Phase 1 Standards (Class 1A and 1B Material)

Regulatory Guidance

November 2024

Australian Government | eSafetyCommissioner

eSafety.gov.au

Contents

Overview of this guidance **3**

Part 1: Legal and regulatory framework for industry codes and standards **5**

What harmful online content is covered by industry codes and standards? 7

What is considered to be class 1A and 1B online material? 8

Addressing systemic risks 11

Which sections of the online industry are regulated and must comply with the industry codes and standards? 11

Which sections of the online industry must comply with the Phase 1 Standards? 13

Part 2: Applying the correct industry standard or code and risk profile **14**

Identifying the applicable industry standard or code 14

Relevant electronic services 15

Designated internet services 16

Risk profiles 17

How are services differentiated under the DIS Standard? 17

How are services differentiated under the RES Standard? 20

Risk assessment 21

Risk categorisation for services that are not pre-assessed or defined 21

eSafety may request information from service providers about risk profiles or categories 29

Part 3: Complying with the Phase 1 Standards **30**

Compliance with general obligations 30

Detecting and removing known child sexual abuse and pro-terror material 30

Disrupting and deterring known and new child sexual exploitation and pro-terror material 36

Notification of child sexual exploitation material and pro-terror material 38

Resourcing trust and safety measures 40

Providing mechanisms to report and make complaints 42

Compliance with specific obligations in the RES Standard 45

Some RES providers to remove class 1 material 45

Compliance with specific obligations in the DIS Standard 45

Scope of a high impact generative AI DIS 45

Part 4: How eSafety can assist service providers **50**

Part 5: Communicating with eSafety **51**

Risk profile notifications 51

Relevant changes to service functionality 52

Referring complaints to eSafety 52

Reporting on compliance....53
Compliance report format....54
Reporting on technical feasibility and reasonable practicability....54
Confidentiality of information in reports....55

Part 6: How do the Phase 1 Standards interact with other regulatory requirements?....56
Interaction with other requirements under the Online Safety Act....56
Basic Online Safety Expectations....56
Online Content Scheme....59
Safety by design....61
Interaction with international regulations and frameworks....63

Part 7: eSafety's approach to assessing compliance and enforcement....64
Monitoring and assessing compliance....64
Complaints from the public....64
Information eSafety will take into account....65
eSafety's approach to assessing compliance....65
Enforcement options....66

Annexure A....69

Overview of this guidance

This guidance is for participants in a particular section of the online industry who are covered by the Relevant Electronic Services – Class 1A and Class 1B Material Industry Standard 2024 (**the RES Standard**) and the Designated Internet Services - Class 1A and Class 1B Material Industry Standard 2024 (**the DIS Standard**) (together, **Phase 1 Standards**), as well as other stakeholders. It provides information about the Phase 1 Standards and the functions of the eSafety Commissioner (**eSafety**) in monitoring and enforcing compliance.

For an overview of each Standard, download the fact sheets.[1]

The Phase 1 Standards aim to minimise and prevent harms associated with distribution, access, and exposure to the most harmful forms of online material on these services (referred to as 'class 1A' and 'class 1B' material).[2]

The Phase 1 Standards apply mandatory obligations to the following two industry sections:

- **Relevant electronic services (RES):** Online services which enable end-users to communicate with one another by email, instant messaging, short message services (SMS), multimedia message services (MMS) or chat services, as well as services that enable end-users to play online games with each other, and online dating services.[3]
- **Designated internet services (DIS):** Online services which allow end-users to access material using an internet carriage service, or which deliver material to persons who have equipment appropriate for receiving that material, where the delivery is by means of an internet carriage service but excludes social media services, relevant electronic services and other identified services. This category includes many apps and websites, as well as file and photo storage services, and some services which deploy or distribute generative artificial intelligence (AI) models.

Standards were introduced for the relevant electronic services and designated internet services industry sections because eSafety considered that draft codes, developed by industry associations in Australia, did not contain appropriate community safeguards, a statutory requirement for registration under sub-section 140(1)(d)(i) of ***Online Safety Act***

[1] eSafety website, Industry Codes and Standards webpage: eSafety.gov.au/industry/codes#understanding-the-phase-1-industry-standards.
[2] Class 1 material is defined in section 106 of the Online Safety Act 2021 (Cth). Class 2 material is defined in section 107 of the Online Safety Act 2021 (Cth).
[3] eSafety interprets the Relevant Electronic Services definition in the Online Safety Act 2021 (Cth) as including dating services. 'Dating service' is defined in the RES Standard under s 6.

2021 (Cth) (the **Act**). The legal and regulatory framework for the establishment of industry codes and standards is provided in Part 1 of this guidance.

The Phase 1 Standards commence on 22 December 2024.

The obligations set out in the Phase 1 Standards are mandatory and enforceable from the commencement date of those standards, which is 22 December 2024. eSafety will take a graduated approach, where appropriate, to compliance and enforcement of the Phase 1 Standards. As outlined under Part 7 – Enforcement Options, eSafety will not undertake any enforcement action until six months has lapsed from the Standards coming into effect, other than in exceptional circumstances such as in response to serious and/or deliberate non-compliance.

This guidance provides information on:

- the legal and regulatory framework for industry codes and standards (Part 1)
- applying the correct industry standard or code and risk profile (Part 2)
- steps that can be taken by service providers to ensure compliance with the Phase 1 Standards obligations (Part 3)
- how eSafety can assist service providers (Part 4)
- communicating with eSafety (Part 5)
- how the Phase 1 Standards interact with other regulatory requirements (Part 6)
- eSafety's approach to assessing compliance and enforcement with the Phase 1 Standards (Part 7).

This guidance should be read alongside the regulatory guidance for the six industry codes for class 1A and 1B material (**Phase 1 Codes**), particularly by providers of multiple services that may be captured by both codes and standards.[4] More information on the industry codes and standards, including the development process, is also available on eSafety's website.[5]

[4] eSafety website, Regulatory Guidance webpage: eSafety.gov.au/about-us/who-we-are/regulatory-schemes#phase-1-industry-codes
[5] eSafety website, Industry Codes and Standards webpage: eSafety.gov.au/industry/codes

Part 1: Legal and regulatory framework for industry codes and standards

eSafety is Australia's independent online safety regulator. Its mandate is to promote and improve online safety for all Australians. The Act provides eSafety with legislative powers to help prevent Australian residents and end-users in Australia from being exposed to harmful online content and activity.

Part 9, Division 7 of the Act provides for the establishment of mandatory industry codes and standards for eight sections of the online industry.

The removal powers in Part 9, Divisions 2 to 6 of the Act relate to specific instances of illegal or restricted online material. These are *ex post* powers, meaning they operate after material has been shared and identified. The industry codes and standards outlined in Part 9, Division 7 are intended to address risks and harms associated with particular types of online material at a systemic level, and are *ex ante*. This means that under industry codes and standards, participants in specific sections of the online industry[6] are required to take steps to address the presence of illegal or restricted online material on their services for people in Australia who are users of the service (end-users in Australia). These industry participants are referred to as 'service providers' in this guidance.

Representative industry associations drafted industry codes for each section of the online industry identified in Table 2 and submitted these to eSafety for registration. As required by the Act, the relevant industry associations consulted with service providers and the public more broadly in the preparation of the Phase 1 industry codes. More information on the industry associations involved in developing the Phase 1 industry codes can be found on eSafety's website.[7]

For an industry code to be registered, the eSafety Commissioner must be satisfied that it meets certain procedural and substantive requirements set out in the Act. In particular, the eSafety Commissioner must be satisfied that an industry code submitted for registration provides appropriate community safeguards for matters of substantial relevance to the community before registering a code.[8]

---

[6] Sections of the online industry are specified in Section 135 of the Act, and are outlined below. Section 136 provides that a person is a participant in a section of the online industry if the person is a member of a group that constitutes a section of the online industry.

[7] eSafety website, Background to the Phase 1 industry codes webpage, eSafety.gov.au/industry/codes/background-to-the-phase-1-industry-codes#what-are-the-industry-codes.

[8] Section 140 of the Act.

According to the commencement provisions contained in clause 7.1 of the head terms of the industry codes,[9] the codes become enforceable six months from the date of registration by eSafety. If a draft code does not meet the statutory requirements, the eSafety Commissioner is able to determine an enforceable industry standard for that section of the online industry.[10]

The eSafety Commissioner registered six of the eight draft codes addressing class 1A and class 1B material but found that codes submitted by industry associations for relevant electronic services and designated internet services did not meet the registration requirements because they did not provide appropriate community safeguards.

As a result, industry standards were developed and registered for relevant electronic services and designated internet services.[11] The development of the Phase 1 Standards built on the extensive work to develop the draft codes for relevant electronic services and designated internet services, and involved consultation with relevant industry associations, service providers, and digital, privacy and child rights groups.[12]

The Phase 1 Standards are risk-based, with requirements placed on providers that are proportionate to the risk their service presents with respect to class 1A and 1B material. The requirements in the Phase 1 Standards are also outcomes-based, setting out the objectives while remaining technology neutral, and allowing providers to choose how best to meet the required outcomes within their existing framework of operations.

Proactive obligations to detect and remove known child sexual abuse and pro-terror material[13] in the Phase 1 Standards are intended to be flexible to enable service providers to take steps and meet their obligations in a way that is suitable for their services.

Non-compliance will be enforceable by civil penalties and other enforcement options.[14] Enforcement is discussed in more detail at Part 7 of this guidance.

---

[9] Head Terms – Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material), available at eSafety.gov.au/industry/codes/register-online-industry-codes-standards#register-of-industry-codes.
[10] Section 145 of the Act.
[11] More information on the status of industry standards can be found at eSafety website, Industry Codes and Standards web page: eSafety.gov.au/industry/codes.
[12] For more information on the amendments made to the Standards post-consultation, download the fact sheets: eSafety.gov.au/industry/codes#understanding-the-phase-1-industry-standards; more information on the public consultation for the Industry Standards is available at eSafety.gov.au/industry/codes/standards-consultation.[12]
[13] DIS Standard s 20-21, RES Standard s 19-20.
[14] Sections 143-144, 146-147 and Part 10 of the Act.

What is the difference between the codes and standards?

Industry codes are developed by industry associations and registered by eSafety if they meet statutory requirements. eSafety is responsible for developing industry standards, not industry associations. Industry standards are legislative instruments that are tabled in parliament and which set out enforceable requirements. They are not technical standards of the kind developed by standards making bodies.

Both industry codes and industry standards are mandatory and enforceable. Available enforcement options include civil penalties, formal warnings, infringement notices, enforceable undertakings and injunctions.

What harmful online content is covered by industry codes and standards?

Industry codes and standards regulate online activities[15] related to class 1 and class 2 material. Class 1 and class 2 material ranges from the most seriously harmful illegal online content, such as videos showing the sexual abuse of children or acts of terrorism, through to content which is inappropriate for children, such as online pornography. eSafety refers broadly to such content as 'illegal and restricted online content'.

Online content can include written, video, audio and/or image-based material. Class 1 and class 2 material is defined under the Act by reference to Australia's National Classification Scheme. The National Classification Scheme is a cooperative arrangement between the Australian Government and state and territory governments. Its purpose is to classify films, publications and computer games. The definitions in the Act apply to films, publications, computer games and any other online material.[16] Additional information on the classification of material is available in the Online Content Scheme Regulatory Guidance on eSafety's website.

To facilitate implementation of the industry codes and standards, eSafety developed subcategories of class 1 material and asked industry to take a two-phased approach to developing industry codes and standards. The purpose of this was to prioritise the implementation of measures to prevent and reduce the most harmful online material. A number of industry and other stakeholders supported this approach, which captures the full range of class 1 and class 2 material in codes and standards, as intended by Parliament, but recognises the differences in the respective harms.

[15] Online Activities are listed in section 134 of the Act.
[16] Other material is material that is not a film, publication or computer game: Sections 106-107 of the Act.

What is considered to be class 1A and 1B online material?

Phase 1 Codes and Standards deal with class 1A and class 1B material. Class 1 material (which includes class 1A and 1B material) is defined in the Act with reference to the National Classification Scheme, and is material that is or would likely be Refused Classification under that Scheme. It includes material that:

- depicts, expresses or otherwise deals with matters of sex, drug misuse or addiction, crime, cruelty, violence or revolting or abhorrent phenomena in such a way that they offend against the standards of morality, decency and propriety generally accepted by reasonable adults to the extent that they should not be classified
- describes or depicts in a way that is likely to cause offence to a reasonable adult, a person who is, or appears to be, a child under 18 (whether the person is engaged in sexual activity or not), or
- promotes, incites or instructs in matters of crime or violence.

Class 1A and 1B material and their sub-categories are defined in the Phase 1 Standards and are set out in Table 1.

- Class 1A material is material that is seriously harmful, including child sexual abuse material, pro-terror material, and extreme crime and violence.
- Class 1B material is also harmful but is more context dependent, and includes crime and violence and drug-related material.

Requirements on class 1A versus class 1B material

As outlined in the supplementary explanatory statements, eSafety recognises that the harms arising from class 1B material can be less severe. This is reflected throughout the Phase 1 Standards, which attach less onerous obligations to class 1B material. The Phase 1 Standards also exclude class 1B material which have justifications from being in scope of the Phase 1 Standards.

Requirements to proactively detect material are limited to types of class 1A material, with service providers only required to have and enforce terms of use in relation to class 1B material and respond to breaches. This reflects our understanding that scalable measures for class 1B material may be more challenging and context dependent, thereby posing more risks to privacy and free expression if applied broadly.

Definitions for types of material have been influenced by the classifications system, in particular the Classification (Publications, Films and Computer Games) Act 1995, National Classification Code, Guidelines for the Classification of Films 2012, Guidelines for the

Classification of Publication 2005 and Guidelines for the Classification of Computer Games 2023.

However, eSafety acknowledges that varying terminology may be used across different jurisdictions and contexts. The standards provide that no particular phrases or words are required in the terms of use. For example, a provider may capture pro-terror material in a service's terms of use with different but similar terminology, such as 'terrorist' and' 'violent extremist' content. eSafety considers that this approach holds online service providers accountable for their own terms of use, and the enforcement of them.

Definitions for class 1A and class 1B, as well as specific types of material such child sexual abuse material, are defined in section 6(1) of the Phase 1 Standards.

The National Classification Scheme requires the nature of the material to be considered, including its literary, artistic, or educational merit and whether it is of a medical, legal or scientific character.[17] This requirement recognises the importance of context in classifying material.

A second phase of industry codes (or if required, industry standards) is being developed to address class 2 (restricted R18+/X18+) material, such as online pornography and other high impact material, as well as material identified by eSafety as class 1C material (certain fetish pornography falling within the definition of class 1 material). For further information please refer to the Phase 2 industry codes position paper.[18]

The subcategories subject to industry codes and standards are outlined in Table 1.

[17] *Classification (Publications, Films and Computer Games) Act 1995*, s 11.
[18] eSafety website, eSafety position paper: Phase 2 industry codes, eSafety.gov.au/industry/codes#esafety-position-paper--phase-2-industry-codes.

Table 1: Sub-categories and eSafety's phased approach

Phase	Sub-category	Material	National Classification Scheme
Phase 1	Class 1A	a) Child sexual exploitation material (CSEM) – material that is child sexual abuse material[19], that contains exploitative descriptions or depictions of a child, or that promotes or provides instruction of paedophile activity. b) Pro-terror material – material that advocates the doing of a terrorist act (including terrorist manifestos). c) Extreme crime and violence material – material that describes, depicts, expresses or otherwise deals with matters of extreme crime, cruelty or violence (including sexual violence) **without justification**.[20] For example, murder, suicide, torture and rape. Material that promotes, incites or instructs in matters of extreme crime or violence.	• Class 1 • Refused Classification (RC)
Phase 1	Class 1B	• Crime and violence material – material that describes, depicts, expresses or otherwise deals with matters of crime, cruelty or violence **without justification**. Material that promotes, incites or instructs in matters of crime or violence. d) Drug-related material – material that describes, depicts, expresses or otherwise deals with matters of drug misuse or addiction, or provides detailed instruction or promotion, **without justification.**	• Class 1 • Refused Classification (RC)
Phase 2	Class 1C	• Online pornography – material that describes or depicts specific fetish practices or fantasies.	• Class 1 • Refused Classification (RC)
Phase 2	Class 2A	• Online pornography – other sexually explicit material that depicts actual (not simulated) sex between consenting adults.	• Class 2 • X18+ • Category 2 restricted

[19] Child sexual abuse material, which shows a sexual assault against a child, is a narrower category and can be considered a sub-set of CSEM.

[20] Reference to 'without justification' highlights that the nature of the material must be considered, including its literary, artistic, or educational merit and whether it serves a medical, legal, social or scientific purpose. Section 11 of the *Classification (Publications, Films and Computer Games) Act 1995* outlines matters to be taken into account in making a decision on classification.

Phase 2	Class 2B	• Online pornography – material which includes realistically simulated sexual activity between adults. Material which includes high-impact[9] nudity. • Other high-impact material which includes high-impact sex, nudity, violence, drug use, language and themes. 'Themes' includes social issues such as crime, suicide, drug and alcohol dependency, death, serious illness, family breakdown and racism.	• Class 2 • R18+ • Category 1 restricted

More information on the National Classification Scheme ratings is available at classification.gov.au/classification-ratings/what-are-ratings.

The National Classification Scheme is currently undergoing a two-staged review process. Stage 1 reforms[21] commenced in 2024, while public consultations on Stage 2 reforms concluded in May 2024[22]. The Online Content Scheme applies to materials which would be rated R18+ (or Category 1 in the Publications Guideline) or higher.

Addressing systemic risks

The aim of industry codes and standards is to address systemic risks. To be effective, compliance measures set out in the Phase 1 Standards focus on proactive and systemic change, rather than the removal of individual items of material after they have already been shared. Similarly, compliance and enforcement activities will target systemic safety failures, rather than focusing on isolated incidents.

Which sections of the online industry are regulated and must comply with the industry codes and standards?

Under the Act, industry codes or standards apply to eight sections of the online industry, as set out in Table 2.

[21]Australian Classification website, National Classification Scheme Stage 1 Reform begins, classification.gov.au/about-us/media-and-news/news/national-classification-scheme-stage-1-reform-begins.

[22] Department of Infrastructure, Transport, Regional Development, Communications and the Arts websites, Modernising Australia's National Classification Scheme—Stage 2 Reforms, infrastructure.gov.au/have-your-say/modernising-australias-national-classification-scheme-stage-2-reforms.

Table 2: Industry sections and services covered by industry codes and standards

Industry section	Examples of services (non-exhaustive)
Relevant electronic services	• instant messaging services • Short Message Services and Multimedia Message Services • chat services • online multi-player gaming services • email services • online dating services • enterprise messaging services
Designated internet services	• file storage services managed by end-users in Australia • websites and apps* *Note: Unless an online service is otherwise considered a social media service or a relevant electronic service.
Social media services[23]	• social networks • public media sharing networks • discussion forums • consumer review networks
Search engine services	• electronic services designed to collect, organise (index) and/or rank information on the World Wide Web in response to end-user queries and return search results to end-user queries Note: Excludes search functionality within platforms where content or information can only be surfaced from that which has been generated/uploaded/created within the platform itself and not from the World Wide Web more broadly.
App distribution services	• services distributing apps that can be accessed by end-users in Australia (for example app stores/marketplaces) Note: Excludes links to apps and download of apps from third party websites.
Hosting services	• services which host stored material in Australia (for example services with data centres located in Australia)
Internet carriage services	• retail internet service providers (ISPs) that supply internet carriage services (including mobile and broadband) to end-users in Australia Note: Excludes providers of wholesale ISP services, including NBN Co.

Equipment services	• manufacturers, suppliers, maintainers and installers of equipment that is used to access online services[11] such as:
	  ○ mobile phones
	  ○ laptops
	  ○ tablets

[23] As outlined on pages 15 to 16 of this guidance, a service provider that offers instant messaging, chat functionality, or any other functionality contained in the definition of relevant electronic services in section 13A of the Act, that service will likely be considered a relevant electronic service. This will affect providers of social media services, regardless of whether they predominantly provide social media functionality.

- internet-enabled devices (such as smart TVs and gaming consoles)
- immersive technologies (such as virtual reality headsets)
- wi-fi routers

Note: This section of the online industry includes manufacturers of these devices, as well as businesses and retail outlets that install, sell and/or repair or maintain such devices.

Which sections of the online industry must comply with the Phase 1 Standards?

The Phase 1 Standards apply to two sections of the online industry, and the legislative instrument for each can be found in the following table. The RES and DIS Standards and their respective explanatory statements are available on eSafety's website.[24]

Table 3: Industry sections covered by Phase 1 Standards

Industry Section	Standard
Relevant electronic services	Relevant Electronic Services – Class 1A and Class 1B Material Industry Standard 2024
Designated internet services	Designated Internet Services - Class 1A and Class 1B Material Industry Standard 2024

[24] eSafety website, Register of industry codes and industry standards for online safety, eSafety.gov.au/industry/codes/register-online-industry-codes-standards

Part 2: Applying the correct industry standard or code and risk profile

Identifying the applicable industry standard or code

Consistent with the head terms of the Phase 1 Industry Codes, service providers are required to comply with the Phase 1 industry code or standard that applies to each separate service.

Service providers should have regard to the following non-exhaustive factors to assist in distinguishing one service that they provide from another:

- The presence of a separate sign-up process, including terms and conditions, for each service.
- The method(s) by which end-users can access each service (for example, whether via the same website or application).
- The functionality of each service and the level of integration of the functionality between the services (such as, what the service can do).

Section 5 of the RES Standard and DIS Standard set out the relevant application tests. The application test for the RES Standard differs from the DIS Standard:

- The DIS Standard requires an assessment of the service's predominant purpose to determine which industry code or standard applies. Section 14(1) of the Act also provides that the definition of a designated internet service excludes services that are social media services and/or relevant electronic services.
- For relevant electronic services, if a service meets the definition of a 'relevant electronic service' under the Act, then, as provided for by section 5(2) of the RES Standard, the service is subject to the standard and will not be required to comply with an industry code in relation to class 1A and class 1B material.

Provision of multiple services

Some service providers may provide multiple electronic services, in which case the service provider would be subject to the industry code or industry standard applicable to each electronic service which they operate. For example, a service provider may separately provide multiple services consisting of:

- one service which manufactures and/or supplies equipment used by end-users in Australia

- one service which makes available an online messaging service on the equipment.

Given that the service provider is the provider of two separate services, they would respectively need to comply with the Equipment Services Code and, once in effect, the RES Standard.

No service provider will have to comply with more than one industry code or industry standard in relation to the same electronic service. The schedule for each industry code may provide further detail as to the intended scope of that code. If a service provider is still unsure which industry code or industry standard is applicable to a particular electronic service, the service provider may seek guidance from the relevant industry association in respect of a particular industry code or from eSafety, although neither can provide legal advice.

Relevant electronic services

The RES Standard applies to a service that meets the definition of a relevant electronic service in section 13A of the Act and where the service is provided to end-users in Australia.

The 'relevant electronic service' definition provides for any service which enables communication with other end-users by means of email, instant messaging services, SMS services, MMS services, or online chat services, as well as services that enable end-users to play online games together. A service provider which meets the definition is considered a relevant electronic service and the provider will be required to comply with the RES Standard, regardless of whether the service also meets the definition of another industry section.

This is because Section 5.2 of the RES Standard provides that it applies 'to the exclusion of any industry code'. The Act also provides that industry standards prevail over industry codes to the extent of any inconsistency.[25]

This creates a more uniform regulatory framework, which recognises the specific risks that messaging, chat, and similar communications features provide, and which provides for greater certainty in the face of converging and evolving technologies.

This currently applies in relation to the Phase 1 Codes and Phase 1 Standards and not the proposed Phase 2 codes, which are intended to deal with different material and harms (including class 1C and class 2 material). The Phase 2 codes, which are under development, can be structured in a manner that is appropriate to the material and harms they are

[25] Section 150 of the Act.

addressing. The degree of alignment with Phase 1 will depend on how the Phase 2 codes, head terms and potentially standards develop. eSafety will develop regulatory guidance on Phase 2 after the registration of Phase 2 codes and/or standards.

Providers that align with both social media services and relevant electronic services

Services are increasingly fulfilling multiple purposes and offering several features for users. In particular, some social media services are also relevant electronic services. For example, a social media service may offer users a range of features that enable social interaction and allow users to post material as outlined in the Act's definition for a social media service. If that service also includes, for example, instant messaging or chat functionality, then that service is likely to meet the definition of 'relevant electronic service' in the Act and therefore be subject to the RES Standard. The presence of other features relating to social media does not affect this categorisation.

Services which fit this scenario have been subject to the Social Media Services Online Safety Code (Class 1A and Class 1B Material), which came into effect on 16 December 2023. Actions taken by service providers in order to comply with the Social Media Services Code will assist in ensuring compliance with the RES Standard because the minimum compliance measures in the Social Media Services Code are largely relevant and applicable to the RES Standard (see Annexure A for an indicative comparison). Providers can adapt any compliance processes established in response to the Social Media Services Code to the RES Standard.

However, once the RES Standard has commenced, services that fall into the definitions of both social media services and relevant electronic services will no longer have to comply with the Social Media Services Code. Instead, they will have to comply with the RES Standard.

Designated internet services

The DIS Standard will apply to a service that meets the definition of a designated internet service in section 14 of the Act[26], but will not apply where the service's predominant purpose is more closely aligned with another Phase 1 Industry Code.

[26] Services which meet the definition of a Social Media Service or Relevant Electronic Service in the Act cannot meet the definition of DIS: Section 14 of the Act.

Service providers operating a designated internet service, and that have more than one online service, should determine the industry code or industry standard that is most clearly aligned with the predominant purpose of each separate service.

For example, an app distribution service may have a website and so would be providing a designated internet service in relation to that website, however, the service's predominant purpose would be more closely aligned with the App Distribution Services Industry Code.

The Designated Internet Services category is broad, and this 'purpose' test aims to avoid the application of the DIS Standard to services that are more appropriately subject to another industry code or industry standard. This is consistent with the standards operating alongside the Phase 1 Codes head terms.[27]

Risk profiles

The Phase 1 Standards consider the level of risk that different kinds of services can pose to end-users in Australia in relation to class 1A and class 1B material. They recognise the different functionalities, capabilities and risks of a very broad range of services. Based on these risk levels and the type of service, the Phase 1 Standards then set out obligations that are proportionate and appropriate to the risk that class 1A or class 1B material will be generated or accessed by, or distributed to end-users in Australia, or stored on the service.

Pre-assessed and defined categories have risk profiles defined by eSafety. These categories differ between the RES and DIS Standards. Service providers that fall into these categories are outlined in the tables below.

Providers of services that do not fall within the pre-assessed or defined categories are required either to conduct their own risk assessments or default to assigning the service a Tier 1 (high) risk profile.

How are services differentiated under the DIS Standard?

Pre-assessed categories

For pre-assessed designated internet service categories, these risk profiles are ranked as either Tier 1 (highest risk) or Tier 3 (lowest risk). No services are pre-assessed as Tier 2 in the DIS Standard.

[27] Each Phase 1 Code has a common set of Head Terms called the: Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material) Head Terms.

Table 3: Pre-assessed categories in the DIS Standard

Pre-assessed categories	
Tier 1	**High impact DIS:** A website or app that has the sole or predominant purpose of enabling access to high impact materials (R18+, X18+ or RC) posted by end-users, such as certain 'gore' sites[28] and pornography sites.
Tier 2	No services are pre-assessed as Tier 2 in the DIS Standard. Tier 2 consists of services which have self-assessed their risk as medium.
Tier 3	**Classified DIS:** A service, such as a website, that has the sole or predominant purpose of providing general entertainment, news or educational content which is (or would be) classified no higher than R18+.[29] **General Purpose DIS:** A website or app that provides information for – or enables transactions related to – business, charitable, professional, health, news reporting, scientific, educational, academic research, government, emergency, counselling, or support service purposes. This category also applies to web browsers, as well as designated internet services that do not fall within other categories under the standard.[30] **Enterprise DIS:** A service provided to an organisation for use in the organisation's activities, such as for internal communication or ordering commercial supplies. This category also includes services which provide pre-trained artificial intelligence or machine learning models for integration into a service deployed or to be deployed by an enterprise customer.

Defined categories

Noting that a broad range of websites and apps meet the 'Designated Internet Service' definition, the DIS Standard seeks to provide clarity by requiring specific measures for defined categories of services with unique risk profiles. As they have unique risk profiles, these defined categories are not attributed a risk tier.

Table 4: Defined categories in the DIS Standard

Defined categories
End-user managed hosting service: An online service primarily designed or adapted to enable end-users to store or manage material, such as cloud storage for files and photos.
High impact generative AI DIS:[31] An online service that uses machine learning models to enable an end-user to generate material, where the service has not incorporated sufficient controls to

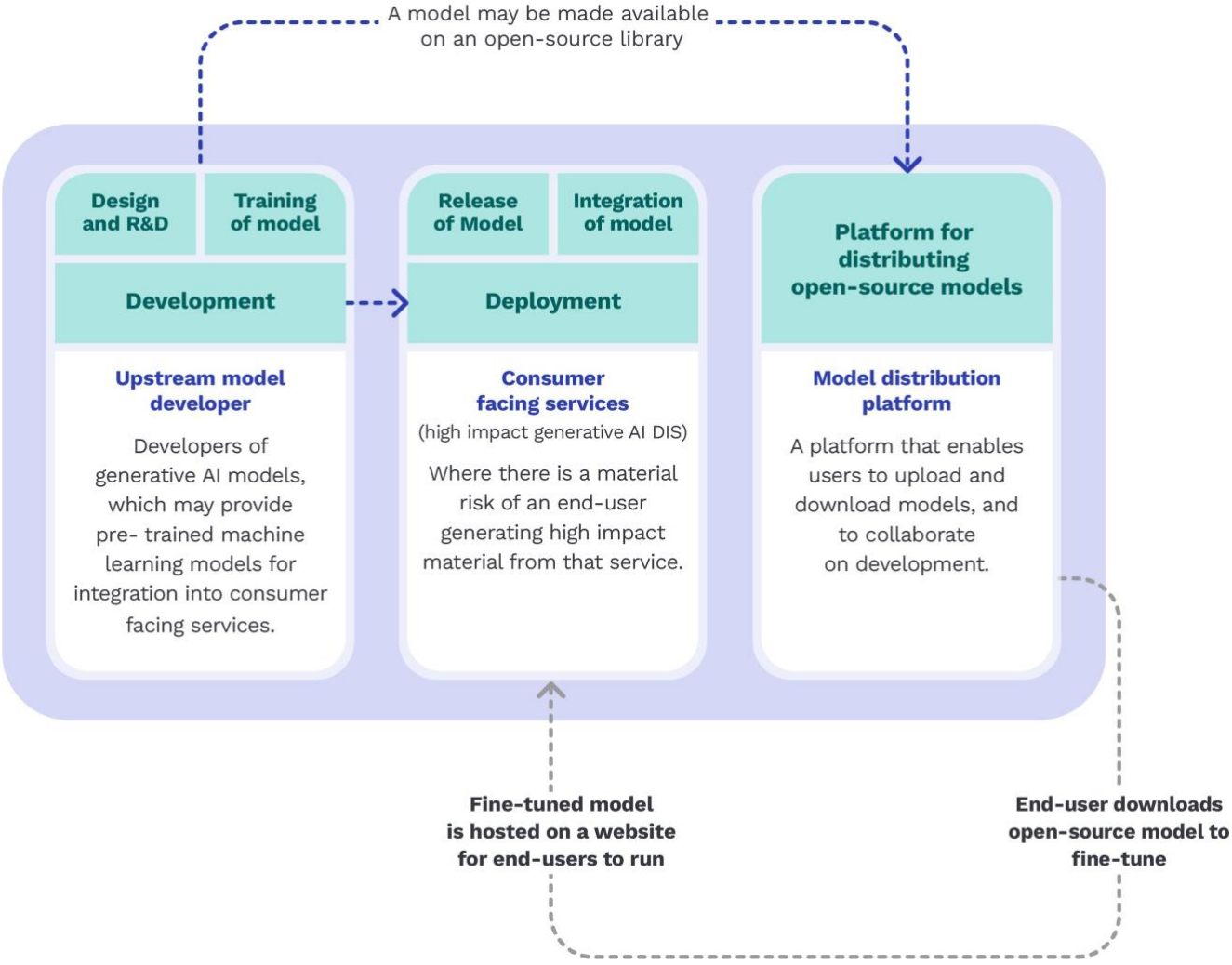
[28] Gore sites serve as digital hubs for the sharing of real-life killings, torture, and other forms of violence, catering primarily to 'gore seekers'; a niche audience searching for graphic and disturbing material (Institute for Strategic Dialogue, 2023).
[29] A classified DIS can only be pre-assessed as tier 3 if it meets the requirements of section 6(2) of the DIS Standard.
[30] A general purpose DIS can only be pre-assessed as tier 3 if it meets the requirements of section 6(2) of the DIS Standard.
[31] eSafety is engaging closely across government to align definitions, including as part of the Safe and Responsible AI approach. Terminology is distinct, reflecting the different scope of the DIS Standard.

reduce the risk of generating synthetic high impact (X18+ or RC) material,[32] such that the risk is not 'immaterial'. This may include some apps that 'nudify' images without effective controls to prevent their application to children, and pornography generators.

Model distribution platform: An online service which has a purpose that includes making available machine learning models, and which allows end-users to upload machine learning models.

The generative AI supply chain is shown in Figure 1. Only high-risk generative AI deployers and platforms distributing open-source models are assessed as defined categories. Upstream model developers are pre-assessed as Tier 3 in the Enterprise DIS category.

Figure 1: Treatment of generative AI services under the Designated Internet Services Standard



[32] There are slightly different thresholds for a high impact DIS (R18+ and up) compared to high impact generative AI DIS (X18+ and up). This is because a service with the predominant purpose of providing material that would be rated R18+ or higher presents a higher risk in relation to class 1A and class 1B material than a generative AI DIS that is merely capable of generating R18+ material.

Services outside the pre-assessed and defined categories

Providers of designated internet services that do not fall within the pre-assessed or defined categories are required either to conduct their own risk assessments, or default assign their service as a Tier 1 (high) risk profile.

How are services differentiated under the RES Standard?

Pre-assessed categories

Table 5 outlines the categories in the RES Standard which have been pre-assessed as being subject to the most comprehensive obligations.

Table 5: Pre-assessed categories in the RES Standard

Pre-assessed categories	
Communication relevant electronic service	This includes services that have a predominant purpose of enabling an end-user to communicate with another end-user or to view, navigate or search for other end-users with or without already having their contact details, which does not fit the other categories in the RES Standard – this includes online messaging and some video conferencing services, as well as some carriage services (email but not text messaging).
Gaming service with communication functionality	A service that enables end-users to play online games with each other, and enables end-users to share URLs, hyperlinks, images and/or videos.
Dating service	A service primarily used for dating that has a messaging function. This category does not include services that have a primary purpose of connecting end-users with escort or sex work services.

Defined categories

Table 6 outlines the categories in the RES Standard which have been defined with unique risk profiles.

Table 6: Defined categories in the RES Standard

Defined categories	
Telephony RES	A Short Message Service (SMS) or Multimedia Messaging Service (MMS) provided over a public mobile telecommunications service.
Enterprise RES	A service provided to an organisation to enable end-users within that organisation to communicate with each other.
Gaming service with limited communication functionality	A service that enables end-users to play online games with each other and allows limited sharing of some kinds of material (for example, in-game images and/or pre-selected messages).

Services outside the pre-assessed and defined categories

Providers of relevant electronic services that do not fall within the pre-assessed or defined categories are required either to conduct their own risk assessments, or default assign their service as a Tier 1 (high) risk profile.

Risk assessment

Risk categorisation for services that are not pre-assessed or defined

Where a service does not fall within a pre-assessed or defined category, we recommend that service providers use the risk profiles outlined in Table 7 and Table 8 as a guide when conducting a risk assessment under Part 3 of the Standards[33]. The tables are designed to support interpretation of the risk methodology matters which must be considered under section 8(5) of both Phase 1 Standards.[34] The risk assessment methodology does not provide any weighting to the matters that must be considered.

Depending on the nature of a service and the context it operates in, providers are likely to have additional risk factors to consider as part of their methodology. Some risk factors may not be applicable to a service. Service providers should consider only relevant risk factors.

Noting that each service is different, this guide offers a sliding scale of potential risk indicators which providers can apply as relevant to their services.

If a risk assessment indicates that the service may be in-between risk tiers, the provider should assign a higher risk profile to that service. This reflects the fact that while some

[33] DIS Standard Part 3, RES Standard Part 3.
[34] DIS Standard s 8(5), RES Standard s 8(5).

risks on a service may be lower, it only requires one higher risk feature or attribute to increase harm.

If a provider assesses its service's risk as high, then it must assign itself a Tier 1 risk profile with the highest level of obligations. A medium risk service should be assigned a Tier 2 profile. A low-risk service, which has minimal to no risk of being misused for class 1A or class 1B material, should be assigned a Tier 3 (low) risk profile with minimal obligations under the standards.

Designated internet services

Table 7: Designated internet services – risk profiles

Risk factor	Indicators of Tier 3	Indicators of Tier 2	Indicators of Tier 1
Predominant purpose[35]	The purpose is provision of a general purpose DIS or a classified DIS.	The purpose is not to provide a general purpose DIS, a classified DIS or a high impact DIS.	The purpose is to enable end-users to post or access high impact materials.
Posting material[36]	The service: e) does not enable end-users in Australia to post or view material to the service or a) enables end-users in Australia to post material only for the purposes of enabling such end-users to review or provide information on products, services, or physical points of interest or locations made available on the service or b) enables end-users in Australia to post or share material only for the purpose of sharing that material with other end-users for a business,	The service enables end-users in Australia to post and view material.	

[35] DIS Standard, s 8(5)(a).
[36] DIS Standard, s 8(5)(b).

	informational or government service or support purpose.		
Functionality – content creation[37]	The service: a) only makes available professionally produced material[38]to end-users and/or b) does not make available generative AI features.	The service makes available: a) professionally produced material and end-user generated material and/or b) generative AI functionality with a risk of producing material which would be classified as R18+ or lower.	The service predominantly makes available to Australian end-users: a) material which has been posted by any end-user. and/or b) generative AI functionality with a risk of producing material which would be classified as R18+, X18+ or RC.[39]
Terms of arrangement for content acquisition[40]	The terms of arrangement (contracts) with third party providers of material to the service: a) prohibit Class 1 material and b) require interventions to that ensure that ensure that the risk of Class 1 material is immaterial. Or c) due to the nature of the material provided under the terms of the arrangement by the third party provider, there is an immaterial risk of Class 1 material being provided.	The terms of arrangement (contracts) with third party providers of material to the service do not: a) prohibit Class 1 material and b) require robust methods to ensure that the risk of Class 1 material is immaterial.	

[37] DIS Standard, s 8(5)(c), (l) & (m).
[38] Professionally produced material is material produced by persons or entities who create such material:
- as a means of livelihood or for a commercial benefit; or
- on commission by the service provider (for example, a musical album by a professional musician, or a graphic design firm/photographer showcasing their portfolio).

[39] The requirements for a Tier 1 DIS and a high impact generative AI DIS are different. However, a service with the Tier 1 risk indicator of having functionality to produce R18+, X18 or RC material would likely also meet the definition of a high impact generative AI DIS.
[40] DIS Standard, s 8(5)(g).

Visibility of material[41]	Any uploaded material is visible only to the Australian end-user and service provider.	Any uploaded material is available to the service provider and the user, and may be made visible and accessible to other end-users of the service.
Terms of use[42]	The designated internet service has clear terms of use prohibiting the use of the service to solicit, access, generate, distribute or store (as applicable, having regard to the purpose and functionality of the service) class 1A material or class 1B material, and which give the designated internet service rights to enforce breaches of its terms of use.	The designated internet service has terms of use prohibiting the use of the service to solicit, access, generate, distribute or store (as applicable, having regard to the purpose and functionality of the service) class 1A material but not class 1B material. Or The designated internet service does not have terms of use prohibiting the use of the service to solicit, access, generate, distribute and store (as applicable, having regarding the purpose and functionality of the service) either class 1A or 1B material
Age of end-users[43]	The service is not likely to be accessed by children.	The service is likely to be accessed by children.
Safety by design guidance and tools[44]	Other information from relevant Safety by Design assessments and guidance, should be used to inform appropriate risk tiering.	
Other factors	The list of factors that can be taken into account when carrying out a risk assessment under section 8 are non-exhaustive. Any other matter relevant to the provider and the context in which they operate can form part of a risk assessment methodology.	

Examples of applying designated internet service risk profiles

Some **consumer-facing generative AI models** may be outside the scope of the defined category for a 'high impact generative AI DIS' if the providers are able to demonstrate that they present an immaterial risk of generating X18+/RC material. The providers of these services would therefore need to self-assess into a risk tier. If end-users are restricted from generating X18+ or RC material, but can generate R18+ material, then in applying the above risk profiles an assessment as Tier 2 may be appropriate.

Pornography services may be outside the scope of the pre-assessed category for a 'high impact DIS' where they are not making available high impact material that has been posted by end-users. For designated internet services which predominantly make

[41] DIS Standard, s 8(5)(e).
[42] DIS Standard, s 8(5)(f).
[43] DIS Standard, s 8(5)(h).
[44] DIS Standard, s 8(5)(j).

available pornography but do not allow end-users to post material, then in applying the above risk profiles an assessment as Tier 3 or Tier 2 may be appropriate.

Relevant electronic services

Table 8: Relevant electronic services – risk profiles

Risk factors	**Indicators of Tier 3**	**Indicators of Tier 2**	**Indicators of Tier 1**
Purpose[45]	The service is predominantly for: a) social interaction within a limited end-user group that has a pro-social common community interest (such as within a school, or neighbourhood or university community or a social or religious organisation or charity or sporting club or association) or b) social interaction within a commercial or public enterprise that is limited to employees and/or customers of the enterprise for the enterprise's stated purpose.	The predominant purpose of the service is to provide a forum for social interaction on a specific topic, such as to enable users to post reviews of products and services or for a limited commercial or public purpose. This may include the crowdfunding of commercial or charitable activities or social causes or to start an online petition for social change.	The predominant purpose is general social interaction, and it is not designed for social interaction in a specific context or for a specific purpose.
Function[46]	The service only enables:		The service enables:

[45] RES Standard, s 8(5)(a).
[46] RES Standard, s 8(5)(b).

	a) sharing of material on a one-to-one basis between end-users, or within a defined group of end-users or b) sharing of ephemeral material (material that lasts or is displayed only for a short time) without a sharing function. And the service does not have a chat or messaging service or a live streaming feature[47].	a) sharing and re-sharing of material to all end-users of the service and the material is permanent (not ephemeral) and/or b) the service has a chat or messaging service, and/or enables live video streaming.
Terms of use[48]	The service has clear terms of use prohibiting the use of the service to solicit, access, generate, distribute or store (as applicable, having regard to the purpose and functionality of the service) class 1A material or class 1B material, and which give the service rights to enforce breaches of its terms of use.	The service has terms of use, which may not be clear, and/or do not prohibit the use of the service to solicit, access, generate, distribute or store (as applicable, having regard to the purpose and functionality of the service): a) class 1A material but not class 1B material or b) neither class 1A or class 1B material. And/or c) the service does not have terms of use which give the service rights to enforce breaches.
Terms of arrangement for content acquisition[49]	The terms of arrangement (contracts) with third party providers of material to the service: a) prohibit Class 1 material and b) require interventions to that ensure that ensure that the risk of Class 1 material is immaterial or c) due to the nature of the	The terms of arrangement (contracts) with third party providers of material to the service do not: a) prohibit Class 1 material and b) require robust methods to ensure that the risk of Class 1 material is immaterial.

[47] Live streaming is live video that can be created and watched on a service by end-users in real time.
[48] RES Standard, s 8(5)(c).
[49] RES Standard, s 8(5)(d).

	material provided under the terms of the arrangement by the third party provider, there is an immaterial risk of Class 1 material being provided.		
Number of end-users in Australia that are monthly active end users[50]	1 to less than 500,000	500,000 to 3 million	Over 3 million
Likelihood of access by children[51]	The service is unlikely to be accessed by children.	The service is likely to be accessed by children.	
Safety by Design guidance and tools[52]	Other information from relevant Safety by Design assessments and guidance is used to inform appropriate risk tiering.		
Risk of generative AI material on the service[53]	The service does not make available generative AI functionality.	The service makes available generative AI functionality with a risk of producing material which would be classified as R18+ or lower.	The service makes available to Australian end-users generative AI functionality with a risk of producing material which would be classified as R18+, X18+ or RC.
Format of materials[54]	The service enables sharing of text or audio only.	The service: a) enables sharing of materials in text, image, audio and video and/or b) is enabled through immersive technologies.[55]	
Visibility of users[56]	The service **typically** only enables end-users to access and communicate with a list of contacts created by the end-user and does	The service enables end-users to do any of the following: a) view or create a list of other	The service enables end-users to: a) search for and contact other end-users or

[50] RES Standard, s 8(5)(b).
[51] RES Standard, s 8(5)(e).
[52] RES Standard, s 8(5)(g).
[53] RES Standard, s 8(5)(h).
[54] RES Standard, s 8(5)(a)-(b).
[55] Immersive technologies enable a user to experience and interact in three-dimensions (3D) with digital content in a way that looks, sounds and feels almost real. These technologies include augmented reality (AR), virtual reality (VR), mixed reality (MR) and haptics (interaction involving touch).
[56] RES Standard, s 8(5)(b).

	not enable end-users to: a) view or create a list of other end-users' individual connections on the service b) search for other end-users on the service using known identifiers (for example, name, username, email address) or connections c) search for other end-users on the service based on interests or keywords d) recommend other contacts to end-users based on interests or shared connections.	users' individual connections on the service b) search for other end-users on the service using known identifiers, but not search for other end-users or discover material on the service based on interests or keywords c) recommend other contacts, or material, to end-users based on interests or shared connections.	discover material on the service based on interests or keywords or b) recommends other contacts or material to end-users based on interests or shared connections.
Other factors	The list of factors that can be taken into account when carrying out a risk assessment under section 8 are non-exhaustive. Any other matter relevant to the provider and the context in which they operate can form part of a risk assessment methodology. For example, a relevant matter is also likely to be the number of global active end-users, as this heightens the risk that the service will be misused and the potential impact of dissemination of certain material, such as the child sexual abuse material of an Australian child.		

Assessing risk when making a material change

In addition to a provider assessing the appropriate category at the time the Standards come into force, section 7(5) of the Phase 1 Standards also requires a risk assessment in accordance with Part 3, where a service makes a material change that will increase the risk of class 1A or 1B material. This applies irrespective of whether the service provider is captured under a defined or pre-assessed category and would be ordinarily exempt from other requirements to undertake a risk assessment. This ensures that services which undergo substantial changes maintain appropriate categorisation under the standard.

Both section 7(5) and the obligation, in section 18 of the RES Standard and section 24 of the DIS Standard, to incorporate appropriate safety features and settings are triggered by a provider making a material change. The difference is that section 7(5) relates to undertaking risk assessments for the purposes of identifying the appropriate categorisation for compliance measures, while sections 18 and 24 relate to the requirement to assess and deploy safety features.

Providers are free to streamline compliance processes with risk assessment processes if they wish. For example, assessment processes around a material change could be streamlined to meet aspects of section 7(5) and the requirement to incorporate safety features and settings simultaneously.

eSafety may request information from service providers about risk profiles or categories

eSafety may require information from a service provider about their risk profile. Where the provider of a service is required to conduct a risk assessment, they supply documents about the risk assessment upon eSafety's request. These documents must include reasons for assigning a particular risk profile or category.

Part 3: Complying with the Phase 1 Standards

This part provides guidance on the steps that service providers can take to ensure compliance with the Phase 1 Standards obligations. This guidance is not exhaustive. It canvasses the major and more complex requirements that apply across both Phase 1 Standards, and then specific obligations under each Standard.

eSafety welcomes feedback on this guidance, and any areas where clarification or further detail would be helpful. This guidance will be periodically updated with further information, including as technology changes and new risks emerge, as this will have an impact on what is appropriate, reasonably practicable and technically feasible for providers to do.

Compliance with general obligations

Detecting and removing known child sexual abuse and pro-terror material

The Phase 1 Standards contain obligations[57] on some relevant electronic service providers and designated internet service providers to implement appropriate systems, processes and technologies to detect and remove known child sexual abuse and pro-terror material, unless doing so would:

- not be technically feasible and reasonably practicable
- introduce a systemic weakness or vulnerability into the service
- (in relation to an end-to-end encrypted service) implement or build a new decryption capability into the service or render methods of encryption used in the service less effective.

The importance of detecting and removing known child sexual abuse material is recognised internationally, for example in the UK's *Online Safety Act 2023* (UK) and Singapore's Code of Practice for Online Safety.

[57] DIS Standard s 20-21, RES Standard s 19-20.

Systems, processes and technologies

The Phase 1 Standards refer to 'systems, processes and technologies' as the basis for some requirements. This reflects the intent of the codes and standards, which is to provide safeguards focused on minimising systemic online harms, rather than only the removal of individual items of content after they have already been shared.

Processes consist of a series of steps which should be documented and followed internally, and which set out how providers may respond to risks, for example through standard operating procedures. Receiving and actioning user reports is an example of a process.

Technologies are capable of automatically taking certain actions, such as matching material against verified lists of child sexual abuse or pro-terror material.

Systems can encompass processes and technologies, as well as other inputs and outputs. For example, a system can integrate the use of hash matching technologies with processes for human review.

Technically feasible

The term 'technically feasible' maintains its ordinary meaning. It is intended that providers will first consider whether a system or technology is technically feasible before considering whether it is reasonably practicable.

The term 'technically feasible' does not contain any measure of practicality, proportionality or reasonableness. If it is possible under current technology for a person to do something, on this test taken in isolation, it will be technically feasible no matter the resources required to do so and any impacts on the person or business.

For providers of end-to-end encrypted service, please refer to the section on 'systemic weaknesses and vulnerabilities' on page 33-34.

For an example of where a measure would not be considered technically feasible, please refer to the 'Example of limitations on telephony RES' on page 45.

Reasonably practicable

What is 'reasonably practicable' is determined objectively. This means that a provider must meet the standard of behaviour expected of a reasonable person in the provider's position who is required to comply with the same obligation.

In determining whether the system or technology is or is not reasonably practicable, any burden in addressing impediments to implementation must be balanced against the severity of risks and harms to end-users.

When determining if a measure is reasonably practicable, providers should consider the risk of any child sexual abuse or pro-terror material being stored on, or distributed by or to, Australian end-users. Providers should also consider whether the system or technology is proportionate to that risk, the costs and practicality of implementation and whether the system or technology is likely to achieve the intended outcome of the Phase 1 Standard.

Interaction between 'technically feasible' and 'reasonably practicable'

In assessing a system or technology, a service provider might find that its implementation is technically feasible, but that there are other significant impediments to implementation that do not justify its implementation in the circumstances where the risk of certain material is low. For example, it may be technically feasible for a start-up relevant electronic service or designated internet service provider to design their systems in a way which enables hash-matching to be deployed. However, it may not be reasonably practicable for an early-stage company to do so while their user base is still small and the risk of harm is low.

Cost considerations

The cost of implementing a system, process or technology is one of the relevant considerations when assessing reasonable practicability. While cost burdens are different for every service, the costs of detecting and removing material should not be disproportionate to the risk.

Many hash-matching technologies are available for no cost. These are examples:

- Microsoft's PhotoDNA and Facebook's PDQ and TMK+PDQF are algorithms for image- and video-matching, which can be applied to child sexual abuse and pro-terror material.
- Google's CSAI Match is used for detecting child sexual abuse material in videos.
- The US-based National Center for Missing and Exploited Children (NCMEC) has a global hash sharing database for child sexual abuse material.
- The Global Internet Forum for Countering Terrorism (GIFCT) and Tech Against Terrorism both maintain databases of confirmed terrorist and violent extremist hashes, which enable providers to detect when this content is uploaded to their services.

It is important that hash matching databases are maintained, for example by implementing regular hash database updates, and ensuring processes are in place to check hash sets for authenticity and accuracy. A recent external audit of 538,922 images

and videos in NCMEC's hash database found that 99.99% met the US federal definition of 'child pornography'.[58]

Systemic weakness or vulnerability

Providers are not required to implement systems or technologies to detect and remove material where doing so would require the provider to implement or build a systemic weakness or systemic vulnerability into the service, or where it would require an end-to-end encrypted service to implement or build a new decryption capability or render methods of encryption used in the service less effective.

If a system or technology would make the encryption of an end-to-end encrypted service less effective, the Phase 1 Standard would not require it, and this applies whether a 'decryption capability' already exists or not. [59]

These exemptions in relation to systemic weaknesses and vulnerabilities are intended to complement those in the Online Safety (Basic Online Safety Expectations) Determination 2022 (Cth) (**BOSE Determination**).

Defining systemic weakness and vulnerability

The terms 'weakness' and 'vulnerability' have significant cross-over. The Australian Cyber Security Centre defines a vulnerability as 'a weakness in a system's security requirements, design, implementation or operation that could be accidentally triggered or intentionally exploited and result in a violation of the system's security policy.'[60]

Many organisations will use industry standard definitions such as those from the National Institute for Standards and Technology, which defines a weakness as a 'defect or characteristic that may lead to undesirable behavior.'[61] It notes examples including missing a requirement or specification, having an architectural or design flaw, and an implementation weakness, including hardware or software defect.

The language of 'systemic' weakness and vulnerability recognises that there is no such thing as a perfectly secure service. Every service provider introduces theoretical weaknesses and vulnerabilities when they implement features for their own business purposes, but responsible service providers seek to build, design and test their services to

[58] National Center for Missing and Exploited Children, 'Concentrix' Audit of NCMEC's Hash List', 2024, URL: https://www.missingkids.org/content/dam/missingkids/pdfs/Concentrix-NCMEC-document.pdf
[59] National Center for Missing and Exploited Children, 'Concentrix' Audit of NCMEC's Hash List', 2024, URL: https://www.missingkids.org/content/dam/missingkids/pdfs/Concentrix-NCMEC-document.pdf
[60] Vulnerability | Cyber.gov.au.
[61] Ross R, McEvilley M, Winstead M (2022) Engineering Trustworthy Secure Systems. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) NIST SP 800-160v1r1. https://doi.org/10.6028/NIST.SP.800-160v1r1.

minimise reasonably foreseeable risks. The explanatory statements recognise this, noting that risks must be 'actual and not merely theoretical'. 'Systemic' is intended to mean that the risk is a material weakness or vulnerability that effects the security of the system as a whole.

These provisions work in tandem with the protections for end-to-end encrypted services that also specify that providers are not required to do anything that would result in a new decryption capability or render methods of encryption used in the service less effective. For the avoidance of doubt, this makes clear that providers are not required to 'build back doors' or undermine end-to-end encryption.

Note: Due to the differences in purpose and technology between legislative schemes, the phrases 'systemic weakness' or 'systemic vulnerability' in the Phase 1 Standards should not be interpreted using the definitions or caselaw relevant to Part 15 of the Telecommunications Act 1997. This is also emphasised in the explanatory statements to the Phase 1 Standards.[62]

Appropriate alternative action

Certain providers are required to implement appropriate systems, processes, and technologies to detect and remove known child sexual abuse material and pro-terror material. Such technologies may include hashing technologies and machine learning. If it is not technically feasible or reasonably practicable for a provider to use such technologies, or doing so would build a systemic weakness/vulnerability or render encryption less effective, they must take appropriate alternative action.

When implementing an appropriate alternative action, for the avoidance of doubt, the provider is not required to implement any action that would render encryption less effective, result in a systemic weakness or vulnerability, or require the implementation of a new decryption capability.

Both the RES and DIS Standards refer to appropriate alternative action, these apply to several sections as follows:

- RES: Section 19 Detecting and removing known child sexual abuse material.
- RES: Section 20 Detecting and removing known pro-terror material.
- DIS: Section 20 Detecting and removing known child sexual abuse material.
- DIS: Section 21 Detecting and removing known pro-terror material.

[62] Federal Register of Legislation, Explanatory Statement of the Online Safety (Designated Internet Services—Class 1A and Class 1B Material) Industry Standard 2024; Explanatory Statement of the Online Safety (Relevant Electronic Services—Class 1A and Class 1B Material) Industry Standard 2024.

The factors which must be considered when determining if something is appropriate are outlined in section 11 of both the RES and DIS Standards. The appropriate alternative action may comprise a suite of additional steps, which when considered holistically in the context of the specific service, provide risk mitigations and appropriate safeguards in lieu of a technology or system.

It is expected that service providers should assess the appropriate alternative actions that can be applied as part of a broader set of risk mitigations to protect the rights and best interests of children and end-users in Australia. eSafety can request a report from a provider describing where it is has deployed appropriate alternative actions with justification for the actions described (see **Part 5 Communicating with eSafety** for more information).[63] It is therefore incumbent on the provider to be capable of explaining how their alternative actions are appropriate in the circumstances.

Examples of appropriate alternative action

eSafety cannot be prescriptive, as what is appropriate to do will depend on the service in question. However, providers should use a combination of these measures and/or other suitable measures in a proportionate way. Some of these measures may also be steps that a provider takes to disrupt and deter end-users from using the service for known and new pro-terror material and child sexual exploitation material.

- When product features create obstacles to reliably detect child sexual abuse and pro-terror material, carrying out and documenting risk assessments to ensure that risks are fully considered and safety measures are built into a service's design, rather than considered afterwards when harms arise.
- Providing features such as interstitial or warning messages and blurring potential child sexual abuse or pro-terror material.
- Providing easily accessible user reporting mechanisms (see also pages 43 to 46 for more details of requirements to enable reports and complaints). User reporting features that enable the provision of recent messages to the service provider. The number of messages shared with the provider should be sufficient to enable the identification of the context of a communication, which may be particularly important for pro-terror material.
- Providing educational or supportive information, including safety information, to end-users.

[63] DIS Standard s 32; RES Standard s 33.

- Using classifiers to detect signals and metadata relevant to unlawful and harmful content (such as behavioural signals related to private group membership, frequency of joining or leaving groups, engagement with children or young people using the service).
- For end-to-end encrypted services, using hashing, machine learning, artificial intelligence and other detection technologies on any parts of the service that are not end-to-end-encrypted (such as profile pictures, content in user reports, group names).

For end-to-end encrypted services, these measures align with section 8 of the BOSE Determination and the reasonable steps outlined in pages 39 to 40 of the Basic Online Safety Expectations Regulatory Guidance.[64] These examples constitute important yet common measures which are already being deployed by some end-to-end encrypted services, as is show in responses to eSafety's transparency notices.[65]

Disrupting and deterring known and new child sexual exploitation and pro-terror material

Certain providers of relevant electronic services and designated internet services must implement systems and processes and, if it is appropriate to do so, technologies that:[66]

- effectively disrupt attempts by end-users to use the service to solicit, generate, access, distribute or otherwise make available, or store, child sexual exploitation material or pro-terror material
- effectively deter end-users from using the service to solicit, generate, access, distribute or otherwise make available, or store, child sexual exploitation material or pro-terror material.

This requirement applies to both new material (which has not previously been verified) and known material (which has previously been verified). It ensures that providers take meaningful steps to effectively disrupt and deter new and known child sexual exploitation and pro-terror material on their services, even if they are limited in their ability due to issues with technical feasibility, reasonable practicability, and systemic weakness/vulnerability.

The requirement is in addition to the complementary obligations to detect and remove known child sexual abuse and known pro-terror material. the 'technically feasible' and 'reasonably practicable' limitations which apply to the requirement to detect and remove known child sexual abuse and pro-terror material do not apply to the disrupt and deter

[64] eSafety website, Regulatory schemes, eSafety.gov.au/about-us/who-we-are/regulatory-schemes#basic-online-safety-expectations.
[65] eSafety website, Responses to transparency notices, eSafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices.
[66] DIS Standard, s 22; RES Standard, s 21.

requirement. It requires the service provider to implement systems and processes, but technologies need only be implemented **where it is appropriate to do so**.

This exception is in recognition that, at present, technologies to disrupt and deter new material may not be as accurate and robust as technologies to detect and remove known material, so service providers might encounter broader impediments in deploying such technologies. Within this provision the use of 'appropriate' is included, which can include a consideration of proportionality.

Examples of disrupting and deterring known and new material

eSafety cannot be prescriptive, as there are a wide range of suitable systems, processes and technologies that vary depending on the service, and on the factors which must be considered when determining if something is appropriate as outlined in section 11 of the Phase 1 Standards. These are some examples of steps that could be taken to disrupt and deter known and new material:

- The blocking of certain keywords and/or search terms that may be associated with child sexual exploitation or terrorism and violent extremism.
- Using machine learning to identify potential child sexual exploitation and terrorism and violent extremism.
- Using signals or indicators to prevent recidivism by end-users who have previously been banned or suspended for breaches of a provider's terms of use for child sexual exploitation or terrorism and violent extremism.
- Providing end-users with clear communication advising if they are engaging in child sexual exploitation or pro-terror material, including conduct that violates terms of use (for example, providing a warning via a pop-up).
- Setting children's accounts to high privacy and safety settings by default to prevent offenders from contacting them.
- Deploying measures under the detect and remove requirements in relation to known material.[67]

As stated in the explanatory statements of the Phase 1 Standards, in considering what an appropriate use of technology may be, providers can consider their specific contexts and user base, including any underrepresented groups which may be at greater risk of

[67] Measures to detect and remove known child sexual abuse and pro-terror material deployed under s 20-21 of the DIS Standard and s 19-20 of the RES Standard can achieve part of the requirements of disrupting and deterring child sexual exploitation and pro-terror material. The requirements to disrupt and deter, however, is broader in scope as they apply to new and known child sexual exploitation material rather than only known child sexual abuse material.

technology systems falsely flagging their material. Providers may also consider varying levels of accuracy which some machine learning classifiers have when classifying complex material at scale. As with any proactive technology, a system and process incorporating human review of outputs of the technology, and the ability of end-users to appeal, are important in helping to mitigate limitations in a tool's accuracy and robustness.

Service providers should invest in and keep abreast of developments in the use of AI in trust and safety, to ensure that their measures are effective, safe and proportionate.

Notification of child sexual exploitation material and pro-terror material

Child sexual exploitation material

The Phase 1 Standards require relevant electronic services and designated internet services to provide notification as soon as practicable, if they become aware of child sexual exploitation material on their service and believe in good faith that it is not already known to an organisation authorised to handle such material.[68]

These are government or non-government organisations, the functions of which include combating child sexual abuse and/or child sexual exploitation. Non-governmental organisations must be generally recognised as expert or authoritative in the context of combating child sexual abuse and exploitation, as well as lawfully authorised in the relevant jurisdiction to handle such material. An example of an organisation to which providers can refer child sexual exploitation material for the purposes of subsection 18(3) is the National Center for Missing and Exploited Children (NCMEC) in the United States.

NCMEC reporting procedure

NCMEC operates the 'CyberTipline' that receives reports from the public and digital service providers. Under United States (US) federal law, service providers that are headquartered in the US that become aware of child sexual abuse material (whether known or new), child sex trafficking, and/or online enticement on their services must report to NCMEC.[69] In 2023, NCMEC received 35,944,826 reports from industry.[70]

Using the automated reporting system available specifically for service providers or the CyberTipline form provided on NCMEC's website, a service provider can file a report with

[68] DIS Standard, section 18(3); RES Standard, section 16(3).
[69] U.S. Code § 2258A, s 18(a)(1)(A), govinfo.gov/app/details/USCODE-2011-title18/USCODE-2011-title18-partI-chap110-sec2258A.
[70] missingkids.org/gethelpnow/cybertipline/cybertiplinedata.

NCMEC. Reports should contain all the necessary information to enable expedient referrals to relevant law enforcement agencies.

Once a report is received, NCMEC will triage and analyse the information and files, including video and image content, to confirm that the material appears to meet the US definition of 'child pornography'. These assessments are triple-vetted by NCMEC analysts and managers. NCMEC then labels files that appear to meet the US definition of 'child pornography' and generate hashes, which are shared with digital platforms and services that voluntarily choose to use NCMEC's hash sharing database to enable more effective detection and removal of known child sexual abuse material.

CyberTipline reports are also made available to law enforcement agencies around the world to support local investigations.

As well as the child sexual abuse database, NCMEC operates other databases, including a database of exploitative imagery and Take It Down (for removal of nude, partially nude, or sexually explicit imagery of children under 18).

Pro-terror material

The Phase 1 Standards[71] also require certain RES and DIS providers to provide notification as soon as practicable, if they become aware of pro-terror material on their service and believe in good faith that it is not already known to an appropriate organisation.

This must be a non-government organisation that verifies material as pro-terror material or is generally recognised as having expertise in counter-terrorism. An example of an organisation to which providers can refer pro-terror material is Tech Against Terrorism.

Tech Against Terrorism

Tech Against Terrorism (TAT) operates a service for participating platforms and services to report pro-terror material so that it can be classified and hashed. Hashes are shared with other platforms to enable detection and removal of known content.

Tech Against Terrorism operates the Terrorist Content Analytics Platform (TCAP) which serves to verify terrorist content, alert URLs to affected services, archive this material, and then hash the terrorist content. The TCAP's scope includes: activity that poses a threat to life and content associated with designated terrorist organisations. Platforms receive alerts when verified terrorist content is detected and also receive hashes of

[71] DIS Standard, s 18(4); RES Standard, s 16(4).

content found across the internet in order to remove known material. The TCAP also provides a 24/7 online crisis response capability.

The TCAP offers a 'Trusted Flagger' portal that enables rapid sharing of verified terrorist content alerts to platforms without the need for the reporter to contact platforms directly. Tech against Terrorism's technology and analysts verify that the reported content is associated with a designated terrorist organisation pursuant to its inclusion policy, before alerting content to platforms, in order to ensure that terrorist content removals processes cannot be erroneously used to remove non-terrorist content from platforms.

Another example of an organisation to which providers can refer pro-terror material is the Global Internet Forum to Counter Terrorism.

Global Internet Forum to Counter Terrorism

The Global Internet Forum to Counter Terrorism (GIFCT) operates a hash-matching database to which its members can contribute. Hashes added to the database must fit within GIFCT's taxonomy. This taxonomy addresses content based on a terrorist or violent extremist entity producing the content, and the type (or behavioural elements) of the content the entity produced.

Amongst other requirements, all members of the GIFCT are required as a condition of membership to establish user reporting mechanisms that allow platforms and services to receive and act upon reports of illegal material and/or violated terms of service and user appeals.

Resourcing trust and safety measures

Some relevant electronic services and designated internet services are required to have and implement management, supervision and internal reporting arrangements. This is to ensure that the provider complies with the requirements of the Phase 1 Standard and can otherwise effectively supervise the online safety of the service.[72]

These arrangements may include duties and responsibilities for personnel, as well as the use of systems, processes and technologies. In practice, this may involve:

- nominating individuals or teams that are accountable for user safety policy creation, evaluation, implementation and operations

[72] DIS Standard, s 19; RES Standard s 17.

- embedding user safety considerations, training and practices into the roles, functions and working practices of all individuals who work with, for, or on behalf of the relevant electronic service or designated internet service provider
- requiring personnel to complete training such as eSafety's Safety by Design assessment tools[73] to understand realistic, actionable and achievable measures that providers of all sizes and stages of maturity can use to safeguard users from online risks and harms.

These services must also have, or have access to, sufficient personnel with the skills, experience and qualifications needed to ensure that the provider always complies with the requirements of the Phase 1 Standard.

Membership associations for trust and safety professionals such as the Trust and Safety Professional Association (TSPA) and industry-led organisations such as the Digital Trust and Safety Partnership (DTSP) provide resources such as curricula and best practice frameworks which may be useful for building foundational knowledge of good practices amongst trust and safety practitioners.

Implementing development programs

Some relevant electronic services and designated internet service providers are required to establish and implement, for the calendar year, a program of investment and development activities in respect of systems, processes and technologies.[74]

An example of an appropriate investment may be providing support, either financial or in kind, to organisations that work to combat child sexual abuse, child sexual exploitation or terrorism. Appropriate organisations could include universities, the WePROTECT Global Alliance, the Tech Coalition and the Global Internet Forum to Counter Terrorism.

Intersections with the Basic Online Safety Expectations

The steps that providers may take to comply with the development program requirements are relevant to section 6(3)(h) of the BOSE Determination, which provides the example of investing in systems, tools and processes to improve the prevention and detection of material or activity on the service that is unlawful or harmful.

As the explanatory statement to the updated *Online Safety (Basic Online Safety Expectations) Amendment Determination 2024* notes, 'investment' is not necessarily limited to financial investment but could include a broad range of initiatives such as participation in and support for research, pilot projects, and collaboration with law enforcement, non-government and government organisations or cross-industry collaboration.

---

[73] eSafety website, Assessment tools, eSafety.gov.au/industry/safety-by-design/assessment-tools.
[74] DIS Standard, s 23; RES Standard s 22.

Providing mechanisms to report and make complaints

Most relevant electronic services and designated internet services are required to provide tools that enable end-users to make a report or complaint identifying or flagging class 1A material or class 1B material, or to complain about the provider's non-compliance with an industry standard.[75]

While not required by the Phase 1 Standards, eSafety notes that it is best practice for service providers to also enable an end-user to appeal where their material has been removed or restrictions imposed on their accounts. Providers are encouraged to have complaints and appeals processes, which can enable end-users to raise examples where a provider may have made an incorrect content moderation decision.

The obligation to establish and operate user complaints and reporting systems is also a key feature of most international online safety frameworks.

Intersections with the Basic Online Safety Expectations

Requirements in Part 4 Division 3 of the RES and DIS Standards complement expectations set out in the following sections of the BOSE Determination:

- Section 13 and 15: mechanisms to report and make complaints about certain material, and breaches of a service's terms of use
    - Providers are expected to have 'clear and readily identifiable mechanisms' that enable end-users to report or make complaints about class 1 material and breaches of the service's terms of use.

Providers are expected to have 'clear and readily identifiable mechanisms' that enable end-users to report or make complaints about class 1 material and breaches of the service's terms of use.

Designated internet services reporting tools

Reporting tools must:

- be easily accessible on or through the service[76] and include, or be accompanied by, clear instructions on how to use them
- enable the complainant to specify the harm associated with the material, or the non-compliance, to which the report or complaint relates.

[75] DIS Standard, s 27; RES Standard, s 28.
[76] DIS Standard, s 27(3); RES Standard s 28(3).

eSafety expects that:

- if the service can be accessed through a website, the tools should be available on or directly through the same website
- if the service is only accessible through an app, the tools should be provided on or directly through the app – if the tools cannot be provided on or directly through the app, the information should be available on a main website that the service uses and end-users of the app should be directed to that website in the app. This requirement is to ensure that end-users can easily report class 1A and 1B material encountered on a service and that there are minimal impediments to doing so. For example, an end-user of an app should be able to easily access and find a direct reporting tool within the service.

In service reporting on relevant electronic services

Relevant electronic services must similarly have clear instructions and harm-specific reporting options. Reporting options must also:

- be easily accessible and easy to use
- be available 'in service', that is, not on a website separate to the website (or app) providing the service – if it is not technically feasible or reasonably practicable for the provider to offer access to the tools 'in service', these tools should be made available on an easily accessible website, with links to the website available through the service where possible.

For example, an 'in service' reporting tool in a messaging application may enable an end-user who receives an unwanted message to click a button on the same screen as the message and/or hold on the message, and be provided with an option that enables them to report the message within the application itself.

Additional rules for reviewing reports and complaints

Additional rules in relation to dealing with reports and complaints apply to providers of some relevant electronic services and designated internet services. Under these rules a complainant can require the provider to conduct a review of the outcome of the investigation into a report or complaint of class 1A or 1B material. The review must be conducted by a person other than the person who conducted the investigation into the report or complaint and the provider must take appropriate action to facilitate the review.[77]

When enforcing provisions, eSafety will consider a range of factors including the size of a service provider.

[77] DIS Standard, s 29(3)(a); RES Standard s 30(3)(a).

Providers must also record in writing their systems, processes and technologies used to conduct investigations and reviews and must ensure that their personnel who investigate reports and complaints, and conduct reviews, have appropriate training and experience, including training in and experience of the provider's applicable policies and procedures.

Appropriate training in the context of class 1A and class 1B material may involve specialist training in the specific harms that are the subject of these categories, and engagement with experts in relation to material that depicts child sexual abuse or pro-terror material. Additionally, appropriate training may include training in escalation processes to refer complex or specialist cases to expert teams.

Intersections with the Basic Online Safety Expectations

The requirement to provide mechanisms to report and make complaints in the Phase 1 Standards are also relevant to sections 6(3) and 14(1) and 14(3) of the BOSE Determination, which relate to resourcing of safety interventions and teams, as well as policies and procedures for dealing with reports and complaints.

As outlined in the Basic Online Safety Expectations Regulatory Guidance, it is important that a service's safety interventions are resourced proportionate to the risks identified and to enable compliance with the Expectations. This should include the following:

- Appropriately resourcing trust and safety teams, to ensure that appropriate safety interventions are in place, that interventions are working effectively, and that safety issues are responded to as a priority.
- Ensuring all relevant staff are suitably trained and supported, including through training on Safety by Design principles – there should be specialist training for trust and safety teams, and trust and safety functions should be subject to oversight and accountability by senior management.
- Ensuring trust and safety teams engage with experts in online safety and technology, as well as victims, to inform policies and processes.
- Having clear and effective escalation processes to refer complex or specialist cases to expert teams.

Additionally, providers are expected to have policies and procedures for dealing with reports and complaints of class 1 material and breaches of a service's terms of use (section 14(1)(c)). The Basic Online Safety Expectations Regulatory Guidance emphasises that these policies and procedures should enable prioritisation of, and response to, reports and complaints that likely relate to unlawful material or activity or present a serious threat to life, health or safety, and that this prioritisation should occur from the point at which the report was made.

Providers are also expected to review and respond to reports and complaints within a reasonable period of time (section 14(3)). The Basic Online Safety Expectations Regulatory Guidance sets out relevant information regarding what a 'reasonable' period of time means, including with reference to any stipulated timeframes for reviewing and responding to reports under the Standards.

Compliance with specific obligations in the RES Standard

Some RES providers to remove class 1 material

Sections 15 (responding to class 1A material) and 24 (responding to breaches of terms of use in relation to class 1B material) of the RES Standard require that providers of certain services remove the material as soon as practicable, unless it is **not technically feasible or reasonably practicable** for the provider to do so.

Providers must also take 'appropriate action' to ensure the service no longer permits access to or distribution of the material, that the breach ceases and that any further risks of breaches are minimised.

This requirement applies to providers of pre-assessed RES (communication, gaming and dating with communication services) telephony RES, Tier 1 and Tier 2 services.

Example of limitations on telephony RES

It would be considered technically infeasible and not reasonably practicable for the provider of a service that cannot access or does not hold the material on their service to be required to remove it. For example, a telephony RES such as a telecommunications carrier may have limitations on removing material. While they provide data and voice transmission services, data that passes through the carrier service may not be 'stored' and therefore may not be accessible to the carrier service provider to remove.

Compliance with specific obligations in the DIS Standard

Scope of a high impact generative AI DIS

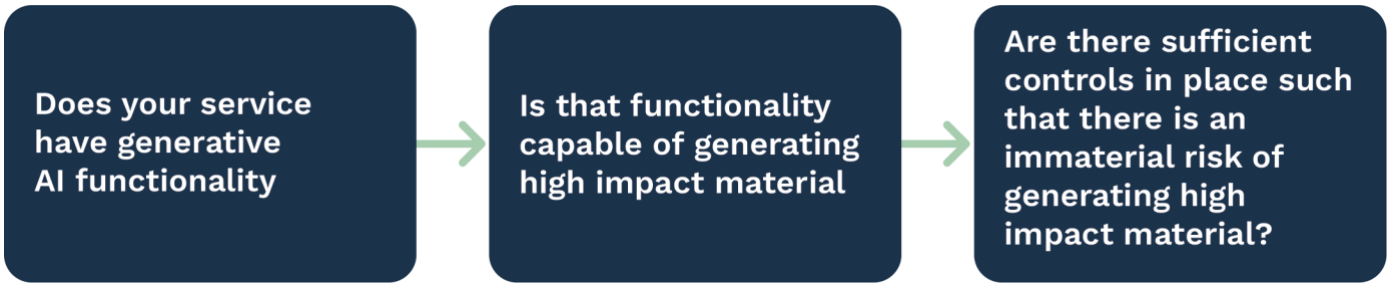
A '**high impact generative AI DIS**' is defined as a designated internet service that uses machine learning models to enable an end-user to produce material and is capable of being used to generate high impact material. **High impact material**, in relation to a high impact

generative AI DIS, is material that would be classified as X18+ Restricted or RC,[78] or in the case of publication, classified as Category 2 Restricted or RC.[79]

In addition to services dedicated to providing only a generative AI functionality to end-users, this category covers services with integrated generative AI functionality that can be used to produce high impact material. This includes completely new material, and material that has been created from editing existing material such as some instances of deepfake child sexual exploitation material.

To determine whether a service is capable of being used to generate high impact material, providers should assess the service as it is provided to end-users.

Figure 2: Providers should consider these questions to determine if their service is a high impact generative AI DIS.



This category is intended to capture services which do not implement appropriate controls, safeguards or interventions to reduce the risk of the service being used to generate high impact material. If the risk of producing high impact material is 'immaterial', then the provider will not be captured by the high impact generative AI DIS category.

Immaterial risk

A key element of the definition for a high impact generative AI DIS is whether the risk of generating synthetic high impact material using the service is 'immaterial'. Services should apply the following guidance on immaterial risk, to assess whether they are in scope as a high impact generative AI DIS.

An immaterial risk is one that is:

- likely to have insignificant consequences, or
- has a very low likelihood of occurring.

[78] X18+ material consists of sexually explicit material that depicts actual (not simulated) sex between consenting adults, and which can be AI-generated. The threshold for a high impact generative AI DIS does not include R18+ material, for example high impact nudity.
[79] DIS Standard, s 6(1).

eSafety considers that there are almost always significant consequences that flow from the risk of a generative AI service being used to generate high impact material such as child sexual abuse material. Therefore, the **likelihood** of a service being used in this way is the most relevant aspect for industry to consider when assessing whether a risk is immaterial.

This means that the risk that end-users can generate synthetic high impact material is only 'immaterial' in relation to a service if there is a very low likelihood of this occurring on the service.

Providers may deploy a range of design features and controls which lead to an immaterial risk of generating high impact material. eSafety's Generative AI Tech Trends Position Statement details the risk mitigations that can be deployed across the full generative AI lifecycle.[80] The report on Safety by Design for Generative AI: Preventing Child Sexual Abuse, developed by Thorn and All Tech is Human, also set out relevant risk mitigations.[81]

eSafety considers that for a general-purpose model to effectively mitigate the risk of generating high impact material, among other measures, relevant matching tools and classifiers should be deployed on training data, in user prompts and model outputs. The most effective solutions will apply safety interventions across each stage – in training data, user prompts and outputs.

Types of services considered high impact generative AI DIS

The risks arising from generative AI can be categorised in three ways:

- Inadvertently without explicit prompting – this occurs when a system unintentionally causes harm by generating incorrect or harmful responses.
- When explicitly prompted to do so – this happens when a model can be exploited for harmful purposes.
- The system has been optimised specifically for the purpose of generating high impact material.

Generative AI services which have been optimised or fine-tuned for high impact material – such as bespoke pornography generators, 'nudify' type models, and virtual girlfriends apps – present a higher risk of generating high impact material. Other services without clear terms of service and processes limiting the generation of potentially high impact materials are also likely to present a material risk.

[80] eSafety website, Generative AI – position statement, eSafety.gov.au/industry/tech-trends-and-challenges/generative-ai.
[81] Thorn website, Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments, thorn.org/blog/generative-ai-principles.

Even general-purpose services which clearly prohibit the generation of high impact materials typically involve some risk of perpetrators circumventing safeguards to use a model to produce class 1 material.

Services outside the scope of the high impact generative AI DIS category

Any provider of a Designated Internet Service with generative AI features that does not fall within the definition of a high impact generative AI DIS or other defined category is required to conduct a risk assessment under subsection 7(1) of the DIS Standard to determine the most suitable risk profile.

As outlined in Part 4, eSafety can require documents justifying the risk profile determined by a service provider and details of their risk assessment. For the DIS Standard, this includes the applicable risk methodology for the most recent risk assessment for the service.[82] Section 8(5) of the DIS Standard lists matters that must be taken into account, which includes the risk that any generative AI features of the service will be used to generate high impact materials.[83] Accordingly, service providers that consider their service to be out of scope of the high impact generative AI DIS category must, when requested, demonstrate their assessment of the risk of their service being used to generate high impact material.

Intersections with the Basic Online Safety Expectations

These requirements in the DIS Standard are also relevant to Section 8A of the BOSE Determination, and outline the expectation that service providers will take reasonable steps regarding the safety of generative AI capabilities.

As outlined in the Basic Online Safety Expectations Regulatory Guidance, it is important that providers:

- take reasonable steps to consider end-user safety at all stages of the life cycle of a generative AI capability
- 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 include:

- ensuring that assessments of safety risks and impacts are undertaken, identified risks are appropriately mitigated, and safety review processes are implemented
- ensuring, to the extent reasonably practicable, that generative AI capabilities can detect and prevent the execution of prompts that generate unlawful or harmful material

[82] DIS Standard, s 31(1)(d).
[83] DIS Standard, s 8(5).

- using educative prompts and nudges when users attempt to misuse generative AI capabilities.

The Australian Government's Safe and Responsible AI agenda

The DIS Standard complements the Government's broader work to ensure AI is developed and used safely and responsibly in Australia.

The Voluntary AI Safety Standard from the Australian Government Department of Industry, Science and Resources provides guidance to organisations on how to use and innovate with AI safely and responsibly.[84] Providers should consider ways to implement the Voluntary AI Safety Standard in tandem with complying with the DIS Standard. The requirements in the DIS Standard align with the voluntary guardrails which are framed at a higher level than the requirements in the DIS Standard. Both the DIS Standard and the voluntary guardrails address testing, transparency and accountability requirements. In addition, both are targeted at deployers of AI systems.

The Australian Government is considering options to mandate guardrails for AI in high-risk settings, including general-purpose AI models, with an emphasis on transparency, testing and accountability. eSafety continues to work with the Department of Industry, Science and Resource to ensure alignment across our respective workstreams, which have different timescales and focus areas.

[84] https://www.industry.gov.au/publications/voluntary-ai-safety-standard.

Part 4: How eSafety can assist service providers

The head terms to the Phase 1 Codes state that service providers may seek guidance and information from eSafety if they are unsure which code or standard applies to them and what steps they should take to meet compliance obligations and measures.

The guidance or information that eSafety will be able to provide in response to such requests will be general in nature. eSafety can provide further information on the interpretation of a provision in an industry code or standard or the Act, but is unable to provide legal advice as to how that provision applies to a specific set of circumstances.

Where service providers are concerned about their legal position with respect to Phase 1 Standards compliance, they should seek their own legal advice.

Industry participants can contact eSafety with general enquiries through the Codes and Standards Industry Compliance Portal. More information about the portal can be found on eSafety's website, including how to request access.

eSafety will also engage – both informally and during compliance assurance activities – with service providers and industry associations to understand the experiences of service providers during implementation of Phase 1 Standards.

eSafety will look to publish updated guidance on how to comply with Phase 1 Standards as particular compliance and enforcement issues are identified.

Part 5: Communicating with eSafety

Certain obligations in the Phase 1 Standards require relevant service providers to communicate with eSafety. The key communication obligations for service providers are:

- notifying eSafety of risk profiles when requested[85]
- updating eSafety about relevant changes to functions and features of their services[86]
- referring unresolved complaints to eSafety[87]
- providing reports about compliance with the Phase 1 Standards when requested[88]
- providing reports relating to technical feasibility and reasonable practicability when requested[89]
- providing reports on outcomes of development programs regarding development activities and investments undertaken in the calendar year.[90]

eSafety's systems will securely store information provided as part of these communications.

eSafety expects service providers covered by Phase 1 Standards to communicate with eSafety in a timely, appropriate and collaborative manner.

Service providers can provide some of the communications through the Codes and Standards Industry Compliance Portal. More information about the portal can be found on eSafety's website, including how to request access. All other communications can be made by contacting eSafety at codes@eSafety.gov.au.

Risk profile notifications

Service providers do not need to proactively notify eSafety of their risk profiles or categories. However, where a provider is required to conduct a risk assessment, eSafety can seek this information from a service provider by requiring documents outlining the risk profile determined by a service provider and details of their risk assessment.[91]

[85] DIS Standard, section 31; RES Standard, section 32.
[86] DIS Standard, sections 33-34; RES Standard, section 34.
[87] DIS Standard, section 30; RES Standard section 31.
[88] DIS Standard, section 36; RES Standard section 36.
[89] DIS Standard, section 32; RES Standard, section 33.
[90] DIS Standard, section 35; RES Standard, section 35.
[91] DIS Standard, section 21; RES Standard, section 21.

Relevant changes to service functionality

The Phase 1 Standards require providers of certain relevant electronic services and designated internet services to provide updates to eSafety on changes to features and functions of their services.

Providers of pre-assessed, Tier 1 and Tier 2 RES[92], Tier 1 DIS, Tier 2 DIS and end-user managed hosting services[93] must notify eSafety as soon as practicable if they decide to add a new feature or setting, or remove or make a feature or function inoperable, that would significantly increase the risk of class 1A and 1B material on the service. Providers of all designated internet services must also notify eSafety as soon as practicable if they decide to add a new feature or function that would significantly increase the risk of the service generating high impact material. [94]

Generally, eSafety considers that it would be good practice for service providers to notify eSafety within two weeks of a new feature or functionality change, which would trigger sections 33 or 34 in the DIS Standard and section 34 in the RES Standard.

In addition, there are separate requirements that apply prior to making material changes to a service. Before making material changes, providers of certain services must assess the kinds of features and settings that could be implemented on the service to minimise the risk that class 1A and 1B material will be accessed, or distributed to end-users, or which would be stored on the service.[95] eSafety can, by written notice, require a provider to provide these assessments within the period specified in the notice[96] On the basis of this assessment, service providers are required to determine and incorporate the most appropriate and effective features and settings for the service.

These obligations sit alongside eSafety's investigatory powers[97], as well as eSafety's powers in connection with the Basic Online Safety Expectations (outlined in Part 6 of this guidance).

Referring complaints to eSafety

Certain service providers are required to refer to eSafety any complaints made about their non-compliance with the applicable industry standard that they have not been able to

[92] RES Standard, section 24.
[93] DIS Standard, section 34.
[94] DIS Standard, section 33.
[95] RES Standard, section 18(2); DIS Standard, section 24(2).
[96] RES Standard, section 32; DIS Standard, section 31.
[97] See generally Part 14 of the Act. The Act enables the Commissioner to require a person to provide documents or information or attend before the Commissioner in relation to an investigation under Section 42, which includes an investigation into whether a service provider has breached a relevant industry code or standard: Section 199.

resolve. In the RES Standard, this applies to providers of pre-assessed and Tier 1 RES; in the DIS Standard, this applies to providers of Tier 1, end-user managed hosting services and high impact generative AI DIS.

This requirement is triggered where a provider becomes aware that the complainant is dissatisfied with the way in which the report or complaint was dealt with or the outcome of the report or complaint. A provider will not satisfy this obligation by merely referring a complainant to eSafety's codes and standards complaint form. It is expected that a provider will communicate directly with the Commissioner outlining:

- the details of the complaint, including the provider's internal reference number
- to the extent that the provider is aware, the reason that the complainant is dissatisfied.

This should be done as soon as reasonably practicable after becoming aware that a complaint is unresolved. Providers can notify eSafety by submitting the information through the Codes and Standards Industry Compliance Portal. More information about the portal can be found on eSafety's website. Service providers can request a link to the portal by contacting eSafety at codes@eSafety.gov.au.

The provider should advise the complainant that details of their complaint have been provided to eSafety to assist with our monitoring and assessment of industry standards. The provider should also advise the complainant that if they wish to provide further information about their complaint to eSafety, they can do so through the codes and standards complaint form (including the provider's internal reference number).

Reporting on compliance

Providers of a service may be required by written notice to provide the Commissioner a report for the most recent calendar year that specifies:[98]

- the steps the provider has taken to comply with the applicable standard and why these steps were appropriate
- the number of complaints made to the provider about the provider's non-compliance with the standard
- depending on the type of service, other information like the average monthly number of active end-users, the volume of child sexual exploitation and pro-terror material identified on the service, and details of the actions taken with regards to the identified material.

[98] DIS Standard, Division 4; RES Standard, Division 4.

The Phase 1 Standards require that service providers give the documents to the Commissioner within the period specified in the notice, which can be extended upon request. Unless a shorter period is reasonably necessary in the circumstances, eSafety will provide a minimum period of 30 days to produce such reports.

Compliance report format

eSafety may require compliance reports in a specified form. To minimise burden, eSafety may elect not to require information on all compliance measures in a report and instead prioritise those of most relevance. eSafety will have regard to other information held by eSafety, including through Basic Online Safety Expectations notices and previous reports in relation to compliance with the Social Media Services Code, before requesting a standards compliance report.

eSafety's website will provide reporting templates and other relevant guidance as it becomes available on the Industry codes and standards compliance page. Service providers regulated by industry standards can also notify and communicate with eSafety by emailing codes@eSafety.gov.au.

Reporting on technical feasibility and reasonable practicability

Where a service provider is required to implement systems, processes and technologies to detect and remove known child sexual abuse or pro-terror material, eSafety can request a report that describes:

- the cases in which it was not, or would not, be technically feasible or reasonably practicable for a provider to implement systems or technologies of a particular kind to comply with its obligations, including other obligations which refer to technical feasibility or reasonably practicability
- the systems or technologies that were or are available but were not, or would not be, implemented to comply because to do so would introduce a systemic weakness or systemic vulnerability, or would require an end-to-end encrypted service to implement a new form of decryption
- the alternative action taken to comply where a provider is required to implement appropriate alternative action.[99]

[99] RES Standard, section 33; DIS Standard, section 32.

Confidentiality of information in reports

Generally, eSafety does not intend to publish compliance reports or confidential information provided by service providers.

eSafety considers that confidential information includes, but is not limited to:

- information that is commercial-in-confidence (including trade secrets)
- other business information that would be unreasonable to publish
- information that could affect law enforcement and public safety
- personally identifiable information.

However, there may be circumstances in which the Act, or another Australian law, requires or authorises eSafety to disclose this material.

The key purpose of the compliance reports that may be requested under the Phase 1 Standards is to assist eSafety to determine compliance with the Phase 1 Standard that applies to any service and identify whether investigation and/or enforcement is appropriate and necessary. eSafety does not intend to publish compliance reports or other reports as required under the Phase 1 Standards as a matter of course. However, the information provided in a compliance report may be relevant to the exercise of statutory powers and functions by eSafety. For example, eSafety may use the information in deciding whether to commence an investigation into a complaint about class 1 material, or to determine the subject matter or recipient of a notice given in connection to the Basic Online Safety Expectations. In these cases, information provided as part of a compliance report may be publicly communicated.

eSafety can also be required to produce material in certain circumstances including:

- in response to a request under the Freedom of Information Act 1982 (Cth)
- at a court's direction or in performance of its duties in court proceedings
- in response to a Minister, house of parliament or another government agency's power to obtain information.

Service providers can also refer to information provided under existing voluntary reporting, or another reporting requirement under the Act. This may include publicly available information or information provided in response to a notice in connection with the Basic Online Safety Expectations. The purpose of this is to reduce the regulatory burden on service providers and potential duplication.

Part 6: How do the Phase 1 Standards interact with other regulatory requirements?

eSafety has a range of legislative functions and powers to regulate harmful online content and activity, including powers to issue removal notices and investigate breaches of service providers' regulatory requirements. Some of these functions and powers interact with regulatory requirements under the codes and standards. This section outlines how Phase 1 Standards may interact with other regulatory requirements in the Act and international regulations and frameworks.

Interaction with other requirements under the Online Safety Act

Basic Online Safety Expectations

The Basic Online Safety Expectations (**the Expectations**) set out the steps the Australian Government expects should be taken by providers of designated internet services, relevant electronic services and social media services to keep end-users in Australia safe online.

Compliance with the Expectations is not enforceable. However, eSafety has powers under the Act to obtain information from the applicable providers, on a periodic or non-periodic basis, about the steps they are taking to comply with the Expectations. eSafety can also publish statements about whether providers have or have not complied with the Expectations and summaries of the information received in response to notices. The aim is to increase the transparency and accountability of providers, thereby helping to incentivise and improve safety standards.[100]

The Expectations are set out in a determination from the Minister for Communications.[101] The BOSE Determination is a legislative instrument. In addition to the Expectations, it includes non-exhaustive examples of reasonable steps that can be taken to meet the Expectations.

The Expectations cover a broader range of online material and activity than the Phase 1 Standards. While some Expectations relate to material that can be required to be removed under the Act, other Expectations require steps in relation to all unlawful or harmful

[100] See generally Part 4 of the Act. A failure to comply with a reporting notice to the extent that a person is able can attract a civil penalty (up to 500 penalty units) in addition to other enforcement action: Section 50 of the Act.
[101] For the complete Expectations, see Online Safety (Basic Online Safety Expectations) Determination 2022 (23 January 2022) and associated Explanatory Statement. Both can be found on the Federal Register of Legislation's website at legislation.gov.au.

material and activity. Examples of unlawful material and activity covered by the Expectations include material that is illegal or refused classification[102], sexual grooming of children, and the sharing of non-consensual intimate images. Other harmful online material and activity covered by the Expectations include:

- all material or activity prohibited by law
- all harmful online material and activity under the Act
- other harmful activity that is prohibited or otherwise addressed in a provider's terms of use, policies and procedures, or standards of conduct for end-users.

Accordingly, the obligations in the Phase 1 Standards are narrower in scope than the Expectations as they focus on class 1A and class 1B material, rather than the broader unlawful and harmful material and activity covered by the Expectations.

In some cases, specific mandatory steps to address class 1 material required under the Phase 1 Standards will be directly relevant to an Expectations, including requirements to:

- undertake risk assessments and ensure safety by design (relevant to section 6 of the BOSE Determination)
- minimise the provision of certain material, including class 1 material (relevant to sections 6 and 11(d) of the BOSE Determination)
- incorporate safety measures in relation to generative AI capabilities (relevant to section 8A of the BOSE Determination)
- provide reporting and complaint mechanisms for end-users, and review and respond to reports and complaints (relevant to sections 13, 14, 15 and 16 of the BOSE Determination)
- ensure the implementation and enforcement of terms of use, policies and procedures that address class 1 material (relevant to sections 14, 15, 17 and 18 of the BOSE Determination).

[102] Under the National Classification Scheme.

Differences between Phase 1 Codes and Standards, and the Expectations

The Expectations apply to content and activity including:

- child sexual exploitation material including child sexual abuse material
- material that relates to terrorism and violent extremism, in addition to the narrower category of 'pro-terror' material.[103]

Where Phase 1 Standards require certain steps in relation to child sexual abuse material or pro-terror material only, such as 'detection', providers should consider whether they are taking reasonable steps to address the other harmful categories of material – such as child sexual exploitation material, terrorism and violent extremism – as set out in the Expectations. Reasonable steps could involve implementing similar steps to address this other material as those required under the Phase 1 Standards, or it could involve alternative reasonable steps.

Where a provider is using hash-matching technology to detect known child sexual abuse material,[104] and a dataset of known child sexual exploitation material is similarly available for the provider to hash-match, eSafety may consider it 'reasonable' for the purposes of the Expectations for the provider to use this technology to detect both categories of material to ensure a broader range of unlawful material is detected and addressed.

Table 1: Comparison of Basic Online Safety Expectations, Industry Codes and Industry Standards.

	Applies to	Applies to unlawful and harmful 'material'	Applies to unlawful and harmful 'activity'	Consequences for failure to comply
Basic Online Safety Expectations	• Social media services • Relevant electronic services • Designated internet services	Yes	Yes	eSafety may prepare, and publish, a Statement of Non-Compliance with one or more Expectations. eSafety has a range of enforcement options in relation to ensuring compliance with a reporting notice or determination.

[103] Pro-terror material is defined in the Standards at section 6. Pro-terror material is tied to class 1 material, and is a narrow subset than terrorist and violent extremist content.
[104] As required under section 20 of the DIS Standard, and section 19 of the RES Standard.

Industry Codes (Phase 1)	• Social media services • App distribution services • Hosting services • Internet carriage services • Manufacturers, maintenance and installation providers of equipment • Search engine services	Applies only to certain categories of class 1 material.	Applies to certain activities that affect the provision of certain categories of class 1 material	eSafety may issue a formal warning or written direction to comply with an industry code. Failure to comply with a direction may result in enforcement through an enforceable undertaking or injunction. It may also result in an infringement notice or civil penalty proceedings.
Industry Standards (Phase 1)	• Relevant electronic services • Designated internet services	Applies only to certain categories of class 1 material.	Applies to certain activities that affect the provision of certain categories of class 1 material.	Failure to comply with an industry standard may result in a formal warning, enforcement through an enforceable undertaking or injunction. It may also result in an infringement notice or civil penalty proceedings.

eSafety's Regulatory Guidance – Basic Online Safety Expectations contains additional information about the Expectations and highlights where the Expectations may overlap with obligations under Phase 1 Codes and Standards.

More information about providing compliance reports under the Phase 1 Standards is in Part 4 of this guidance.

Online Content Scheme

The Online Content Scheme under the Act gives eSafety a range of powers to deal with class 1 and class 2 material.

This part of the Act includes the framework for the development of industry codes and standards that are focused on reducing, at a systemic level, the risks associated with class 1 and class 2 material on online services.

It also focuses on using the Online Content Scheme for removal or restriction of specific instances of highly harmful material, referred to as 'Illegal and restricted content'. Key features include the following:

- A complaints scheme for online material that may be illegal or for which access should be restricted.
- Investigation and information gathering powers which allow eSafety to receive complaints about class 1 and class 2 material and investigate the provision of class 1 and class 2 material, whether in relation to a complaint or on eSafety's own initiative.
- Removal and restriction powers which allow eSafety to, in certain circumstances, give notices that require providers of social media services, relevant electronic services, designated internet services and hosting services to remove class 1 material and certain class 2 material from their services or ensure that access to certain types of class 2 material is age restricted.
- Powers related to compliance and enforcement of removal notices or notices requiring the restriction of material. This includes formal warnings, civil penalties, injunctions and seeking Federal Court orders to require a person to cease providing a social media service, relevant electronic service, designated internet service or internet carriage service.

Interactions between Phase 1 Codes and Standards, and other powers in the Online Content Scheme

The Phase 1 Industry Codes and Standards deal with class 1A and class 1B material on online services at a **systemic level** while the other powers under the Online Content Scheme relate to **specific identified examples** of class 1 and class 2 material.

Under the Online Content Scheme, eSafety may give a notice to providers of social media services, relevant electronic services, designated internet services or hosting services to take all reasonable steps to remove class 1 material within 24 hours, or a longer timeframe specified by eSafety.[105]

In addition, eSafety may give a written notice to an app distribution service requiring it to cease enabling the download of a particular app when certain requirements under the Act are met.[106]

eSafety can also give a link deletion notice to providers of search engine services requiring the service to stop providing a link that enables access to class 1 material within 24 hours, or a longer timeframe specified by eSafety, when certain requirements under the Act are met.[107]

[105] Sections 109-110 of the OSA.
[106] Section 128 of the OSA
[107] Section 124 of the OSA.

These powers under the Online Content Scheme complement the measures that service providers are required to comply with under the Phase 1 Codes and Standards.

More information about the Online Content Scheme can be found in our Regulatory Guidance – Online Content Scheme.

Safety by design

eSafety encourages service providers to adopt a safety by design approach when complying with the Phase 1 Standards, to ensure that user safety is positioned as a fundamental design consideration. In particular, the Phase 1 Standards require the following:

- Providers of some relevant electronic services and designated internet services must carry out an assessment of the kinds of safety features and settings that could be incorporated into the service to minimise the risk that class 1A material or 1B material, before making a material change to the service. They also require providers to incorporate these features and settings.[108]
- Providers of services which fall into a defined or pre-assessed category must carry out a risk assessment in accordance with a documented risk methodology. This risk methodology must consider safety by design guidance and tools published or made available by a government agency international body.[109]

Principles

At the heart of the Safety by Design initiative led by eSafety, there are three principles that provide platforms and services with guidance as they incorporate, assess and enhance user safety.

- **Service provider responsibility:** the burden of safety should never fall solely upon the user. Every attempt must be made to ensure that online harms are understood, assessed and addressed in the design and provision of online platforms and services.
- **User empowerment and autonomy:** the dignity of users is of central importance. Products and services should align with the best interests of users.
- **Transparency and accountability:** transparency and accountability are hallmarks of a robust approach to safety. They not only provide assurances that platforms and services are operating according to their published safety objectives, but also assist in educating and empowering users about steps they can take to address safety concerns.

[108] DIS Standard section 24, RES Standard section 18.
[109] DIS Standard section 8(5)(j), RES Standard section 8(5)(g).

Resources

Our Safety by Design assessment tools are intended to provide both a safety health check and a learning resource that helps companies continually improve online safety. The assessment tools take service providers through sets of targeted multiple choice questions, as well as information that is relevant to the overarching stream they select. The multiple choice questions ask service providers about the systems, processes and practices that are in place at their company. The responses generate a tailored report that identifies opportunities to improve user safety.

Integration of Phase 1 Codes and Standards, and Safety by Design

Safety by Design principles and tools can be used by service providers to support compliance with the Phase 1 Standards.

Example 1: Internal processes

Service providers should put in place infrastructure that supports internal and external triaging of safety issues, clear escalation pathways and reporting on all user safety concerns, alongside readily accessible mechanisms for end-users to flag and report concerns and violations at the point they occur.

Example 2: Default settings

Service providers should create default settings with the highest possible safety level and consider additional safety settings for children's accounts. This can include having settings that share profile information or location data switched to private mode by default. It can also include having friend lists and chat functions that are only accessible to approved contacts.

Example 3: Reporting and blocking

Service providers should make it easy for end-users to report accounts and block messages from other end-users. These features should be easy to find and use, for instance by being available in-app.

Example 4: Community guidelines

Having community guidelines helps to create and maintain a safe and positive service. They define a set of norms and expectations and a common code of conduct for users of a platform or service, including on safety issues. They often expand on the rules outlined in a platform's terms of service or terms of use, explaining them in a way that can be easily understood by all users.

The community guidelines should clearly articulate prohibited content and behaviour, and the consequences for violating the guidelines. They should be easy to find on the service's app or website.

More information on the Safety by Design initiative and tools can be found on eSafety's website.

Interaction with international regulations and frameworks

Countries around the world have introduced legislation and regulation to combat online harms. While each nation operates within different political and cultural contexts and norms, online safety frameworks and regulatory mandates share many key requirements.

Preventing access, generation and distribution of harmful content is a primary goal of online safety regulations in many countries, including Australia. There is also a growing focus on regulations that promote user empowerment – one of the principles of Safety by Design which is addressed and embedded into many of the obligations of the Phase 1 Standards.

eSafety has engaged with international regulators in the development of the Phase 1 Standards and has considered requirements under international standards to address similar online content and its associated harms. eSafety's commitment to regulatory coherence is also set out in the Global Online Safety Regulators Network's position statement, published in May 2024.

More information on how eSafety works with online safety regulators and organisations around the world is available on eSafety's website.[110]

[110] eSafety.gov.au/about-us/who-we-are/international-engagement.

Part 7: eSafety's approach to assessing compliance and enforcement

Monitoring and assessing compliance

eSafety recognises the importance of service providers being given sufficient time to prepare for and meet the required obligations of the Phase 1 Standards. This is reflected in the six-month transition period from registration to commencement. **eSafety will also not take enforcement action in the first six months of the Phase 1 Standards coming into effect, apart from exceptional circumstances such as in response to serious or deliberate non-compliance.**

eSafety will monitor compliance with Phase 1 Standards, effective 22 December 2024. This monitoring will inform any decision eSafety makes to commence an investigation into non-compliance with a Phase 1 Standard under the Act.[111]

eSafety may assess and/or investigate, on its own initiative or in response to complaints, whether a service provider has complied with the applicable Phase 1 Standard.

eSafety can require the provision of relevant information through examination or the production of documents from any person for the purpose of an investigation under the Act.[112] A refusal or failure to provide the required information or documents may be subject to criminal or civil penalties where an appropriate exemption to the requirement cannot be demonstrated.

Complaints from the public

eSafety can receive complaints about non-compliance with an industry code or standard from Australian residents, including end-users of services and the general public. We also welcome complaints and engagement on compliance issues with researchers, civil society organisations and other expert groups. eSafety will use the information provided in complaints to identify and address potential systemic issues in the online industry, to help keep Australians safer online.

End-users seeking to make a complaint about non-compliance with a Phase 1 Standard should be referred to eSafety's codes and standards complaint form.

[111] Section 146 of the Act.
[112] See generally Part 14 of the Act.

Information eSafety will take into account

eSafety may take a range of information into account when monitoring and assessing the compliance of service providers with the Phase 1 Standards. These are some examples:

- Complaints made directly to eSafety about potential non-compliance with obligations.
- Information from unresolved user complaints about potential non-compliance.
- Service providers' compliance reports provided to eSafety.
- Reports relating to technical feasibility and reasonable practicability of compliance relevant to certain provisions in the Phase 1 Standards.
- Information obtained through eSafety's other regulatory mechanisms (such as the Industry Codes, the Basic Online Safety Expectations and complaints data about illegal and restricted online content).
- Information that service providers already publish voluntarily or as part of international transparency initiatives.
- Information from stakeholders such as researchers, non-government organisations, law enforcement agencies and/or other governments.
- Information obtained through any routine assessment initiated by eSafety. For example, eSafety may check whether providers of applicable relevant electronic services and designated internet services have reporting and complaints mechanisms in place for class 1A and class 1B material, and may assess and test other features.

eSafety's approach to assessing compliance

In assessing a service provider's compliance with a Phase 1 Standard, eSafety will consider whether the actions a service provider has taken fulfil the requirements of applicable compliance obligations. Service providers are responsible for demonstrating that they meet the requirements of applicable obligations.

eSafety will seek to work cooperatively with service providers and will place significant weight on good faith and reasonable efforts by a service provider to comply with the applicable industry standard to the benefit of end-users in Australia, particularly in the early stages of implementation. eSafety will have regard to documented commitments by providers to take specific measures to implement the Phase 1 Standards, where deficiencies may exist.

In assessing compliance, eSafety:

- will take a fair and evidence-based approach
- will focus on the impact that compliance measures are having on the generation, access and distribution of class 1A material as a priority. To the extent that compliance with obligations is targeted to specific harms, eSafety will focus on obligations related to addressing and mitigating the most seriously harmful material (predominantly child sexual exploitation material and pro-terror material) and the detection and removal of that material
- will consider enforcement in relation to class 1B material only where there is serious non-compliance, where that information is available
- may use information gathered from monitoring, assessment and enforcement action to identify specific priority areas for compliance during subsequent years
- will communicate any priority areas publicly to encourage proactive compliance.

eSafety may approach a service provider to obtain further information and/or, where an investigation has commenced, use the information-gathering powers under the Act.[113] The steps taken by eSafety will depend on the nature of the potential breach being assessed and/or investigated, the information already available to eSafety and other factors and circumstances.

In some cases, eSafety may decide education and/or an informal request to seek rectification of a compliance issue is appropriate and likely to achieve compliance quickly and effectively. eSafety's compliance and enforcement policy also sets out more information regarding eSafety's approach.

Service providers may seek guidance and information from eSafety, noting the limitations around the advice eSafety can provide outlined at Part 4.

Enforcement options

eSafety takes a graduated approach, where appropriate, to compliance and enforcement. We strive to balance the protection of Australians against ensuring no undue burden is imposed on service providers.

eSafety will initially focus on working with industry associations and service providers to raise awareness of their obligations under the Phase 1 Standards.

[113] Sections 199, 203 of the Act.

If and when enforcement is necessary, eSafety has a range of options under the Act for addressing non-compliance with the Phase 1 Standards.

Enforcement options include the following:

- **Formal warnings:** A formal warning can be issued to advise a service provider that they have failed to comply with a Phase 1 Standard, and they could face further consequences if they continue to fail to comply.
- **Enforceable undertakings**: A service provider may enter into an agreement with eSafety to ensure compliance with a Phase 1 Standard. Once accepted by eSafety, the undertakings that a service provider has agreed to can be enforced by a Court.[114]
- **Injunctions:** An injunction is an order granted by the Federal Court of Australia or the Federal Circuit Court of Australia to compel a service provider to take certain actions, or to refrain from taking certain actions. An injunction is available where a service provider has not complied with a Phase 1 Standard.[115]
- **Infringement notices:** Infringement notices are notices that set out the particulars of an alleged contravention and specify an amount to be paid. If it is not paid, eSafety may commence civil penalty proceedings. Infringement notices may be issued by eSafety and do not require the involvement of a Court.[116]
- **Civil penalty orders:** These require payment of a financial penalty and can be directed towards service providers that do not comply with a Phase 1 Standard.[117] A civil penalty order can only be made by a Court following civil penalty proceedings.
- **Seeking Federal Court orders to require a person to cease providing a service**: eSafety may apply to the Federal Court of Australia to seek an order that a particular provider of a relevant electronic service or designated internet service stop providing that service in Australia. To apply for the order, eSafety must be satisfied that a service failed to comply with a civil penalty provision under the Online Content Scheme (such as breaching a Phase 1 Standard) on two or more occasions over the past 12 months, and that continued operation of the service poses a significant community safety risk. To grant the order, the Federal Court of Australia must also be satisfied of those factors.[118] eSafety will usually only pursue this option in relation to non-compliance with the industry codes or standards in the most extreme circumstances, such as where there is continuous non-compliance with no evidence that a service provider intends to take steps to comply.

[114]Section 164 of the Act.
[115] Section 165 of the Act.
[116] Subject to requirements in the *Regulatory Powers (Standard Provisions) Act 2014*.
[117] Sections 162-163 of the Act

More information about eSafety's approach to enforcement and investigative powers can be found on our website in our Compliance and Enforcement Policy.

Annexure A

Indicative comparison between the requirements in the Social Media Service Code and the RES Standard.

Obligation	Application to RES Standard	Application to SMS Code
Division 2—Compliance measures		
Terms of Use	✓ s 13	✓ Minimum compliance measures (MCM) 30, 31
Systems and processes for responding to breaches of terms of use: class 1A material	✓ s 14	✓ MCM 2
Responding to class 1A material	✓ s 15	✓ MCM 3
Notification of child sexual exploitation material and pro-terror material	✓ s 16	✓ MCM 1
Resourcing trust and safety functions	✓ s 17	✓ MCM 4
Safety features and settings	✓ s 18	✓ MCM 6, 7
Detecting and removing known child sexual abuse material and pro-terror material	✓ s 19-20 Requires systems, processes and technologies.	✓ MCM 8, 9 Requires systems, processes and/or technologies.
Disrupting and deterring child sexual exploitation/abuse material and pro-terror material	✓ s 21	✓[2] MCM 10
Development programs	✓ s 22	*Partly*[3] MCM 14, 16
Participation in an annual forum	✗	✓ MCM 15
Systems and processes for responding to breaches of terms of use: class 1B material	✓ s 23	✓ MCM 11
Responding to breaches of terms of use: class 1B material	✓ s 24	✓ MCM 12
Giving information about the Commissioner to end-users in Australia	✓ s 25	✓ MCM 21

Division 2—Compliance measures (CONTINUED)		
Responding to and referring certain unresolved complaints to the Commissioner	✓ s 26	✗
Dedicated section of service for online safety information	✓ s 27	✓ MCM 22
Division 3—Reports and complaints		
Mechanisms for end-users and account holders to report, and make complaints, to providers	✓ s 28	✓ MCM 25
Dealing with reports and complaints—general rules	✓ s 29	✓ MCM 26-28
Dealing with reports and complaints—additional rules for some services	✓ s 30	✗
At least annual reviews of effectiveness of its reporting systems and processes	✗	✓ MCM 29
Unresolved complaints about non-compliance to be referred to the Commissioner	✓ s 31	✓ MCM 18
Division 4—Requirements for reporting to the Commissioner		
Documents about risk assessments and other information	✓ s 32	✓ MCM 32, 33
Reports relating to technical feasibility and practicability of compliance with provisions of Division 2	✓ s 33	✗
Notifying changes to features and functions	✓ s 34	✓ MCM 19
Reports on outcomes of development programs	✓ s 35	✗
Annual compliance reporting by providers of Tier 1 Social Media Service	✗	✓ MCM 32
Commissioner may require compliance reports	✓ s 36	✗

eSafety.gov.au