Fact sheet: Registration of the Designated Internet Services Standard

June 2024

The eSafety Commissioner (eSafety) has registered the Online Safety (Designated Internet Services — Class 1A and Class 1B Material) Industry Standard 2024 (DIS Standard). The DIS Standard minimises and prevents harms associated with access and exposure to the most harmful forms of online material on these services (referred to as 'class 1A' and 'class 1B' material). It addresses gaps and enhances the safeguards in industry's proposed code for these services, which was not accepted by eSafety. Background on this process can be found at: eSafety.gov.au/industry/codes/standards-consultation.

Which services does the DIS Standard cover?

The DIS Standard applies to designated internet services (DIS) provided to end-users in Australia. Designated internet services are defined by the *Online Safety Act 2021*[1] as online services such as websites and apps, which are not social media services[2] or relevant electronic services[3] (noting that relevant electronic services mainly provide online communication services).

What content is covered by the DIS Standard?

The DIS Standard covers two types of class 1 material that are associated with serious harms. These are:

- class 1A material, such as child sexual exploitation material (including child sexual abuse material) and pro-terror material
- class 1B material, such as crime and violence material and drug related material.

Australia's Online Safety Act 2021 defines these as materials that have been (or would be) Refused Classification under the National Classification Scheme, which can allow for legitimate uses of material (such as in art, research and journalism).[4]

[1] Section 13A of the Online Safety Act 2021.
[2] Class 1A and class 1B content on social media services is covered by the Social Media Services Online Safety Code (Class 1A and 1B Material).
[3] Class 1A and class 1B content on relevant electronic services is covered by the Online Safety (Relevant Electronic Services—Class 1A and Class 1B Material) Industry Standard.
[4] The National Classification Scheme is implemented through the Classification (Publications, Films and Computer Games) Act 1995.

How are services differentiated under the DIS Standard?

The DIS Standard recognises the different functionalities, capabilities and risks of a very broad range of online services covered by it. It sets out obligations that are proportionate and appropriate to the risk that class 1A or class 1B material will be accessed or generated by end-users of the service in Australia, or distributed to them using the service, or stored on the service.

Pre-assessed and **defined** categories have risk profiles deemed by eSafety. For pre-assessed categories, these risk profiles are ranked as either Tier 1 (highest risk) or Tier 3 (lowest risk). Designated internet services that do not fall within the pre-assessed or defined categories are required either to conduct their own risk assessments or assign themselves a Tier 1 risk profile.

If a service assesses its risk as high, it must assign itself a Tier 1 risk profile with the highest level of obligations. A medium-risk service can assign itself a Tier 2 profile – for example, this could be an online service that does not fall within a pre-assessed or defined category, where the material posted is only visible to an end-user's subscribers or clients. A low-risk service, which has minimal to no risk of being misused for class 1A or class 1B material, can assign itself a Tier 3 risk profile with minimal obligations under the DIS Standard – for example, a retail website which has minimal to no risk of being misused for class 1A or class 1B material will have minimal obligations under 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[5] 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+. **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,

[5] 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).

	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. **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
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:** [6] 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 reduce the risk of generating synthetic high impact (X18+ or RC)[7] material. 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.

What are the key obligations of the DIS Standard?

All designated internet services are required under the DIS Standard to have policies against using the service for class 1A material or class 1B material. In addition, in-service user reporting tools are required for the defined categories and Tier 1 and Tier 2 services (but not services that are pre-assessed or self-assessed as Tier 3 services).

There are further obligations for specific types of services. Some of these are covered in the following lists (but they may apply to more than one type of service).

[6] 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.

[7] There are slightly different thresholds for 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.

Online file and photo storage services (end-user managed hosting services)

Key obligations of these services include the following:

- Implement appropriate systems, processes and technologies to detect and remove known (pre-verified) child sexual abuse material on their service.[8] Providers of end-to-end encrypted services are not required to do anything that would break or weaken encryption.
- Implement appropriate systems, processes and technologies to detect and remove known (pre-verified) pro-terror material where the service suspects the end-user is storing pro-terror material on the service and the account is being accessed by more than one person. Providers of end-to-end encrypted services are not required to do anything that would break or weaken encryption.
- Take appropriate alternative action where exceptions are applied in relation to the detection and removal obligations.
- Implement systems, processes and – where appropriate – technologies to disrupt and deter end-users from using the service to make available or store child sexual abuse material or pro-terror material.
- Put in place, publish and enforce policies against end-users to stop them using the service to solicit, access, generate, distribute or store class 1A or class 1B material.

High-impact generative artificial intelligence (AI) services

Key obligations of websites or apps that use machine learning models which enable end-users to generate material, where the service has not incorporated sufficient controls to reduce the risk of generating synthetic high impact (X 18+ or RC) material include the following:

- Implement appropriate systems, processes and technologies to detect and remove known (pre-verified) child sexual abuse material and pro-terror material such as in user prompts, or take other appropriate action where this is not technically feasible or reasonably practicable.

[8] Determining whether something is ‘reasonably practicable’ involves a balancing of the impediments that a service provider might encounter in implementing a technology, such as cost or business model limitations, against the severity of risks and harms to end-users.

- Implement systems, processes and – if it is appropriate – technologies to effectively disrupt and deter end-users from using the service to make available or store child sexual exploitation material or pro-terror material.
- Disrupt and deter the generation of child sexual exploitation material and pro-terror material by:
    - implementing systems, processes and technologies that prevent generative AI features from being used to generate outputs of child sexual exploitation material and pro-terror material
    - regularly reviewing and testing models and promptly making adjustments or deploying mitigations
    - ensuring that end-users in Australia specifically seeking child sexual abuse material are presented with prominent messaging that outlines the potential risk and criminality of accessing such material
    - implementing systems, processes and technologies that differentiate AI outputs generated by the model
    - having systems, processes, and technologies to automatically detect and take appropriate action for child sexual abuse material in training data, user prompts and outputs.

Model distribution platforms

Key obligations of platforms that act as distributors or marketplaces for machine learning models include the following:

- Put in place and enforce terms of use that prohibit end-users from using the service to solicit, generate, distribute or store class 1A and 1B material, and that require end-users to take appropriate steps to minimise the risk of a model being used to generate child sexual exploitation material or pro-terror material.
- Implement systems, processes and – if it is appropriate – technologies to disrupt and deter end-users from using the service to make available or store child sexual exploitation material or pro-terror material.
- Take appropriate action for breaches of terms of use for the creation of child sexual exploitation material or pro-terror material, including by imposing restrictions on an end-user or terminating an account for repeated breaches.
- Provide in-service reporting tools and respond promptly to end-user reports with appropriate action.

Tier 1 services (such as high-impact pornography or gore sites)

Key obligations of Tier 1 services include the following:

- Implement appropriate systems, processes and technologies to detect and remove known (pre-verified) child sexual abuse material and pro-terror material on their service, or take other appropriate action where this is not technically feasible or reasonably practicable, or where to do so or would jeopardise the security of the service.
- Implement systems, processes and – where appropriate – technologies to disrupt and deter end-users from using the service to make available or store child sexual abuse material or pro-terror material.
- Put in place, publish and enforce policies against end-users using the service to solicit, access, generate, distribute or store class 1A material or class 1B material.
- Prevent end-users known to be under 18 from using high impact services, and require that only account holders can post or distribute material on the services.
- Provide in-service reporting tools and respond promptly to end-user reports with appropriate action.

Tier 2 and 3 services

The obligations on these services are significantly less than on other services.

Tier 3 services (including enterprise services and general purpose services such as websites or apps which provide general information such as news, educational content and health advice) are required only to:

- respond to risk assessment requests from eSafety.
- notify eSafety where a change increases the risk that the service could be used to generate R18+, X18+ or RC material
- keep records of the actions taken to comply with the standard.

Tier 2 services (with a medium risk profile) have fewer obligations than Tier 1 but more than Tier 3. Their key obligations include:

- having and enforcing policies that prohibit end-users from using the service to solicit, access, generate, distribute or store class 1A material or class 1B material.

- providing in-service reporting tools and responding promptly to end-user reports with appropriate action.
- responding to notices from eSafety requiring compliance reports.

Why are there specific obligations on generative AI services?

Generative AI is already being used to generate highly harmful content such as deepfake child sexual exploitation material which falls within eSafety's regulatory remit.[9] For this reason, the DIS Standard puts in place specific obligations for generative AI services which are at high risk of being used to generate X18+ and RC material, as well as for key distributors of high-risk models.

Designated internet service providers should also be aware of the expectations that apply to them under the new Basic Online Safety Expectations Determination, including the expectation that end-user safety is considered in the design and operation of generative AI.

Additional obligations on generative AI services may be proposed by the Australian Government to ensure the development and use of these services is safe and responsible, reducing risks of harms. The Department of Industry, Science and Resource (DISR) is considering mandatory guardrails focused on transparency, testing and accountability of services including using generative AI. eSafety continues to work with DISR to ensure alignment across our respective workstreams, which have different timescales and focus areas.

When do services have to comply with the DIS Standard?

The DIS Standard will come into effect on 22 December 2024[*], six months after its registration by the Office of Parliamentary Counsel.

[*] Date corrected 26 June 2024.

[9] For background information on generative AI and the online safety risks associated with this technology, see eSafety's Tech Trends position statement on generative AI (2023).

What are the consequences of not being compliant?

In the event of non-compliance, eSafety can issue a formal warning, issue an infringement notice, accept an enforceable undertaking, or seek an injunction or civil penalty in court.

After the DIS Standard takes effect, if you are concerned that one or more of the obligations are not being complied with, you can make a complaint to eSafety through our website at: eSafety.gov.au/industry/codes/complaints. Complaints help eSafety to identity potential non-compliance by a service and understand wider issues around compliance with the industry codes and standards, so we can help to keep Australians safer online.

However, eSafety cannot use the industry codes and standards to resolve disputes between online services and end-users.

If you have encountered illegal content online, we encourage you to lodge a report with eSafety so this content can be removed. For more information on reporting, visit: eSafety.gov.au/report.

Where can I find more information?

eSafety will publish regulatory guidance on the DIS Standard prior to its commencement. All relevant information on the industry codes and standards can be found on our website at: eSafety.gov.au/industry/codes.

Industry participants may seek guidance and information from eSafety if they are unsure which industry codes or standards apply to them. However, eSafety cannot provide legal advice and industry participants concerned about their compliance with the DIS Standard should seek their own legal advice.

Industry participants can contact eSafety at: codes@eSafety.gov.au.

eSafety.gov.au