# Fact sheet

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

November 2023

Australian Government | eSafetyCommissioner

eSafety.gov.au

# Contents

# Overview

The eSafety Commissioner has released for consultation the draft Online Safety (Designated Internet Services – Class 1A and 1B Material) Industry Standard 2024, referred to in this Fact Sheet as the **Designated Internet Services Standard**. The information in this Fact Sheet should be read in conjunction with the Discussion Paper, to inform submissions.

# Which services will need to comply with the Standard?

- The draft Designated Internet Services Standard will cover 'designated industry services' as defined in section 14 of the Online Safety Act 2021 (Cth) (**the Act**). This is a very broad category which includes most internet websites and apps, unless a service is otherwise considered a social media service or a relevant electronic service (which covers online communication services).
- Noting that a broad range of websites and apps are designated internet services, the draft Designated Internet Services Standard seeks to provide clarity by requiring specific measures for **defined categories** of services with unique risk profiles. For example:
    - websites offering generative AI functionality which meet the proposed threshold
    - end-user managed hosting services (online file and photo storage services).
- **Pre-assessed categories** are exempt from risk assessments because they are deemed to have a particular risk tier. Pre-assessed low risk services, or services which have determined they are low risk after conducting and documenting their risk assessment, will have no obligations under the Designated Internet Services Standard unless there is a change to the service's risk profile.
- If a designated internet service does not fall within a defined or pre-assessed category in the Standard, the service would need to undertake a risk assessment and would be classified in one of the following:
    - Tier 1 designated internet service: high risk
    - Tier 2 designated internet service: medium risk
    - Tier 3 designated internet service: low risk

- This tiered approach provides flexibility to cover future designated internet services which may not fall within one of the specified categories.

**Table 1: Defined and pre-assessed categories and risk tiers**

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

# What material is covered by the draft Standard?

- The draft Designated Internet Services Standard puts in place minimum compliance measures to address, minimise and prevent harms associated with access and exposure to the most harmful forms of online material. It covers:
    - class 1A material, which comprises child sexual exploitation material, pro-terror material, and extreme crime and violence material
    - class 1B material, which comprises crime and violence material and drug-related material.[2]
- These types of material are subcategories of class 1 material under the Online Safety Act, which is material that has been or would be refused classification

[1] Synthetic material which would be classified as X18+ or RC.
[2] 1 Importantly, the nature of the material, including its literary, artistic or educational merit, and whether it serves a medical, legal, social or scientific purpose, is relevant to the assessment of whether it is class 1B material. Material only falls within class 1B if there is no justification for the material.

(**RC**) under the Classification Act. Serious harms are associated with class 1A and 1B material whenever it is produced, distributed or consumed.

- A future industry code or industry standard will be developed to address class 2 material under the Act, such as online pornography.

# How is the draft Standard different to the draft Code?

- Many of the draft Designated Internet Services Standard provisions will look familiar to those involved in industry development of the draft Designated Internet Services Code which the eSafety Commissioner declined to register (referred to in this Fact Sheet as the **draft Code**) – including parts of the Head Terms[3], the overall approach to designated internet service categories and risk tiers, various definitions and minimum compliance measures.
- In creating the draft Designated Internet Services Standards, eSafety sought to build on the extensive work of industry in developing and consulting on the draft Code. This means that where appropriate, eSafety has used elements of the draft Code as an initial basis for the Designated Internet Services Standard.
- However, the draft Designated Internet Services Standard seeks to address the concerns about the draft Code that were set out in the eSafety Commissioner's Statement of Reasons on 31 May 2023, as well as additional issues identified by eSafety.
- The draft Designated Internet Services Standard has also been prepared in accordance with good practice for legislative instruments, as well as relevant requirements for effective regulation. This means that it does not have the same wording and format as the industry's draft Code. For example, detailed guidance and examples will be contained in the explanatory statement to the Designated Internet Services Standard and the regulatory guidance, instead of in the Standard itself.
- The draft Designated Internet Services Standard has strengthened requirements on certain providers including end-user managed hosting services to:

[3] Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material) Head Terms In force – latest version. https://www.esafety.gov.au/sites/default/files/2023-09/Consolidated-Industry-Codes-of-Practice-Head-Terms-12-September-23.pdf.

- detect and remove 'known' (verified) child sexual abuse material and 'known' (verified) pro-terror material where technically feasible
- disrupt and deter 'known' and 'new' (unverified) child sexual abuse material and pro-terror material
- invest in systems, processes and technologies that enhance the ability of the service to detect, disrupt and deter 'known' and 'new' (unverified) child sexual abuse material and pro-terror material
- enforce their own policies and terms of service relating to class 1A and 1B material.

# How does the draft Standard deal with the risks presented by generative AI?

- The draft Designated Internet Services Standard contains defined categories to reduce the evolving risks presented by generative AI in a clear and a targeted manner, while providing certainty over which generative AI services are captured. These are:
    - **High impact generative AI designated internet service**, a provider offering generative AI features (such as web and app-based image/audio/visual generators and conversational agents) that enable an end-user to produce material where it is reasonably foreseeable that the service could be used to produce synthetic high impact material.
    - **Machine learning model platform service**, a platform which distributes machine learning models by enabling end-users to upload, share and download models. The obligations that attach to this category are proportionate, and reflect the service provider's key role in driving both the development and the distribution of open-source generative AI services.
    - **Enterprise designated internet services that also provide upstream generative AI services**. The enterprise designated internet service category includes a broad array of services offered to enterprises (corporations, government or organisations) which may have no involvement in generative AI. This broad category may include upstream providers of services which develop generative AI models. Under the draft Designated Internet Services Standard, section 23(4) proposes requirements on those enterprise providers which specifically provide

pre-trained machine learning models for integration into consumer facing services. The proposed obligations for these providers are proportionate to their position in the generative AI ecosystem. The proposals recognise both the limited visibility and control they have over the downstream uses, and the capability of such providers to build in impactful protections.

- The draft Designated Internet Services Standard proposes specific obligations for each category of service, taking into account the role performed by each type of service in the generative AI lifecycle and their capability to disrupt and deter the use of generative AI models to produce child sexual exploitation and abuse material and pro-terror material.

# What is a high impact generative AI designated internet service?

- This category includes web- and app-based image/audio/video generators and conversational agents. It also encompasses services with generative AI functionality to produce new material that has been created from existing material (for example, deepfakes of child sexual abuse material).
- Importantly, a service will only be considered a high impact generative AI designated internet service if it is reasonably foreseeable that the model could generate either X18+ or RC material. A model that is designed with safeguards that effectively minimise the risk that the model can produce high impact material is unlikely to fall within this category. eSafety considers this appropriate given a provider of a model may be unable to effectively guarantee that an end-user could not manipulate the model and produce harmful material, despite safeguards being built in.
- This approach is intended to encourage services to proactively consider end-user safety and implement appropriate safeguards, without necessarily the need for formal regulation.

# Why is there a machine learning model platform service category?

- Open-source generative AI presents a significant risk in relation to AI generated child sexual abuse material and pro-terror material, in part due to the ability for safeguards to be removed. Platforms distributing open-source models therefore have an important role to play in this digital ecosystem.
- This category comprises platforms which distribute machine learning models by enabling end-users to upload, share and download machine learning models. These platforms may also offer an active development environment for end-users.
- Recognising that machine learning model platform services are not capable of reaching into the models themselves, the draft Designated Internet Services Standard does not impose obligations on them to improve or adjust a model. Obligations do, however, reflect that platforms can and do moderate what models they distribute.
- A high impact generative AI designated internet service is distinct from a machine learning model platform service because it does not distribute models, and because a machine learning model platform service lacks the capability to adjust individual models.
- Machine learning model platform services are also intended to exclude upstream generative AI model developers offering their products to enterprise customers for integration into downstream applications. They are addressed separately under the enterprise designated internet services category.

# How does the draft Standard differentiate services based on risk?

- Consistent with the registered codes for other sections of the online industry, the draft Designated Internet Services Standard adopts an outcomes- and risk-based approach. The measures contained in the Standard are proportionate to the risk a service presents in respect of class 1A and class 1B material.
- Similar to the draft Code, the draft Designated Internet Services Standard proposes compliance measures on a tiered basis, based on each designated internet service provider's self-assessment of the risk profile of its service.

  - Tier 1 designated internet services are considered higher risk, and thus attract more obligations, than other designated internet service tiers. They include pornography sites which enable end-users to post material for other end-users to access.
  - Tier 2 designated internet services have some compliance obligations. They could be a website or app that makes available both professionally produced material and end-user generated material, and where posted material is only visible to the service provider or a list of contacts created by an Australian end-user. An example of this could be a fan fiction website which makes available professionally produced material and allows end-users to post self-authored publications to the service.
  - Tier 3 designated internet services are not subject to any compliance measures under the draft Designated Internet Services Standard (unless they implement a significant new feature that would take them to a higher risk tier). Tier 3 designated internet services include services which do not share user-generated material and are focused on providing professionally produced material. These services include 'classified designated internet service' providers and 'general purpose designated internet service' providers.

- Certain pre-assessed categories of designated internet services are deemed to have a particular risk profile, and are therefore not required to conduct a risk assessment. These include:
  - websites or apps with the purpose of providing pornography, or material related to crime and violence ('high impact designated internet service', deemed to be Tier 1)
  - websites providing general entertainment, news or educational content that would be classified R18+ or lower ('classified designated internet service', deemed to be Tier 3) – for example, a video streaming service or a personal blog
  - news, educational or health websites and apps ('general purpose designated internet service' deemed to be Tier 3), for example a news website – this limits the compliance burden on a vast range of low-risk services that support commerce, and public purposes such as health and support services.
- There are also separate defined categories for services with risk profiles that typically reflect the nature of the services they provide. These categories are:
  - High impact generative AI designated internet service

  - Machine learning model platform service
  - Enterprise designated internet service
  - End-user managed hosting service.
- There are specific requirements that attach to these defined category services in accordance with their characteristics.

# Will service providers be required to monitor the contents of a person's cloud storage?

- One of the proposed compliance measures under the draft Designated Internet Services Standard is to require end-user managed hosting services to use systems, processes and technologies that detect and remove known, verified child sexual abuse material. These are images that have been verified against databases that are managed by well-recognised organisations whose functions are to combat child sexual abuse.
- eSafety recognises the importance of privacy regarding online file/photo storage, and **does not** advocate building in weaknesses or back doors to undermine privacy and security on end-to-end encrypted services. However, there are privacy preserving tools capable of detecting known child sexual abuse material and known pro-terror material that are widely available and frequently used by cloud storage services.
- In some cases, it may not be technically feasible for certain end-user managed hosting services, such as those using end-to-end encryption, to deploy tools to automatically detect this known material. We do not expect providers to design systematic vulnerabilities or weaknesses into end-to-end-encrypted services. However, other requirements apply in these circumstances.

# If it's not technically feasible for a service provider to detect and remove harmful material, what requirements will it be expected to meet?

- Where it is technically infeasible for the provider to deploy tools to automatically detect and remove known child sexual abuse material and pro-terror material on the service, the provider is required to take appropriate alternative action. At eSafety's request, the provider must specify where it is technically infeasible to comply, and the appropriate alternative actions taken.
- The draft Designated Internet Services Standard is technology-neutral and outcomes-based and does not specify particular actions and technologies to be deployed.
- Where it is not technically feasible for a service to detect and remove known child sexual abuse material and pro-terror material, the types of appropriate alternative actions that an end-user managed hosting service for example could take include:
    - using hashing, machine learning, artificial intelligence and other detection technologies on parts of the service that are not end-to-end-encrypted (such as content in end-user reports and usernames)
    - having clear and readily identifiable end-user reporting mechanisms so that people can seek help and alert services to breaches of terms of service, and promptly actioning end-user reports
    - using AI or machine learning techniques to detect key words, metadata, and /or behavioural signals indicating that an account may be engaging in illegal activity so that account can be flagged for human review
    - interventions that detect end-users likely to store this material on the service, for example, by acquiring and using off-platform information to help identify and block the registration of potential end-users that have distributed known child sexual abuse material and pro-terror material in other environments.
- Service providers are also required to disrupt and deter both known and new child sexual abuse material and pro-terror material.

# Will a service provider be required to comply with multiple standards and/or codes?

- Consistent with the principle in the Head Terms,[4] no service provider will have to comply with more than one industry code or industry standard in relation to the same electronic service. This is reflected in section 5 of the draft Designated Internet Services Standard.
- Providers of multiple online services will be subject to the industry code or industry standard applicable for each service.
- Where a single online service could fall within the scope of more than one industry code or industry standard, the code or standard that will apply is the code or standard that the service's predominant functionality is most closely aligned with.

# What happens if a service provider doesn't comply with the Standard?

- The draft Designated Internet Services Standard sets out minimum compliance measures which are enforceable and backed by civil penalties, enforceable undertakings and injunctions.
- If a designated internet service fails to comply with the Designated Internet Services Standard, then eSafety may make use of its enforcement powers under the Act. Unlike the Codes for other sections of the industry, under the Designated Internet Services Standard eSafety can take enforcement action without first directing the provider to comply with a requirement.
- eSafety will take a graduated and proportionate approach to enforcement. eSafety's approach to enforcement will be set out in its regulatory guidance for the Designated Internet Services Standard.
- eSafety will be able to receive complaints and investigate potential breaches of the Designated Internet Services Standard. When assessing whether adopted

[4] Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material) Head Terms In force – latest version. Page 4. https://www.esafety.gov.au/sites/default/files/2023-09/Consolidated-Industry-Codes-of-Practice-Head-Terms-12-September-23.pdf

compliance measures are reasonable, eSafety will consider a range of factors including the capability and size of a provider.

# When will the Standard come into effect?

- After the public consultation closes, eSafety will carefully consider all submissions and, where appropriate, amend the draft Designated Internet Services Standard.
- Depending on the public consultation, eSafety expects the Designated Internet Services Standard will be finalised and registered on the Federal Register of Legislation in April 2024.
- eSafety currently proposes that the Designated Internet Services Standard will commence six months from the time it is registered, to allow time for service providers to prepare for implementation and eSafety to provide regulatory guidance.

# How do the Industry Standards and Codes fit with the Basic Online Safety Expectations?

- The Industry Codes and Standards will impose enforceable obligations on eight sections of the online industry in relation to class 1A and class 1B material (and, in future, Class 2 material). By contrast, the Basic Online Safety Expectations provide a benchmark for preventing a broader range of online harms, setting out the Australian Government's expectations for three specific sections of the online industry: designated internet services, relevant electronic services and social media services.
- Designated internet services are covered by both the Industry Standards and the Basic Online Safety Expectations, which are designed to complement each other. Compliance with the requirements of the DIS Standard will be pertinent to a service provider's implementation of some of the Expectations but will not determine whether it meets the Basic Online Safety Expectations.

eSafety.gov.au