(2)

Repeal the subsection, substitute:

(2) 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.

15 After subsection 14(2)

Insert:

(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) take reasonable steps to 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 from 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.

16 Subsection 15(2)

Repeal the subsection, substitute:

(2) 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.

17 After subsection 20(4)

Additional expectation

(5) 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 (disaggregated into active end-users who are children and those who are adult end-users) during a specified period, the provider will comply with the request within 30 days after the notice of request is given.

18 After subsection 21(1)

Insert:

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

EXPLANATORY STATEMENT

Issued by the Authority of the Minister for Communications

Online Safety Act 2021

Online Safety (Basic Online Safety Expectations) Amendment Determination 2024

Authority

The *Online Safety (Basic Online Safety Expectations) Amendment Determination 2024* (Amendment Determination) is made under section 45 of the *Online Safety Act 2021* (the Act). The Amendment Determination is a legislative instrument for the purposes of the *Legislation Act 2003*.

Purpose

The Amendment Determination amends the *Online Safety (Basic Online Safety Expectations) Determination 2022* (the Determination). Under the Determination, online service providers, including social media services, relevant electronic services and designated internet services are expected to take steps to protect Australians from unlawful and harmful material and activity that falls within the remit of the enabling legislation, the *Online Safety Act 2021* (the Act), or impedes the online safety of Australians, and to report on these steps as required.

Amendments to the Determination explicitly address online harms that have developed since it was made, including those associated with generative artificial intelligence and recommender systems, as well as the need to strengthen protections for children online. Amendments also seek to further improve the scheme's overall operation and address gaps identified through the eSafety Commissioner's (the Commissioner's) exercise of existing powers.

Summary of the Amendment Determination

The Determination includes both 'core expectations', established by the Act, and 'additional expectations,' and also provides 'examples of reasonable steps' to meet the expectations. The Amendment Determination includes both new additional expectations and new examples of reasonable steps. Additional expectations added by the Amendment Determination include that services:

- Ensure that the best interests of the child are a primary consideration in the design and operation of any service likely to be accessed by children;
- Make available controls that give users choice and autonomy in deciding who they interact with, the content they receive, and their level of privacy;
- Consider user safety in the design and operation of generative artificial intelligence capabilities, and proactively minimise the extent to which they are used to produce or facilitate unlawful or harmful material (including deepfake images) and activity;
- Consider user-safety in the design and operation of recommender systems, and proactively minimise the extent to which they amplify unlawful or harmful material;
- Take reasonable steps (including proactive steps) to detect breaches of their terms of use, and respond to user complaints about unlawful and harmful material within a reasonable period of time; and

- Provide, on request of the Commissioner, a report on the number of active end-users of the service in Australia – broken down according to numbers of users who are children or adults.

Examples of reasonable steps that services could take to meet expectations added through the amendments include:

- Publishing regular transparency reports about the measures they are taking to keep Australian end users safe on their services;
- Having processes for detecting and addressing hate speech which breaches their terms of use;
- 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 take steps to mitigate the impact;
- Having staff, systems, tools and processes in place to action complaints, and invest in systems, tools and processes for detecting unlawful or harmful material or activity; and
- Continually seeking to improve technologies and processes for preventing access by children to class 2 material.

The Amendment Determination was issued by the Minister under section 45 of the *Online Safety Act 2021* (the Act).

Details of the instrument are set out in Attachment A.

Consultation

Section 47 of the Act requires the Minister to undertake specified consultation prior to making or varying a determination under section 45 of the Act. The Act requires that the Minister must make a copy of the draft determination available on the Department's website and invite written comments for a period of at least 30 days. The Minister must also have due regard to those comments in making or varying the determination.

Public consultation on a draft instrument, as is required under section 47 of the Act, was conducted from 22 November 2023 to 16 February 2024 and 52 submissions were received. The most significant revision following consultation was scaling down the publication of regular transparency reports from an 'additional expectation' to an 'example of a reasonable step' to ensure safe use of a service, due to industry concerns about efficacy relative to impost. In addition, a proposed definition of hate speech was removed from the draft instrument due to concerns about imposition on free speech, but included in the Explanatory Statement so that guidance could be given outside of the legislative context. Minor revisions were also made to improve clarity and flexibility.

The Office of Impact Assessment has confirmed that the preparation of a Regulation Impact Statement is not necessary, as the regulatory burden associated with reporting on the Determination was considered in the Regulation Impact Statement for the Online Safety Act (RIS ID 25408).

Statement of Compatibility with Human Rights

A statement of compatibility with human rights for the purposes of Part 3 of the *Human Rights (Parliamentary Scrutiny) Act 2011* is set out at Attachment B.

Commencement

The Amendment Determination commences on the day after it is registered on the Federal Register of Legislation.

Attachment A

Details of the Online Safety (Basic Online Safety Expectations) Amendment Determination 2024

Section 1 – *Online Safety (Basic Online Safety Expectations) Determination 2024*

This section provides that the name of the Amendment Determination is the *Online Safety (Basic Online Safety Expectations) Determination 2024*.

Section 2 – Commencement

This section provides that each provision of the Amendment Determination is to commence on the day after the Amendment Determination is registered on the Federal Register of Legislation.

Section 3 – Authority

This section provides that the Amendment Determination is made under section 45 of the *Online Safety Act 2021* (the Act).

Section 4 – Schedules

This section provides that each instrument that is specified in a Schedule to this instrument is amended or repealed as set out in the applicable items in the Schedule concerned, and any other item in a Schedule to this instrument has effect according to its terms.

Schedule 1 – Amendments

Schedule 1 sets out amendments to the *Online Safety (Basic Online Safety Expectations) Determination 2022* (the Determination).

Item 1 inserts a new additional expectation at subsection 6(2A), after subsection 6(2) in the Determination. It provides that the provider of the service will take reasonable steps to ensure that the best interests of the child (defined under the Act as an individual who has not reached 18 years of age) are a primary consideration in the design and operation of any service that is likely to be accessed by children.

In the digital environment, where children may be particularly susceptible to certain harms, protective rights such as privacy, protection from abuse and exploitation are highly relevant.

This expectation is worded to align with Article 3(1) of the Convention of the Rights of the Child, which provides that "[i]n all actions concerning children, the best interests of the child shall be a primary consideration." The principle requires welfare institutions and legislative, administrative and judicial bodies to take active measures to protect children's rights, promote their wellbeing and consider how children's rights and interests are, or will be, affected by their decisions and actions.

This principle is applied to providers of social media, relevant electronic and designated internet services that are likely to be accessed by children. These service providers are expected to give high priority to protecting and promoting the full enjoyment by children of all their rights, recognising their particular vulnerabilities and state of development. They are expected to act in the best interests of children, and assess how their actions impact the best

interests of children, throughout the lifecycle of a service, in respect of all aspects of a service's design and operation, including technologies, and a service provider's systems, tools and processes.

Subsection 6(2A) applies to services that are "likely to be accessed by children". This establishes a standard and threshold that is intended to align with the UK Information Commissioner's Age Appropriate Design Code, making it simpler for services to assess whether or to what extent the expectation applies to them. In assessing whether a service is likely to be accessed by children, service providers should consider factors such as:

- The nature and content of the service, and whether it has a particular appeal to children;
- Market research, current evidence on user behaviour, the user base of similar or existing services and service types; and
- The way in which the service is accessed, and whether any measures put in place are effective in preventing children from accessing the service.

Service providers are expected to proactively assess the likelihood that their service is accessed by children, and should not disregard this expectation on the grounds that their service is not explicitly targeted at children.

The expectation applies broadly to services that are likely to be accessed by children, and should be applied in meeting other relevant expectations in the Determination (as amended by the Amendment Determination), including the expectations regarding generative artificial intelligence and recommender systems at sections 8A and 8B under Item 7.

Item 2 amends subsection 6(3) to include reference to new additional expectation 6(2A). This is to provide that the examples of reasonable steps included in subsection 6(3) (such as having content moderation processes, highest default privacy and safety settings for children, having staff trained in online safety, continually improving technologies and practices, conducting safety risk assessments, and the steps outlined under Item 5) are also reasonable steps that could be taken for the purposes of additional expectation 6(2A) in relation to the best interests of the child.

Item 3 repeals and replaces paragraph 6(3)(b), replacing "targeted at, or being used by," with "likely to be accessed by". This amendment brings the language of the existing paragraph about a children's service in line with that used in subsection 6(2A).

Item 4 repeals and replaces paragraph 6(3)(e) to provide that:

- Assessments of safety risks and impacts undertaken include child safety risk assessments, for the purpose of the additional expectation at 6(2A); and
- Risks identified through these assessments are appropriately mitigated, and clarifies that service providers should not only identify safety risks and impacts through assessments, but also consider and implement appropriate mitigations. Mitigation measures should be commensurate with identified risks and be subject to consultation with relevant stakeholders, such as trust and safety staff and external experts.

Item 5 inserts paragraphs 6(3)(f), 6(3)(g), 6(3)(h), 6(3)(i), 6(3)(j) and subsections 6(5) and 6(6) after paragraph 6(3)(e). These paragraphs provide new examples of reasonable steps that

could be taken for the purposes of meeting the expectations that providers will take reasonable steps to ensure safe use in subsections 6(1), 6(2) and 6(2A).

Paragraph 6(3)(f) provides that service providers could help meet the expectations in subsections 6(1), 6(2) and 6(2A) by 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. Decisions made by service providers which affect the operation of the service should be made so as to not increase the likelihood of prevalence of unlawful or harmful material or activity, adversely affect vulnerable users such as children, or otherwise make the service less safe. Relevant decisions that ought to be assessed for their safety implications may include, but are not limited to:

- Significant changes to a service's terms of use, policies and procedures and standards of conduct;
- The creation of different subscription tiers or account types with different safety features;
- Major staffing changes, such as reductions in trust and safety staff; and
- Changes to a service's technical architecture, features or functions that affect a service's ability to detect and address unlawful or harmful content.

Paragraph 6(3)(g) provides that an example of a reasonable step service providers could take in meeting the expectations set out in subsections 6(1), 6(2) and 6(2A) is having staff, systems, tools and processes to action reports and complaints within a reasonable period of time in accordance with subsection 14(3). This is linked to the new additional expectation at Item 15 (new subsection 14(3)) which provides that services should respond to reports and complaints within a 'reasonable period of time'. The safety of users of a service is contingent upon services having an appropriate level of resourcing and effective systems, tools and processes to review and respond to reports and complaints efficiently.

Paragraph 6(3)(h) provides that an example of a reasonable step service providers could take in meeting the expectations set out in subsections 6(1), 6(2) and 6(2A) is investing in systems, tools and processes to improve the prevention and detection of material or activity on the service that is unlawful or harmful. "Investment" is not necessarily limited to financial investment, but could include a broad range of initiatives, including participation in and support for research, pilot projects, and collaboration with law enforcement, non-government and government organisations, or cross-industry collaboration.

Paragraph 6(3)(i) provides that an example of a reasonable step service providers could take in meeting the expectations set out in subsections 6(1), 6(2) and 6(2A) is 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.

Hate speech is communication or conduct 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 or conduct which expresses hate against a person or group of people. Expressions of hate against a person or group of people can be 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. This definition is non-exhaustive and is intended to provide broad guidance on what

hate speech can include. Services may vary in how they define hate speech or hateful conduct in their terms, policies or standards of conduct.

Paragraph 6(3)(j) provides that an example of a reasonable step service providers could take in meeting the expectations set out in subsections 6(1), 6(2) and 6(2A) is preparing and publishing regular transparency reports that outline the steps the service is taking to protect Australians online, including information regarding:

- the safety tools and processes deployed by the services
- metrics on the prevalence of harms on the service
- metrics on the effectiveness of the service's safety features
- a service's responsiveness to reports and complaints, and how the service is enforcing its terms of use, policies and procedures and standards of conduct mentioned in section 14.

Subsection 6(5) is an additional expectation, and provides that the provider of a service will take reasonable steps to make available controls that give end-users the choice and autonomy to support safe online interactions.

Subsection 6(6) provides examples of reasonable steps that could be taken for the purposes of subsection 6(5), and include:

- Making available "blocking" and "muting" controls for end-users (paragraph 6(6)(a)). 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, including things like content and content tags.
- Making available opt-in and opt-out measures regarding the types of content that end-users can receive (paragraph 6(6)(b)). 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 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, and types of content, they do not want to see, and can be particularly helpful in the context of empowering users to control the content provided through recommender systems (see also paragraph 8B(3)(d) under Item 7).
- Enabling end-users to make changes to their privacy and safety settings (paragraph 6(6)(c)). Providing users with autonomy over their privacy and safety settings is critical to allow end-users choice to engage with a service in a manner that suits their particular preferences, context and concerns.

The examples mentioned above are not an exhaustive list of the user empowerment controls that service providers could employ.

Item 6 repeals and replaces paragraph 8(2)(a) in the Determination, to replace "systematic" with "systemic". The word "systemic" has been substituted to better reflect the intention of this paragraph – that is, making clear that subsection 8(1) does not require online services to implement or build a whole-of system weakness or vulnerability into a form of encrypted service that could undermine the integrity of the encryption services used by online services.

Item 7 inserts sections 8A and 8B after section 8. The sections include additional expectations and examples of reasonable steps regarding the use by services of generative artificial intelligence (generative AI) and recommender systems. The expectations at subsections 8A(1), 8A(2), 8B(1) and 8B(2) are intended to reflect the expectations at subsections 6(1) and 6(2) in the determination, applying to these technologies the principles of safety by design and proactive minimisation of unlawful and harmful material or activity. The inclusion of specific sections regarding generative AI and recommender systems is not intended to imply that these (or other technologies) were previously outside of the scope of the Determination, but recognises the increased risks of such technologies in adversely affecting online safety, e.g. by enabling or amplifying the provision of unlawful or harmful material on a service.

Section 8A provides two additional expectations (subsections 8A(1) and 8A(2)) and examples of reasonable steps (subsection 8A(3)) regarding generative AI capabilities. Generative AI is distinguished from other applications of artificial intelligence by its capacity to generate novel material, such as text, images, videos, audio, or a combination of these.

Subsection 8A(1) is an additional expectation. It provides that if a service uses or enables the use of generative AI 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 generative AI capabilities on the service. This expectation provides that user safety must be considered, safety features incorporated, and safety risks minimised at all stages during the life cycle of a service's generative AI product or capability. As this expectation applies to all stages of the development and deployment cycle or 'stack' of a product or capability, it applies to all relevant electronic, designated internet, and social media services involved in the development and deployment of generative AI. Each service in this cycle has a role in ensuring that the final product made available to end-users promotes user safety. However, where a service sits in the cycle can impact what steps it can take to protect end-users. Whether a service provider is taking 'reasonable steps' to meet this expectation will depend on the matters within its control.

Subsection 8A(2) is an additional expectation. It provides that if a service uses or enables the use of generative AI capabilities, the provider of the service will take reasonable steps to proactively minimise the extent to which generative AI capabilities may be used to produce material or facilitate activity that is unlawful or harmful. This expectation covers material such as the production of non-consensual 'deepfake' intimate images or videos, class 1 material (such as child sexual exploitation or abuse material and terrorist and violent extremist material), or the generation of images, video, audio or text to facilitate cyber abuse or hate speech.

Subsection 8A(3) provides examples of reasonable steps that could be taken for the purposes of subsections 8A(1) and 8A(2), which include:

- 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 AI capabilities (paragraph 8A(3)(a)). This step aligns with the example of a reasonable step contained in paragraph 6(3)(e), but applies specifically to the design, deployment and post-deployment of the generative AI capabilities. This means that service providers

should take steps to identify how such capabilities are working in practice to minimise unlawful and harmful material and activity. Assessments could identify risks, harms, and gaps in safeguards of generative AI capabilities so that providers can then take appropriate steps to mitigate them.

- Providing educational or explanatory tools (including when new features are integrated) to end-users that promote understanding of generative AI capabilities on the service and any risks associated with the capabilities (paragraph 8A(3)(b)). These materials should be updated as new features are integrated into the capability. These measures will enable end-users to make informed decisions about engaging with generative AI capabilities on a service and to have greater awareness about any risks associated with material that is generated.
- Ensuring, to the extent reasonably practicable, that training materials for generative AI capabilities and models do not contain unlawful or harmful material (paragraph 8A(3)(c)). Service providers that develop generative AI capabilities must make important decisions regarding the input data that is used to train their 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, and non-consensual intimate images of individuals. This can potentially lead to the generation of unlawful or harmful material by the capability. To the extent reasonably practicable, service providers should proactively remove such content from training material so as to mitigate the risk of online safety risks emerging later in the lifecycle of the generative AI capability.
- Ensuring, to the extent reasonably practicable, that generative AI capabilities can detect and prevent the execution of prompts that are associated with unlawful or harmful material (paragraph 8A(3)(d)). Where possible services should endeavour to detect, and prevent the execution of, prompts that are clearly intended to produce unlawful or harmful material. However, measures such as educative prompts or nudges can also provide an opportunity for end-users to reconsider their engagement with the generative AI capability and curb misuse through clear signalling from the service that the user's engagement is not appropriate. As with paragraph 8A(3)(c), these preventive efforts should be ongoing throughout the lifecycle of the generative AI capability.

Section 8B provides two additional expectations, subsection 8B(1) and 8B(2), and examples of reasonable steps, subsection 8B(3), regarding recommender systems. Recommender systems are systems that prioritise content or make personalised 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. For the purposes of this section, recommender systems include not only systems that recommend material, but also systems which make recommendations of other users, accounts or profiles to follow, or which recommend a user's account or profile to others.

Subsection 8B(1) provides that if a service uses recommender systems, the provider of the service will take reasonable steps to consider end-user safety and incorporate safety measures into the design, implementation and maintenance of recommender systems on the service. As with subsection 8A(1) above, this expectation provides that user safety must be considered, safety features incorporated, and safety risks minimised at all stages during the development and operation of a recommender system.

Subsection 8B(2) provides that if a service uses recommender systems, the provider of the service will take reasonable steps to proactively minimise the extent to which recommender systems amplify material or activity on the service that is unlawful or harmful. The design and operation of recommender systems help to determine what material is promoted to end-users, and has the potential to amplify illegal or harmful content. Services are responsible for the algorithms that recommender systems are founded on and should consider and take steps to address the potential for virality and amplification of harmful material in the design and operation of recommender systems.

Subsection 8B(3) provides examples of reasonable steps that could be taken for the purposes of subsections 8B(1) and 8B(2), and include:

- 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 (paragraph 8B(3)(a)). Consistent with paragraph 8A(3)(a), this step is intended to ensure service providers identify gaps in the safety associated with recommender systems so as to appropriately mitigate them.
- Providing educational or explanatory tools to end-users that promote understanding of recommender systems on the service, their objectives, and any risks associated with such systems (including when new features are integrated) (paragraph 8B(3)(b)). Consistent with paragraph 8A(3)(a), this step is intended 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.
- 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 (paragraph 8B(3)(c)). Allowing users to raise concerns that a recommender system is promoting unlawful or harmful material or activity (in addition to merely allowing them to complain about the material or activity itself) would assist service providers in identifying underlying risks in the operation of their recommender systems. Services could then take more effective steps to prevent the provision of such material or activity on the service in the future.
- Where technically feasible, enabling end-users to opt-out of receiving recommended content, or providing alternative curation options (paragraph 8B(3)(d)). Providing users with greater control over how a service promotes material to them, and the kinds of material promoted, may better allow them to avoid material that may be particularly harmful for them.

Item 8 repeals and replaces paragraph 9(2)(a) to include reference to having proactive processes to prevent the same person from repeatedly using anonymous (or pseudonymous) accounts to post material, or 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. Where feasible, and consistent with the principle of anonymity, services should 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 – including those who have been banned or suspended from the service.

Item 9 repeals and replaces subsection 10(1) to clarify that the additional expectation requires providers of a service to take reasonable steps to consult and cooperate between all relevant services provided by that provider, in addition to providers of other services. This amendment clarifies that the additional expectation is intended to apply to consultation and cooperation between services belonging to the same provider, such as where a parent company owns and operates more than one social media, relevant electronic or designated internet service.

Item 10 repeals and replaces paragraphs 10(2)(a) and 10(2)(b), which are examples of reasonable steps for the purposes of subsection 10(1). They have been amended to align with the clarifying amendment to subsection 10(1) in Item 9.

Item 11 repeals and replaces paragraph 12(2)(a) and provides that age assurance mechanisms implemented in accordance with the paragraph should be appropriate age assurance mechanisms. The inclusion of the word 'appropriate' clarifies that age assurance mechanisms to prevent children's access to class 2 material should be calibrated according to factors such as:

- The effectiveness of the age assurance mechanisms;
- The extent to which class 2 material is provided on the service; and
- The likelihood of children accessing the material on the service.

This means that in some instances, asking users to self-declare 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 that can be accessed on the service.

Item 12 inserts paragraph 12(2)(c) after paragraph 12(2)(b). This paragraph provides as a new reasonable step for the purposes of subsection 12(1) that service providers continually seek to implement improved technologies and processes for preventing access by children to class 2 material. Technologies and processes for age assurance and age appropriate design are continuously evolving and service providers should continue to seek ways to improve upon or refine their existing approaches. They can do this through developing their own improved technologies and processes or through supporting or sourcing those developed by others for use on their services.

Item 13 inserts subsection 14(1A) after subsection 14(1) of the Determination. It provides a new additional expectation that service providers take reasonable steps, including proactive steps, to detect breaches of terms of use, policies and procedures, and standards of conduct. This sets the expectation that service providers should not rely solely on user reports and complaints to identify and address such material and activity that breaches its rules of conduct in relation to online safety.

The Amendment Determination does not specify examples of reasonable steps for this expectation. However, reasonable steps that service providers could take include:

- Technological interventions that detect activity or material 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 terrorist and violent extremist material, AI classifiers that identify new material that could be unlawful or harmful which then get prioritised for human review, and technologies such as language or text analysis.
- Remaining alert and detecting ongoing patterns of unlawful and harmful activity that breaches their terms of use, policies and procedures and standards of conduct once such conduct has been reported by others.

Item 14 repeals and replaces subsection 14(2) to:

- Clarify that service providers are expected to enforce, in addition to terms of use, the policies and procedures and standards of conduct mentioned in subsection 14(1). Where "terms of use" is referred to in the Determination, it should be interpreted broadly to include all of these, as well as the core expectations set out in the Act at subsections 13(1) and 20(1) where the shorthand usage of "terms of use" has not been amended.
- Specify that reasonable steps include proactive steps. In enforcing terms, policies and standards of conduct against all accounts held or created by an end-user, services should not rely solely on user reports and complaints, but should take proactive steps, such as through detecting when users evade enforcement through using new or alternative accounts.

Item 15 inserts subsections 14(3), 14(4) and 14(5) after subsection 14(2).

Subsection 14(3) is an additional expectation. Paragraph 14(3)(a) provides that service providers will review and respond to reports and complaints regarding material and activity specified in section 13 regarding specified material, and those mentioned in section 15 regarding breaches of a service's terms of use and (where applicable), breaches of policies and procedures and standard of conduct. Paragraph 14(3)(b) provides that service providers will take reasonable steps to give feedback on the action taken, within a reasonable period of time. It is important that services address reports of harm on their service as quickly as is reasonably possible, so as to minimise the potential harm, and to provide feedback to the complainant on the outcome of their report. Taking timely action and informing users on decisions taken is not merely important for the effectiveness and transparency of a service's moderation policy, but will assist users who may wish to subsequently report material to the Commissioner.

Examples of providing feedback on action taken by the service could include information regarding:

- actions taken in responding to the report (including a decision to take no action); and
- the reasons as to why certain action was taken or no action was taken. This is particularly important where a decision has been taken to take no further action on the reported activity or material, or to take alternative action to that which was requested.

Services are expected to take reasonable steps to provide feedback on actions taken. Services may receive a high volume of reports, with some being potentially vexatious or without merit. The efficiency and effectiveness of moderation may be impaired if detailed feedback is expected for every complaint. The feasibility of providing feedback may depend on factors such as the size of a service, the volume of complaints, and the sophistication of the service's

systems (such as the capacity for automating feedback). Where feasible, however, feedback should be sufficient to inform a complainant of where a service's response stands with respect to a service's policies. Clear and transparent feedback in this manner may in fact discourage a vexatious complainant from continuing to make similar reports.

Subsection 14(4) provides guidance about the factors that service providers should consider when determining what is a reasonable period of time to review, respond, and provide feedback in relation to the complaint or report. The nature and impact of the harm that is the subject of the report or complaint would generally be the primary consideration for determining a reasonable period of time (paragraph 14(4)(a)). However, the complexity of the report or complaint may also affect the time required to address a complaint (paragraph 14(4)(b)). Some material may need to be actioned immediately to prevent ongoing harm, whereas other reports and complaints may require a more nuanced analysis of the context and circumstances of the relevant conduct. The timely resolution of reports and complaints can 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. Other relevant matters a service must have regard to could include referring to any guidance material that is made available by the Commissioner (paragraph 14(4)(c)).

Subsection 14(5) provides a definition of "review" and "respond" for the purposes of paragraph 14(3)(a) where:

- review refers to a service's procedure for considering reports and complaints when they are made; and
- respond refers to the decision itself and implementing the decision following the review procedure, such as the removal of content, reporting of illegal content to authorities, banning or suspension of the accounts of offending users, deprioritisation of material on a recommender system, or no action.

Item 16 repeals and replaces subsection 15(2) to extend the reference to terms of use to explicitly refer to policies and procedures and standards of conduct mentioned in section 14.

Item 17 inserts a new additional expectation at subsection 20(5). It provides that the Commissioner may, by written notice to a provider of a service, request that the provider give to the Commissioner a report on the number of active end-users of the service in Australia, disaggregated into active end-users who are children and those who are adult. The provider will be expected to comply with the request within 30 days after the notice of the request is given. Information about the number of Australian end-users 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 Australian adults and children. This will improve the Commissioner's capacity to support Australians by identifying where Australians are most likely to need support, and support the Commissioner's regulatory functions.

Item 18 inserts a note after subsection 21(1) about the expectation to have a designated contact point. The explanatory note confirms that service providers are expected to comply with the expectation (at section 21 of the Determination) to provide the Commissioner with a designated contact point for the purposes of the Act, irrespective of whether the service provider has staff physically located in Australia. This is also intended to clarify that where a service provider with end-users in Australia previously had staff physically located in

Australia, but decides to cut staff or move staff overseas, or has never had staff physically located in Australia, it is still expected to provide the Commissioner with a designated contact point.

Attachment B

Statement of Compatibility with Human Rights

Prepared in accordance with Part 3 of the Human Rights (Parliamentary Scrutiny) Act 2011.

Online Safety (Basic Online Safety Expectations) Amendment Determination 2024

The Online Safety (Basic Online Safety Expectations) Amendment Determination 2024 (the Amendment Determination) is compatible with the human rights and freedoms recognised or declared in the international instruments listed in Section 3 of the *Human Rights (Parliamentary Scrutiny) Act 2011*.

Overview of the Amendment Determination

The Amendment Determination was issued by the Australian Government under Section 45 of the *Online Safety Act 2021* (the Act).

The Amendment Determination amends the *Online Safety (Basic Online Safety Expectations) Determination 2022* (the Determination). Under the Determination, online service providers, including social media services, relevant electronic services and designated internet services are expected to take steps to protect Australians from unlawful and harmful material and activity that falls within the remit of the enabling legislation the *Online Safety Act 2021* (the Act), or impedes the online safety of Australians, and to report on these steps as required.

Amendments to the Determination address online harms that have grown or emerged since it was made, including those associated with generative artificial intelligence and recommender systems, as well as the need to strengthen protections for children online. Amendments also seek to improve the scheme's overall operation and address gaps identified through the eSafety Commissioner's (the Commissioner's) exercise of existing powers.

Therefore, in relation to human rights, the amendments have the overall effect of strengthening the support of these rights provided under the original Determination.

Human rights implications

The amendments made under the Amendment Determination engage the same principal human rights engaged by the original Determination, which are also engaged by the Act. These are:

- The right to freedom of expression primarily contained in Article 19 of the *International Covenant on Civil and Political Rights* (the ICCPR), and also referred to in Articles 12 and 13 of the *Convention on the Rights of the Child* (the CROC) and Article 21 of the *Convention on the Rights of Persons with Disabilities* (the CRPD);
- The prohibition on interference with privacy and attacks on reputation primarily contained in Article 17 of the ICCPR, and also referred to in Article 16 of the CROC, and Article 22 of the CRPD;
- The right to protection from exploitation, violence and abuse primarily contained in Article 20(2) of the ICCPR, and also referred to in Article 19(1) of the CROC and Article 16(1) of the CRPD; and

- The best interests of the child, contained in Article 3(1) of the CROC.

In addition, through increased focus on online protections for children, the Amendment Determination engages additional articles of the CROC:

- The right to survival and development contained in Article 6(2); and
- The right to protection from information and material injurious to his or her well-being contained in article 17.

These rights, and how they are engaged in the Amendment Determination, are discussed below.

Freedom of expression

Rights relating to freedom of expression are recognised and protected by Article 19 of the ICCPR and, in relation to children specifically, by Articles 12 and 13 of the CROC.

Paragraph 1 of Article 19 of the ICCPR recognises that everyone shall have the right to hold opinions without interference. Paragraph 2 states that everyone shall have the right to freedom of expression. Paragraph 3 recognises that the exercise of the rights provided for in Paragraph 2 may be subject to certain restrictions. Paragraph 3 of that article as well as Paragraph 2 of Article 13 of the of the CROC limits the types of restrictions that may be imposed. Restrictions are as provided for by law and are necessary either in respect of the rights or reputations of others or for the protection of national security, public order, health and morals.

The Amendment Determination adds or strengthens a number of provisions that support the right to hold opinions without interference and in so doing support freedom of expression, by addressing negative behaviour of end users towards other end users and addressing threats to safe use of online spaces more broadly. Relevant measures include:

- Clarifying the additional expectation that providers cooperate to promote safe use to specify that information on all services should be shared, and that all services provided by a provider should work together, to detect high-volume or 'pile on' attacks against individuals or small groups;
- Expanding an additional expectation regarding setting and enforcement of terms of use to provide that services should proactively seek to detect breaches, including in relation to end user conduct; and
- A new additional expectation that complaints will be responded to in a reasonable timeframe and feedback provided on action taken.

The Amendment Determination also strengthens certain necessary restrictions on freedom of expression. In addition to those above, which in effect curtail the freedom of expression of end users engaged in harmful conduct, several other provisions which seek to restrict access to or remove certain content deemed to be harmful also necessarily restrict freedom of expression. Amendments include:

- New examples of reasonable steps to ensure safe use and proactively minimise unlawful or harmful material, such as risk assessment, mitigation and review of services, resourcing and processes for timely responses to reports, investment in

prevention and detection of harmful material or activity including hate speech, and regular transparency reports;

- An additional expectation to take reasonable steps to ensure that the best interests of the child are a primary consideration in the design and operation of a service likely to be accessed by children;
- An additional expectation to make available controls that give end-users the choice and autonomy to support safe online interactions, with related examples of reasonable steps relating to blocking, muting, opting in or out of content and ability for users to change privacy and safety settings;
- An additional expectation to proactively minimise the extent to which generative intelligence capabilities may be used to produce material or facilitate activity that is unlawful or harmful; with associated examples of reasonable steps;
- An additional expectation to proactively minimise the extent to which recommender systems amplify material or activity on the service that is unlawful or harmful, with associated examples of a reasonable steps;
- Amending an additional expectation to include proactive processes as an example of processes to precent the use of anonymous accounts to post material, or engage in activity, that is unlawful or harmful;
- Adding against the core expectation to ensure technological or other measures are in effect to prevent access by children to class 2 material (as defined under the Online Content Scheme in the Act), an example of a reasonable step that services continually seek to develop, support or source, and implement improved technologies and processes for preventing access by children to class 2 material; and
- An additional expectation that services review and respond to reports within a reasonable period of time and an associated example of a reasonable step.

The Amendment Determination allows service providers to determine how unlawful or harmful material or activity will be dealt with, in a way that is consistent with achieving the intended policy outcome of responding to this material or activity. In particular, the example of a reasonable step regarding hate speech only engages the right to freedom of expression insofar as online services themselves have policies restricting hate speech, and does not create a positive expectation or obligation that they have such policies.

In any circumstances, the Amendment Determination does not prescribe the manner in which this harmful or unlawful material must be handled, nor are the expectations mandatory for services.

The right to protection from exploitation, violence and abuse

The right to protection from exploitation, violence and abuse is primarily contained in Article 20(2) of the ICCPR and other related conventions. The ICCPR and related conventions requires Australia to take measures to protect persons from exploitation, violence and abuse.

The Amendment Determination strengthens and adds provisions that engage these rights and obligations, noting that most expectations under the original Determination engage broadly with promoting safety of end users, including protection from and recourse in relation to exploitation, violence and abuse.

Specific provisions in the Amendment Determination which engage these rights and obligations include:

- A new example of a reasonable step to detect and address hate speech;
- A new additional expectation to make available controls that give end-users the choice and autonomy to support safe online interactions;
- New additional expectations to consider end-user safety and incorporate safety measures in design, implementation and maintenance of generative artificial intelligence capabilities; and 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;
- New additional expectations to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of recommender systems; and to proactively minimise the extent to which recommender systems amplify material or activity on the service that is unlawful or harmful;
- Clarifying the additional expectation that providers cooperate to promote safe use to specify that information on all services should be shared, and that services provided by a provider should work together, to detect high-volume or 'pile on' attacks against individuals or small groups;
- Expanding an additional expectation regarding setting and enforcement of terms of use to provide that services should proactively seek to detect breaches, including in relation to end user conduct; and
- A new additional expectation that complaints will be responded to in a reasonable timeframe and provide feedback on action taken.

Prohibition on interference with privacy and attacks on reputation

Paragraph 1 of Article 16 of the CROC recognises, among other things, the right of a child not to be subjected to unlawful interference with privacy or unlawful attacks on their honour and reputation. Paragraph 2 recognises that children have the right to the protection of the law against such interference or attacks. Article 22 of the CRPD and Article 17 of the ICCPR contain similar rights with regards to persons with disabilities and all persons respectively.

The Amendment Determination engages with the right to privacy, by strengthening privacy provisions for all users. In relation to the new additional expectation regarding user controls to support safe online interactions noted above, it includes as an example of a reasonable step that end users have the ability to make changes to privacy and safety settings. It also clarifies an existing example of a reasonable step to ensure safe use, that default privacy and safety settings are robust and set to most restrictive on services likely to be accessed by children.

The Amendment Determination also includes the following new provisions regarding attacks on reputation for all users:

- Additional expectations to consider end-user safety and incorporate safety measures in design, implementation and maintenance of generative artificial intelligence capabilities; and to proactively minimise the extent to which generative intelligence capabilities may be used to produce material or facilitate activity that is unlawful or harmful, and associated examples of reasonable steps; and

- Strengthening the additional expectation relating to cooperation between service providers to include cooperation between all services by a provider to promote safe use; and examples of reasonable steps that relate specifically to 'pile on' attacks on individuals or groups.

The best interests of the child and related articles

Article 3(1) of the CROC provides that in all actions concerning children, the best interests of the child shall be the primary consideration. The principle requires legislative, administrative and judicial bodies to take active measures to protect children's rights, promote their wellbeing and consider how children's rights and interests are or will be affected by their decisions and actions.

Several provisions under the Amendment Determination engage this principle explicitly. These include:

- A new additional expectation that providers of services 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 likely to be accessed by children;
- Two new child-specific examples of reasonable steps that could be taken by service providers to meet expectations, including clarifying that default privacy and safety settings are robust and set to most restrictive on services likely to be accessed by children and specifying child safety risk assessments as a type of risk assessment to be undertaken and responded to; and
- New examples of reasonable steps to meet expectations, one to meet the core expectation of preventing access by children to class 2 material and the other as an example of a reasonable step to meet a new additional expectation to provide information about the number of active end users to the Commissioner, in which separate figures would be provided for adults and children.

In addition, many amendments under the Amendment Determination, while not expressly targeted at children, seek to improve responsiveness of service providers to matters which are known to significantly impact children and young people, including in relation to controls to support safe online interactions, the deployment of recommender systems, cooperation with other service providers in the event of 'pile-on' attacks on individuals and proactive detection by services of breaches of their terms of use.

Further, measures relating to the deployment of recommender systems in particular, which risk funnelling distorted, incorrect and harmful information to children, engage article 17 of the CROC, concerning the child's right to protection from information and material injurious to his or her well-being.

More broadly, by strengthening protections for children from online harms, the Amendment Determination engages with the right of a child to survival and development contained in article 6 of the CROC.

Conclusion

The Amendment Determination is compatible with the human rights and freedoms recognised or declared in the international instruments listed in Section 3 of the *Human Rights (Parliamentary Scrutiny) Act 2011*. The measures in the Amendment Determination promote the right to protection from exploitation, violence and abuse and the best interests of the child.

To the extent to which the measures in the Amendment Determination may engage with the right to freedom of expression and the prohibition on interference with privacy and attacks on reputation, any limitation is reasonable, necessary and proportionate to the goal of promoting and improving transparency and accountability of online services and improving online safety for Australians.