Discussion paper

Draft Online Safety (Relevant Electronic Services – Class 1A and 1B Material) Industry Standard 2024

and

Draft Online Safety (Designated Internet Services – Class 1A and 1B Material) Industry Standard 2024

November 2023

Australian Government | eSafetyCommissioner

eSafety.gov.au

Contents

Introduction

The Online Safety Act 2021 (Cth) (**the Act**) commenced on 23 January 2022. The Act provides for the establishment of new mandatory industry codes and standards for eight sections of the online industry to regulate the most harmful types of online content.

This discussion paper is about the establishment of two new industry standards under the Act: Online Safety (Relevant Electronic Services – Class 1A and 1B Material) Industry Standard 2024 (**the Relevant Electronic Services Standard**) and the Online Safety (Designated Internet Services Standard – Class 1A and 1B Material) Industry Standard 2024 (**the Designated Internet Services Standard**).

We are seeking the views of the online industry, advocacy groups, the general public and other interested stakeholders on the exposure drafts of the Relevant Electronic Services Standard and the Designated Internet Services Standard.

This discussion paper:

- sets out the legislative framework for the Relevant Electronic Service Standard and the Designated Internet Services Standard (**the Standards**)
- outlines eSafety's overarching approach to the Standards
- sets out consultation questions on key elements of the Standards
- explains the consultation process.

To assist stakeholders, we have provided a Fact Sheet for each draft Standard which sets out eSafety's position in response to key questions. This discussion paper should also be read alongside the draft Standards, to assist with understanding the key questions.

Background

The legal framework for the Standards

Part 9, Division 7 of the Act provides for the establishment of industry codes and industry standards. The Act provides for industry bodies to develop codes to regulate harmful online material, and for eSafety to register the codes if they meet the statutory requirements. The codes become enforceable when registered by the eSafety Commissioner. If a draft code does not meet the statutory requirements, the

eSafety Commissioner is able to determine a standard for that section of the online industry.

The current set of codes and standards deals with class 1A and 1B online material:

- class 1A - child sexual exploitation material, pro-terror material, and extreme crime and violence material
- class 1B - crime and violence material, and drug-related material.

These types of material are subcategories of class 1 material under the Act, which is material that has been or would be refused classification (**RC**) under the Classification Act. Serious harms are associated with class 1A and 1B material whenever it is produced, distributed or consumed.

Future industry codes or industry standards will be developed to address class 2 (restricted) material under the Act, such as online pornography.

The Relevant Electronic Service Standard will apply to the section of the online industry that provides relevant electronic services. This is a wide category of services that enable end-users to communicate with each other online. A 'relevant electronic service' is defined in the Act as a service that enables end-users to communicate with one another by email, instant messaging, short message services (commonly known as '**SMS**'), multimedia message services (commonly known as '**MMS**') or chat services, as well as services that enable end-users to play online games with each other. A relevant electronic service also includes an online dating service, as outlined by eSafety in its September 2021 position paper.[1]

The Designated Internet Services Standard will apply to the section of the online industry that provides designated internet services. A 'designated internet service' is defined in the Act as a service which allows end-users to access material using an internet carriage service or which delivers material to persons who have equipment appropriate for receiving that material, where the delivery is by means of an internet carriage service. This is a very broad category of services that includes many apps and websites, as well as online storage services which are used by end-users to upload, store and manage their files including photos and other media. Designated internet services exclude social media services, relevant electronic services and other identified services.

[1] eSafety, Development of industry codes under the Online Safety Act – Position Paper.

Why these industry Standards are necessary

The Designated Internet Services Standard and Relevant Electronic Services Standard are necessary to meet the objectives of the Online Content Scheme under Part 9 of the Act, that is to put in place codes or standards for each sector of the online industry.

Earlier in 2023, the eSafety Commissioner registered industry codes dealing with class 1A and class 1B material for the six industry sectors. The draft Relevant Electronic Services Code and Designated Internet Services Code were not registered as they were found not to contain appropriate community safeguards for end-users in Australia in relation to matters which are of substantial relevance to the community.

As a consequence, it was necessary for the eSafety Commissioner to determine industry standards for relevant electronic services and designated internet services to ensure that each sector of the online industry offers meaningful protections to end-users in Australia in respect of class 1A and class 1B material.

The registered industry codes cover the following industry sectors:

- social media services
- internet carriage services (also known as internet service providers)
- equipment providers
- app distribution services
- hosting services
- internet search engine services.

Our approach to the Standards

The Relevant Electronic Services Standard and Designated Internet Services Standard will operate alongside the registered industry codes and impose a set of mandatory compliance measures, requiring service providers to:

- take proactive steps to create and maintain a safe online environment
- empower end-users in Australia to manage access and exposure to class 1A and class 1B material
- strengthen transparency of, and accountability for, class 1A and class 1B material on their services.

Consistent with the registered industry codes, the draft Standards adopt an outcomes- and risk-based approach. The requirements proposed in the Standards are proportionate to the risk a service presents in respect of class 1A and 1B material. The requirements are also outcomes-based, in that they set out what they are intended to achieve while providing flexibility in how those outcomes are to be achieved. This approach recognises that:

- different services and technologies may have different risk profiles
- compliance measures should be proportionate to the level of risk associated with a particular service and to the size and capacity of the service provider responsible for that online activity or service
- compliance measures should be flexible, to enable effective implementation, recognising the differences between unique services, and to adapt to changes in technology and in the risk environment.

In creating the draft Standards, eSafety has sought to build on the extensive work of industry bodies in developing and consulting on the draft Relevant Electronic Services Code and Designated Internet Services Code. Where appropriate, eSafety has used elements of the draft codes as an initial basis for the Standards requirements, while addressing the Commissioner's concerns with the draft codes and also developing new measures to address risks posed by generative artificial intelligence (**generative AI**).[2]

[2] For background information on generative AI and the online safety risks associated with this technology, see eSafety's Tech Trends position statement on generative AI.

Public consultation on the draft Standards

Timeline for consultation

The following timeline outlines the key steps the eSafety Commissioner will take to determine the Standards.

Table 1: Timeline for consultation and key steps in determining the Standards

Date	Key steps
20 November 2023	Public consultation on the two draft Standards opens (for 31 days in accordance with section 148 of the Act)
21 December 2023	Public consultation closes
January–February 2024	Review public submissions and amend Standards as required
End of March 2024	Planned Lodgement of Standards and explanatory statements with the Office of Parliamentary Counsel

Submissions

The eSafety Commissioner invites submissions from industry and the public on the draft Standards. Views expressed through this consultation will assist the eSafety Commissioner to finalise the draft Relevant Electronic Services Standard and draft Designated Internet Services Standard under the Online Safety Act 2021.

How to make a submission

We welcome written submissions, which can be made via:

Email: submissions@esafety.gov.au

Mail: Executive Manager, Industry Regulation and Legal Services, Office of the eSafety Commissioner PO Box Q500, Queen Victoria Building NSW 1230

Each submission should be accompanied by:

- the name of the individual or organisation making the submission
- a name for publication purposes (this can be the name of the individual or organisation, or a pseudonym, or ‘anonymous’)

- contact details (such as a telephone number, postal address or email address)

A submitter may claim confidentiality over their name or contact details (see 'Publication of submissions' and 'Confidentiality').

The closing date for submissions is **21 December 2023.**

Please contact us via email at codes@eSafety.gov.au if you require a time extension to make a submission. eSafety recognises that some of the proposals such as those specific to generative AI technology were not raised in the previous industry codes consultation process and that more time may be required to respond to those proposals.

If you require an alternative method of making a submission, please contact eSafety.

Discussion questions

eSafety is seeking views on the effectiveness of the draft Relevant Electronic Services Standard and draft Designated Internet Services Standard in providing appropriate community safeguards for class 1A and class 1B material.

In preparing the draft Standards, eSafety has considered issues raised in industry's consultation during the codes development process for relevant electronic services and designated internet services. Some of these issues are addressed in our Fact Sheets.

Discussion questions have been provided in this paper to assist in focusing submissions. They are a guide only and not intended to limit the scope of submissions. Responses can be provided to all, any or none of the questions.

eSafety requests that submitters provide reasons to support any views expressed. Practical examples, research and other evidence are welcomed.

Publication of submissions

In the interests of transparency, eSafety intends to publish submissions we receive on our website, including any personal information in the submissions. As submissions will be made public, please ensure that you do not include any personal information that you do not want published.

Confidentiality

We prefer to receive submissions that are not claimed to be confidential. However, we accept that people may sometimes wish to provide information in confidence. In

these circumstances, we ask you to identify the material (including any personal information) over which confidentiality is claimed and provide a written explanation for the claim. eSafety will consider each confidentiality claim on a case-by-case basis. If we accept a claim, we will not publish the confidential information unless authorised or required by law to do so. We will notify people if we agree or disagree with the claim.

Privacy information

We collect personal information for the purpose of considering the issues raised in the discussion paper and to contribute to the transparency of the consultation process by clarifying, where appropriate, whose views are represented by a submission. We may also use your details to contact you regarding your submission. eSafety will not use the personal information collected for any other purpose, unless the submitter has provided their consent or eSafety is otherwise permitted to do so under the Privacy Act 1988 (Cth) (Privacy Act). For more information, please see our Privacy Policy and Collection Notification.

eSafety's Privacy Policy contains details about how an individual may access personal information about them that is held by eSafety and seek the correction of such information. It also explains how an individual may complain about a breach of the Privacy Act and how eSafety will deal with such a complaint.

Release of submissions

Any submissions provided to eSafety may be released under the Freedom of Information Act 1982 (Cth) (unless an exemption applies) or shared with other Australian Government agencies and certain other parties under Part 15 of the Act. eSafety may also be required to release submissions for other reasons including for the purpose of Parliamentary processes or where otherwise required by law (for example, under a court subpoena). While eSafety seeks to consult submitters of confidential information before that information is provided to another party, eSafety cannot guarantee that confidential information will not be released through these or other legal means.

Discussion questions – both Standards

Risk assessments are an important way to reduce the risks of class 1A and class 1B material being generated, posted, stored or distributed. The draft Standards propose that providers of certain services self-assess their risk to identify their risk tier and consequent legal obligations.

The draft Standards set out risk assessment obligations, including a requirement that certain service providers conduct a risk assessment within 6 months of the commencement of the applicable Standard or, for a new service, within 6 months before starting to provide the service to end-users in Australia. Service providers are also required to conduct a risk assessment if they propose to make a material change to the service.

Service providers that fall into a pre-assessed category or meet the requirements of a defined category of designated internet services are not required to conduct a risk assessment unless they make a material change which would increase the risk of class 1A material or class 1B material being accessed or generated by, or distributed to, end-users.

- **Pre-assessed categories** are exempt from risk assessments because they are deemed to have a particular risk tier.
- **Defined categories** are also exempt, as each category comprises providers whose risk profiles typically reflect the category of services they provide.

Question 1

Are the requirements for risk assessment in the draft Standards targeted at the right services and at the right points in a service's development journey? Are the risk factors appropriate?

The approach to minimum compliance measures

In setting out the proposed obligations, eSafety has taken into account a range of considerations, in particular:

- the importance of striking a balance between flexibility and enforceability, so service providers, eSafety and third parties have clarity about required outcomes

- the principle of risk-based, outcomes-based and technology neutral regulation so providers can implement measures that reflect the characteristics of their service and are responsive to rapidly shifting technologies
- ensuring obligations are meaningful as well as technically feasible, practical and – where appropriate – able to be deployed at scale
- the importance of human rights, including the right to freedom of expression, the right not to be subjected to arbitrary or unlawful interference with privacy, the right to protection from exploitation, violence and abuse, and the rights and best interests of children.

Question 2

Do the obligations on each relevant electronic service and designated internet service category appropriately reflect the above considerations? Are other considerations relevant?

Discussion questions – draft Relevant Electronic Services Standard

Applicability of the Standard

The Relevant Electronic Services Standard will apply to providers of relevant electronic services to end-users in Australia.[3]

Online service providers with multiple services are required to identify and comply with the industry code or standard that applies to each service.

Section 5 of the draft Relevant Electronic Services Standard provides that, where a single electronic service could potentially fall within the scope of the Relevant Electronic Services Standard and also an industry code or another standard under the Act, the service provider will only be required to comply with one code or one standard in relation to that specific service (until the relevant industry code or standard dealing with class 2 content comes into effect).

The applicable industry code or standard will be the one that the service's predominant functionality is most closely aligned with.

Question 3

Is the test in section 5 workable? Is further guidance required to assist providers to determine whether this standard, or another code or standard, applies to a particular online service?

Requirements to address 'known' material and 'new' material

eSafety recognises that there are differences in the threat to safety posed by 'new' material compared to known material and in the technology and systems available to identify such material.

'New' material refers to newly produced or recorded child sexual exploitation material or pro-terror material which has not been previously verified.

[3] Section 135(2)(c) of the OSA.

'Known' material refers to child sexual exploitation material or pro-terror material that has been previously verified, for example by a recognised child protection organisation or by an organisation with expertise in counter-terrorism.

New child sexual exploitation material is more likely to involve immediate safety risks to a child, for example where the child is being groomed and coerced into producing exploitative material. New child sexual exploitation material may be indicative of recent or ongoing sexual abuse of a child.

New pro-terror material is also likely to indicate more immediate risks to national security. New pro-terror material may have been produced very recently and could provide an important signal that an attack is underway now.

There are also significant harms associated with the distribution and consumption of known child sexual exploitation material and known pro-terror material.

The draft Standard places specific obligations on service providers in relation to these two types of material.

Detecting and removing known material

Sections 20 and 21 of the draft Relevant Electronic Services Standard set out requirements for providers of specified categories of relevant electronic services to proactively detect and remove known child sexual abuse material and known pro-terror material if it is technically feasible. Child sexual abuse material is a sub-set of child exploitation material, and is the focus of key provisions of the Relevant Electronic Services Standard.

Under the draft Relevant Electronic Services Standard, these requirements apply to closed communication relevant electronic services, including email services and private online messaging services, but exclude short messaging services (SMS) and multi-media messaging services (MMS). Closed communication relevant electronic services were not subject to requirements to detect and remove this material in the draft Relevant Electronic Services Code even though many email services and messaging services already detect known child sexual abuse material. The Relevant Electronic Service Standard ensures that this voluntary practice and the level of online safety it provides is recognised and made a minimum standard across industry.

eSafety recognises that some closed communication services, particularly in an end-to-end encrypted environment, may face technical limitations in detecting known child sexual abuse material and known pro-terror material. Therefore, this requirement only applies if it is technically feasible for the service to detect and

remove the material. It does not require encryption to be weakened or subverted. eSafety considers this is a more appropriate and flexible way of dealing with these limitations rather than having a separate category for end-to-end encrypted services.

While certain services may face technical limitations and challenges to detecting known child sexual abuse material and pro-terror material, there are still meaningful steps that providers can take to combat the misuse of their services for this material. Under the Relevant Electronic Services Standard, where it is not technically feasible for a service provider to deploy tools to automatically detect and remove known child sexual abuse material and pro-terror material on the service, the provider is required to take appropriate alternative action. At eSafety's request, the provider must specify where it is technically infeasible to comply, and the appropriate alternative actions taken. It will also be required to meet obligations to disrupt and deter the dissemination of known child sexual abuse material and pro-terror material. Further, service providers currently unable to meet the requirements in sections 20 and 21 may find that detection becomes feasible as new technologies are developed and tested.

Question 4

Is the technical feasibility exception in the obligation to detect and remove known child sexual abuse material and pro-terror material appropriate? How effective will this obligation be with this exception?

Relevant electronic service providers are not required to deploy a particular technology or process in order to comply with the requirement to proactively detect and remove the material under sections 20 and 21. It is open to a service provider to implement the approach most appropriate for their service.

Examples of systems, processes and technologies identified in the draft Relevant Electronic Services Standard for detecting, flagging and/or removing this material include hash matching technologies, machine learning and artificial intelligence tools that scan for known material. For end-to-end encrypted services, these detection technologies may be used on the parts of the services that are not encrypted, such as profile pictures, content in end-user reports (complaints) or group names. These technologies may be accompanied by other steps which disrupt and deter the distribution of child sexual abuse material and pro-terror material, such as those covered in the following paragraphs.

The Relevant Electronic Services Standard does not require service providers to design systematic vulnerabilities or weaknesses into end-to-end encrypted services.

Question 5

Are there other examples of systems, processes and technologies that can detect, flag and/or remove known child sexual abuse material and known pro-terror material at scale, which should be highlighted in the Standards or accompanying guidance?

Disrupting and deterring material

The draft Relevant Electronic Services Standard at section 22 requires certain relevant electronic service providers to take action to disrupt and deter end-users from using the service to solicit, create, post, or disseminate child sexual abuse material and pro-terror material. This requirement applies to both new generation and known material.

Similar to sections 20 and 21, this requirement does not mandate a particular approach or technology, so the service provider has the flexibility to implement systems and deploy technology that is appropriate to the characteristics of its service including the scale of the service. As set out in the draft Relevant Electronic Services Standard, examples of actions providers may take to disrupt and deter include using artificial intelligence or machine learning techniques (such as behavioural signals) to detect activity and remove accounts accessing or sharing such material.

Other examples that are not contained in the Relevant Electronic Services Standard itself, that could be included in the explanatory statement, include the following:

- Interventions that are targeted at preventing end-users from making this material available on the service, for example by acquiring and using off-platform information that can help identify and block the registration of potential end-users who have distributed child sexual abuse material and/or pro-terror materials in other environments. This could mean providers taking into account credible information published, provided or validated by another service, about significant threats related to that end-user, such as those related to child sexual exploitation and abuse or terrorism.
- Deploying safety tools that disrupt or deter the distribution of child sexual abuse material and/or pro-terror material. Examples include interstitial warnings, blurring or blocking content, or providing safety information to end-users.

This requirement also applies to gaming services with communication functionality. While this category of relevant electronic services was not included in the draft

Relevant Electronic Services Code in relation to similar requirements, eSafety considers this inclusion in the Standard appropriate given the risk of gaming services being used to generate and distribute harmful material, including first generation material. Research indicates that the communications features in gaming services, such as chat, voice calls and livestreams, are used to approach children and coerce them into creating child sexual abuse material.[4] Research also suggests that gaming services could be used to circulate pro-terror material, with some gamers encountering extremist content while playing multiplayer games.[5]

Question 6

Are there any limitations which would prevent certain service providers from deploying systems, processes and technologies to disrupt and deter child sexual abuse material and pro-terror material on relevant electronic services? If there are limitations, how might these be overcome?

Is it appropriate for this requirement to apply to gaming services with communication functionality?

Question 7

Are there other examples of systems, processes and technologies that can disrupt and deter the use of a relevant electronic service to solicit, generate, distribute or access child sexual abuse material and pro-terror material, which should be highlighted in the guidance?

Investment requirement for large relevant electronic services

Under section 23 of the Relevant Electronic Services Standard, services with more than 1 million monthly active users in Australia are required to have a program of investment and development to disrupt and deter child sexual abuse material and pro-terror material, including first generation material.

eSafety considers it appropriate to limit this requirement to large relevant electronic services, reflecting both the greater risk of dissemination on services with a large

[4] NSPCC 2020, The impact of the coronavirus pandemic on child welfare: online abuse (nspcc.org.uk); WeProtect Global Alliance 2021, Global-Threat-Assessment-2021.pdf (weprotect.org)

[5] NYU Stern Centre 2023, Gaming The System: How Extremists Exploit Gaming Sites And What Can Be Done To Counter Them — NYU Stern Center for Business and Human Rights

number of end-users as well as their likely access to greater resources and capabilities.

Question 8

Do you agree with the monthly active user threshold for the investment obligation? Are there other appropriate thresholds that should be considered to ensure the obligation is proportionate to the size and reach of the relevant electronic service?

Reporting and complaints mechanisms

Section 27 of the draft Relevant Electronic Services Standard requires that certain types of services must provide a reporting tool or mechanism to enable complaints to be reported to the service about material accessible on the service in breach of their terms of use relating to class 1A or class 1B materials.

This requirement applies to all services of the type identified, including closed communication relevant electronic services. While some services may not be able to definitively ascertain whether there is a breach of terms and conditions due to technical features of their service such as end-to-end encryption, having content reporting mechanisms is critical. For example, reporting mechanisms provide an avenue for individuals including a child who is being sexually exploited to inform the service provider, and it is important for the service to be alerted to the risk that its platform is being used in this way. eSafety's view is that all such service providers should be required to provide their end-users the ability to report a complaint. These reporting mechanisms are intended to work with provisions that require providers to take appropriate action to respond to reports of breaches of terms of use, such as warning the account holder who is in breach.

The draft Relevant Electronic Services Standard strengthens safety by requiring that complaint mechanisms are available within the service interface and do not require the complainant to report via a separate web page or email address. End-user reporting mechanisms must also enable end-users to specify the harm/s they are reporting.

Question 9

Are the end-user reporting requirements workable for the relevant service providers? Are there practical barriers to implementation?

Systems and processes for breaches

Sections 17 and 25 require all types of relevant electronic services to take appropriate action to engage with reports of class 1A and 1B materials and determine whether terms of use or policies have potentially been breached. Sections 16 and 24 require certain types of relevant electronic services to implement systems and processes to respond to breaches of terms of use relating to class 1A and class 1B material. 'Terms of use' includes community standards and acceptable use policies.

In the draft Relevant Electronic Services Standard these requirements apply to closed communication relevant electronic services, addressing a key concern with the draft Relevant Electronic Services Code.

While some services, including some closed communication relevant electronic services, may be limited in their ability to investigate complaints and be unable to definitively ascertain whether class 1A or class 1B material is on their service, the general requirement under the Relevant Electronic Services Standard is that appropriate action be taken in response to these breaches. Steps in the service provider's terms of use, if taken, could be appropriate action.

As set out in section 14, the terms of use are to allow the provider to take one of the following steps in the case of a breach:

- suspending the provision of the service to the end-user
- imposing specified restrictions on the use of the service by a specified end-user for a specified period
- terminating the agreement under which the service is provided.

Question 10

Should the requirement on certain relevant electronic services to respond to reports of class 1A and class 1B material on their service be limited to a requirement to take 'appropriate action'?

Compliance costs

eSafety recognises that the establishment of the Relevant Electronic Services Standard will likely impact a range of providers, including through compliance costs. Compliance costs include administrative costs, such as record keeping costs and costs of reporting to the regulator. Compliance costs also include substantive compliance costs, such as the costs of putting in place new systems and processes

to meet regulatory requirements including end-user reporting and complaints systems.

Some service providers may already have measures in place that would meet regulatory requirements or require relatively small changes to be compliant, meaning that their substantive compliance costs could potentially be minimal.

Question 11

What are your views on the likely compliance costs and, in particular, the impact of compliance costs on potential new entrants?

Additional information

We encourage you to provide further information that may not be covered in answers to the previous questions.

Question 12

Is there any additional information eSafety should consider in determining the Relevant Electronic Services Standard?

Discussion questions – draft Designated Internet Services Standard

Designated internet service categories

The Designated Internet Services Standard will apply to providers of designated internet services to end-users in Australia.[6]

Given this is a very broad range of online services, the draft Standard seeks to identify specific categories that have particular risk profiles and propose requirements which are proportionate and appropriate for each category. A broad risk assessment methodology is proposed for those designated internet services that do not fall within the defined categories.

Low risk services, for example a retail website which has minimal to no risk of being misused for class 1A or class 1B material, will have no obligations under the Standard.

Table 2: Defined and pre-assessed categories and risk tiers

Defined categories of designated internet services (DIS)	
• **High impact generative AI DIS**, for services with generative AI functionality to produce completely or partially synthetic high impact material[7] • **Machine learning model platform service**, for services distributing machine learning models • **Enterprise DIS**, for example websites for ordering commercial supplies, and services being deployed by other organisations for use by their end-users • **End-user managed hosting service**, for example cloud storage for files or photos	
Pre-assessed categories	
Tier 1	**High impact DIS**, for example 'gore' sites, pornography sites
Tier 2	DIS which are not Tier 1, Tier 3 or otherwise fall within a defined or pre-assessed category, for example a DIS which makes available professionally produced material and end-user generated material
Tier 3	**Classified DIS**, for example websites providing general entertainment that would be classified at least R18+ **General Purpose DIS**, for example news, educational and health websites

[6] Section 135(2)(c) of the OSA.
[7] Synthetic material which would be classified as X18+ or RC.

Question 13

Are the categories in Table 2 sufficiently clear for designated internet service providers to identify which category they fall within and therefore what obligations apply? What are the benefits and/or challenges of the categories as they are currently proposed?

Risk assessment

Designated internet services that do not fall into a defined category or pre-assessed category listed in Table 2 are required to carry out a risk assessment as set out in section 8. The risk assessment must be undertaken in accordance with a plan and methodology that takes into account the principle matters in section 9(5). These include factors like the predominant functionality of the service.

Question 14

Are the section 9(5) matters in the draft Standard appropriate and sufficiently clear to help designated internet service providers accurately self-assess which tier their service falls within?

Generative AI

Categories addressing generative AI

Generative AI offers significant opportunities across the economy and society. However, as outlined in eSafety's Generative AI Position Statement[8], there are risks that generative AI functionality could be misused to generate class 1A or class 1B material.

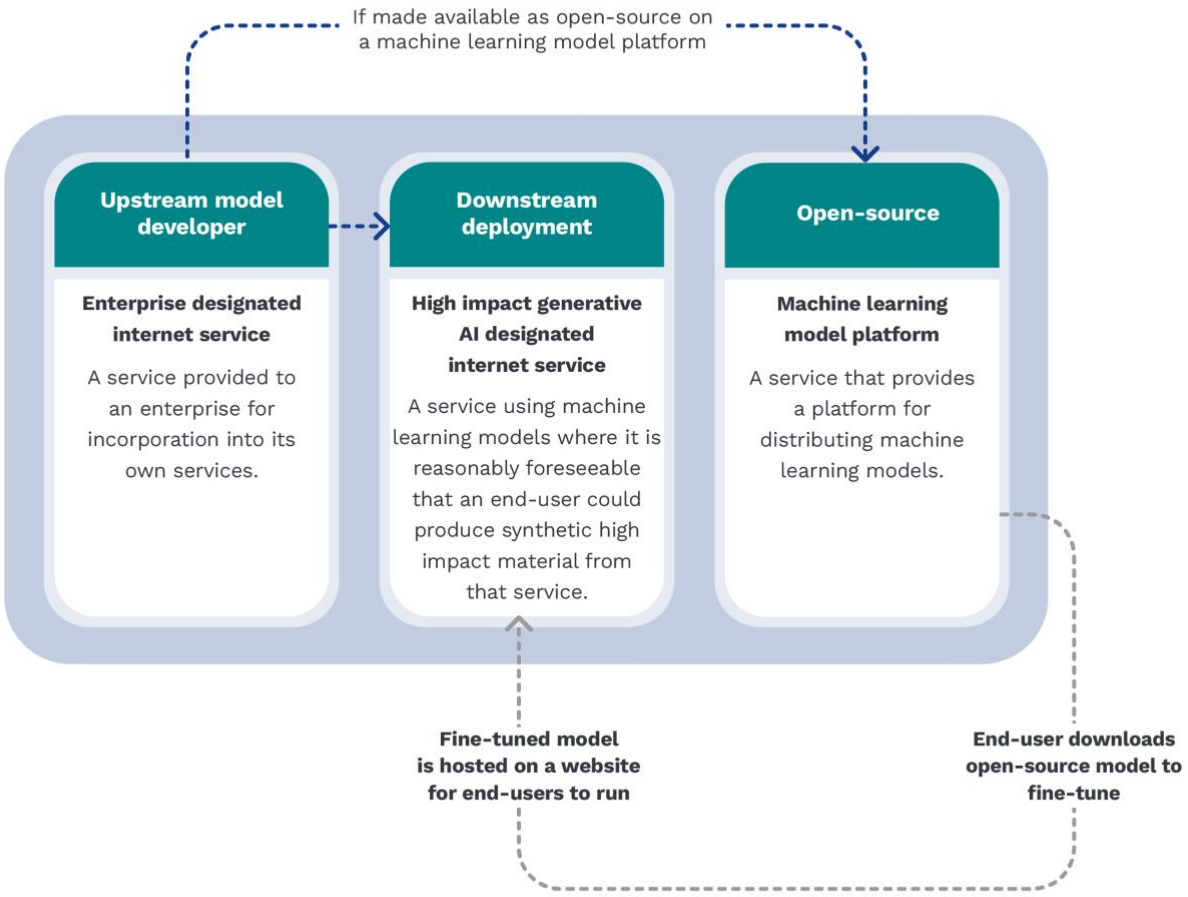
In particular, the availably of this technology can be used by perpetrators to generate synthetic child sexual abuse material and pro-terror material more easily and at scale. Generative AI can be used to create both completely synthetic material and deepfake images or videos of actual children. In addition, perpetrators can train models on existing child sexual abuse material to generate further material of victims. Perpetrators can also use this technology to scale sexual grooming and sexual extortion activity, including by creating material used to threaten a child.

[8] https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai

In order to address the evolving risks associated with generative AI, and ensure obligations are proportionate and targeted, the draft Standard places requirements on specific designated internet service categories:

- **High impact generative AI designated internet service,** a provider offering generative AI features (such as web and app-based image/audio/visual generators and conversational agents) that enable an end-user to produce material where it is reasonably foreseeable that the service could be used to produce synthetic high impact material.
- **Machine learning model platform service**, a platform which distributes machine learning models by enabling end-users to upload, share and download models. The obligations that attach to this category are proportionate, and reflect the service provider's key role in driving both the development and the distribution of open-source generative AI services.
- **Enterprise designated internet services that also provide upstream generative AI services**. The enterprise designated internet service category includes a broad array of services offered to enterprises (corporations, government or organisations) which may have no involvement in generative AI. This broad category may include upstream providers of services which develop generative AI models. Under the draft Designated Internet Services Standard, section 23(4) proposes requirements on those enterprise providers which specifically provide pre-trained machine learning models for integration into consumer facing services. The proposed obligations for these providers are proportionate to their position in the generative AI ecosystem. The proposals recognise both the limited visibility and control they have over the downstream uses, and the capability of such providers to build in impactful protections.

Figure 1: How services with generative AI functionality are proposed to be covered



Machine learning model platform services

Open-source generative AI presents a significant risk in relation to AI generated child sexual abuse material and pro-terror material, in part due to the ability for safeguards to be removed. Platforms distributing open-source models therefore have an important role to play in this digital ecosystem.

This category comprises platforms which distribute machine learning models by enabling end-users to upload, share and download machine learning models. These platforms may also offer an active development environment for end-users.

Recognising that machine learning model platform services are not capable of reaching into the models themselves, the draft Designated Internet Services Standard does not impose obligations on them to improve or adjust a model.

Obligations do, however, reflect that such platforms can and do moderate what models they distribute.

A machine learning model platform service is distinct from both a high impact generative AI designated internet service and an upstream generative AI model developer. A machine learning model platform service effectively distributes or makes available models. eSafety recognises that a platform's control and capabilities over the models it makes available is more limited and the proposed obligations reflect this.

Proportionality in relation to high impact generative AI designated internet services

It is important that obligations are proportionate to a service provider's risk level. A service will meet the definition of a high impact generative AI designated internet service if it is reasonably foreseeable that the model may generate high impact material – that is, synthetic material which would be classified as X18+ or RC. Lower risk services, including those with effective safeguards which minimise the risk of producing high impact material are unlikely to meet this threshold.

Importantly, eSafety is not proposing that a designated internet service with generative AI features needs to completely rule out the possibility of high impact material ever being generated on its service. Given the nature of generative AI models, there may be risks that an end-user could, with sufficient effort, manipulate the model producing harmful material despite safeguards being built in.

Where it is not reasonably foreseeable that X18+ and/or RC material could be generated on a service, the service will not fall within a defined category and will need to carry out a risk assessment as required under the Standard.

Question 15

eSafety is seeking to place requirements on service providers that are best-placed to prevent the use of generative AI features to create and disseminate class 1A and class 1B material. Does the proposal achieve this?

Question 16

Do the draft definitions for a high impact generative AI designated internet service and machine learning model platform service capture the right services? Are there types of providers that should not be included or should be excluded?

Question 17

The high impact generative AI designated internet services category only captures models that meet a high impact threshold. It must be reasonably foreseeable that a service can be used to generate synthetic high impact material that would be classified as X18+ or RC. Is this threshold:

(a) appropriate for differentiating high impact and therefore high risk models?

(b) sufficiently clear to enable service providers to assess whether or not they meet the definition?

eSafety welcomes views on alternative thresholds which may be more suitable.

Generative AI obligations

eSafety is interested in views on the effectiveness of obligations on high impact generative AI designated internet services and machine learning model platform services. Given generative AI is rapidly evolving, an outcomes- and risk-based regulatory approach helps to ensure that service providers can continue to make positive online safety innovations.

There is a risk that generative AI models are used to create synthetic child sexual abuse material and pro-terror material. Section 23 seeks to address this risk with minimum compliance measures required of high impact generative AI designated internet services, machine learning model platform services and certain enterprise designated internet services.

Question 18

In relation to high impact generative AI designated internet services, do the proposed obligations (in particular, the section 21 obligation to 'detect and remove' and the section 23 obligation to 'disrupt and deter' child sexual abuse material and pro-terror material) provide appropriate safeguards? Are there specific challenges to deploying these measures in a generative AI context?

Question 19

In relation to machine learning model platform services, do the proposed obligations (in particular, the section 23 obligation to 'disrupt and deter') provide appropriate safeguards? Are there specific challenges to deploying these measures?

Question 20

In relation to relevant enterprise providers, do the proposed obligations (in particular, the section 23 obligation to 'disrupt and deter') provide appropriate safeguards? Are there specific challenges to deploying these measures?

Other compliance measures

Systems and processes for breaches

Sections 16 and 18 require certain designated internet service providers to have systems and processes in place to respond to breaches of class 1A and 1B material. Sections 17 and 19 require certain providers to respond to breaches of class 1A and 1B.

Section 14 is also relevant to these provisions, as it requires providers to have in place terms of use which give them the following rights in relation to breaches:

- suspend the provision of the service to the end-user
- impose specified restrictions on the use of the service by a specified end-user for a specified period
- terminate the agreement under which the service is provided.

For end-user managed hosting services, the relevant sections require these providers to have standard operating procedures that enable them to take appropriate action to engage with end-user reports, and assess and respond to potential breaches. This requirement builds in flexibility, so services which may be limited in their ability to investigate complaints, and unable to definitively ascertain whether class 1A or 1B material is being hosted and shared on their service, can still take action which meets this requirement. For example, they can:

- make appropriate enquiries into any reports of class 1A and class 1B material
- analyse metadata and information available such as usernames and file titles.

Question 21

Do sections 16 to 19 effectively reflect the considerations on minimum compliance measures outlined on pages 10-11 of this discussion paper?

Detecting and removing known material

Section 23 requires providers of end-user managed hosting services, Tier 1 designated internet services and high impact generative AI designated internet services to implement systems, processes and technologies designed to detect, flag and remove from the service instances of known child sexual abuse material and known pro-terror material.

This may involve using hash matching, machine learning, artificial intelligence, or other safety technologies. In the case of a high impact generative AI designated internet service this can mean detecting and flagging known material in training data or in end-user prompts.

Importantly, these requirements apply to the various parts of a service. For example, an end-user managed hosting service is required to detect known child sexual abuse material on all relevant parts of a service, including both stored and shared content. In this example, that means where it is not technically feasible for the provider of an end-user managed hosting service to hash match stored content (or deploy another technology capable of detection), the service provider will still be expected to hash match (or deploy another technology) on cloud storage files and directories that are shared while at the same time undertaking appropriate alternative measures on stored content.

Question 22

Do the obligations for detecting and removing child sexual abuse material and pro-terror material effectively reflect the considerations on minimum compliance measures outlined on pages 10-11 of this discussion paper?

Technical feasibility

Similar to the position under the Relevant Electronic Services Standard, eSafety recognises that in some cases, designated internet service providers will face technical limitations in detecting known child sexual abuse material and known pro-terror material.

Where it is technically infeasible for the provider to deploy tools to automatically detect and remove known child sexual abuse material and/or pro-terror material on the service, the provider is required to take appropriate alternative action. At eSafety's request, the provider must specify where it is technically infeasible to comply, and the appropriate alternative actions taken.

Service providers are also required to disrupt and deter both known and new child sexual abuse material and pro-terror material.

Examples of appropriate alternative actions for end-user managed hosting services include:

- receiving and actioning end-user reports
- using hash matching, machine learning, artificial intelligence and other detection technologies on parts of the service that are not end-to-end encrypted (such as content in end-user reports and group names)
- using AI or machine learning techniques to detect key words, metadata, and/or behavioural signals, associated with child sexual abuse material and pro-terror materials
- interventions that prevent end-users from storing this material on the service – for example, by acquiring and using off-platform information to help identify and block the registration of potential end-users who have distributed known child sexual abuse material and pro-terror material in other environments. This could mean providers taking into account credible information published, provided or validated by another service, about significant threats related to that end-user.

Question 23

Is the technical feasibility exception in the obligation to detect and remove known child sexual abuse material and pro-terror material appropriate? How effective will this obligation be with this exception?

Investment requirement for large designated internet services

- Section 24 requires Tier 1 designated internet services, end-user managed hosting services, and high impact generative AI designated internet services to make investments to improve their services. To ensure the Relevant Electronic Services Standard does not disproportionately burden smaller service providers, these obligations apply only where these service providers meet the following Australian monthly active user base threshold:
  - 1 million for Tier 1 designated internet services, high impact generative AI designated internet services and machine learning model platform services

- 500,000 for end-user managed hosting services.

Question 24

Do you agree with this monthly active user threshold, or are there other thresholds which can be deployed to ensure this obligation is proportionate?

Compliance costs

eSafety recognises that the establishment of the Designated Internet Services Standard will likely impact a range of service providers including through compliance costs. Compliance costs include administrative costs, such as record keeping costs and costs of reporting to the regulator. Compliance costs include substantive compliance costs, such as the costs of putting in place new systems and processes to meet regulatory requirements including end-user reporting and complaints systems.

Some service providers may already have measures in place that would meet regulatory requirements or require relatively small changes to be compliant, meaning that their substantive compliance costs could potentially be minimal.

Question 25

What are your views on the likely compliance costs for service providers and, in particular, the impact of compliance costs on potential new entrants?

Additional information

We encourage you to provide further information that may not be covered in answers to the previous questions.

Question 26

Is there any additional information eSafety should consider in determining the Designated Internet Services Standard?

eSafety.gov.au