Schedule 4 – Social Media Services (Core Features) Online Safety Code (Class 1C and Class 2 Material)

1 Structure

This Code is comprised of the terms of this Schedule together with the Online Safety Code (Class 1C and Class 2 Material) Head Terms (**Head Terms**).

2 Scope

(a) This Code applies to the provider of a social media service, so far as materials on that service are provided to Australian end-users.

(b) Social media services include a wide variety of unique services from community-based services with a local user base to larger platforms with international user bases.

(c) Social media services may include social networks, public media sharing networks, discussion forums, and consumer review networks, to the extent that they satisfy the criteria of a social media service as outlined in the OSA.

(d) If a social media service includes a messaging feature, this Code does not apply to the messaging feature.

3 Definitions

3.1 General

Unless otherwise indicated, terms used in this Code have the meanings given in the Head Terms or as set out in this clause 3.

3.2 Definition of social media service

(a) **social media service** means an electronic service that:

(i) satisfies the following conditions:

(A) the sole or primary purpose of the service is to enable online social interaction between 2 or more end-users;

(B) the service allows end-users to link to, or interact with, some or all other end-users;

(C) the service allows end-users to post material on the service;

(D) such other conditions (if any) as are set out in the legislative rules; or

(ii) is an electronic service specified in the legislative rules;

but does not include an exempt service (as defined by clause 3.2(c)).

Note: Online social interaction does not include (for example) online business interaction.

(b) **online social interaction** includes online interaction that enables end-users to share material for social purposes.

Note: Social purposes does not include (for example) business purposes.

(c) A service is an **exempt service** if:

(i) none of the material on the service is accessible to, or delivered to, one or more end-users in Australia; or

(ii) the service is specified in the legislative rules made under the OSA.

(d) In determining whether the condition set out in clause 3.2(a)(i)(A) is satisfied, disregard any of the following purposes:

(i) the provision of advertising material on the service;

(ii) the generation of revenue from the provision of advertising material on the service.

3.3 Other definitions

(a) **messaging feature** means an instant messaging feature of a social media service that enables private communication between two or more end-users of the service;

Note: A feature that enables end-users to (i) post material to their followers or community on the service or (ii) post comments in association with other content posted on a social media service, is not an instant messaging feature. These features will still be part of the social media service, but will not be treated as a 'messaging feature' under this Code.

4 Risk profile

4.1 Risk assessment: AI companion chatbot features

(a) Subject to clause 4.1(b) and clause 4.4, the provider of a social media service that includes an AI companion chatbot feature must undertake a risk assessment in respect of each generative AI restricted category of material in accordance with the following tables (as applicable):

If the risk that online pornography will be generated using the AI companion chatbot feature by Australian children is…	**the risk profile of the AI companion chatbot feature in relation to online pornography is …**
High	Tier 1
Moderate	Tier 2
Low	Tier 3

If the risk that high impact sexually explicit material will be generated using the AI companion chatbot feature by Australian children is…	**the risk profile of the AI companion chatbot feature in relation to high impact sexually explicit material is …**
High	Tier 1
Moderate	Tier 2
Low	Tier 3

If the risk that self-harm material will be generated using the AI companion chatbot feature by Australian children is…	the risk profile of the AI companion chatbot feature in relation to self-harm material is …
High	Tier 1
Moderate	Tier 2
Low	Tier 3

If the risk that high impact violence material will be generated using the AI companion chatbot feature by Australian children is…	the risk profile of the AI companion chatbot feature in relation to high impact violence material is …
High	Tier 1
Moderate	Tier 2
Low	Tier 3

If the risk that violence instruction material will be generated using the AI companion chatbot feature by Australian children is…	the risk profile of the AI companion chatbot feature in relation to violence instruction material is …
High	Tier 1
Moderate	Tier 2
Low	Tier 3

(b) If the sole or predominant purpose of an AI companion chatbot feature is to generate material in a particular generative AI restricted category, then:

(i) the service provider is not required to undertake a risk assessment in relation to that category and will automatically have a Tier 1 risk profile in respect of that category; and

(ii) is still required to undertake a risk assessment in relation of other categories.

Note: For example, if an AI companion chatbot feature has the sole or predominant purpose of enabling end-users to generate high impact violence material, the service provider will not be required to undertake a risk assessment in respect of that material. That feature will be required to comply with the measures for features with a Tier 1 risk profile, but only in respect of high impact violence material and not the other categories of material it does not have the sole or predominant purpose in respect of. The service provider will still need to conduct a risk assessment in respect of other relevant generative AI restricted categories of material to determine its risk profile in respect of those categories.

4.2 Risk assessment: other features

(a) How this Code applies to a social media service, other than in relation to any AI companion chatbot feature, depends on the risk that Australian children will access or be exposed to

online pornography, self-harm material or high-impact violence material on the social media service as follows:

(i) if the posting of online pornography, self-harm material or high-impact violence material is allowed under the applicable terms of use for the social media service, then the service provider will need to comply with compliance measures for that material as set out in clause 6 and the table in clause 7; and

(ii) if the posting of online pornography, self-harm material or high-impact violence material is not allowed under the applicable terms of use for the social media service, then the service provider must assess the risk that material in those categories will be accessed, distributed or stored by an Australian child on that service. Taking into account the outcome of the risk assessment, the service provider will need to comply with compliance measures for that material as set out in clause 6 and the table in clause 8.

Note: The scope of any risk assessment that is required may vary depending on the treatment of online pornography, self-harm material or high-impact violence material. For example, if the terms of use for a social media service permit the posting of online pornography but do not permit the posting of self-harm material, then the service provider would not need to carry out a risk assessment for online pornography but would need to carry out a risk assessment for self-harm material.

(b) Subject to clause 4.4 and except where the service provider chooses to automatically apply a Tier 1 risk profile in accordance with with section 5.2(a)(ii) of the Head Terms, if the provider of a social media service is obliged under clause 4.1 to carry out a risk assessment in relation to online pornography, self-harm material or high-impact violence material, the service provider will determine a risk profile for the service in accordance with the following tables (as applicable):

If the risk that Australian children will access or be exposed to online pornography material on a service is …	**the risk profile of the service in relation to online pornography is …**
High	Tier 1
Moderate	Tier 2
Low	Tier 3

If the risk that Australian children will access or be exposed to self-harm material on a service is …	**the risk profile of the service in relation to self-harm material is …**
High	Tier 1
Moderate	Tier 2
Low	Tier 3

If the risk that Australian children will access or be exposed to high-impact violence material on a service is …	the risk profile of the service in relation to high-impact violence material is …
High	Tier 1
Moderate	Tier 2
Low	Tier 3

For the avoidance of doubt, where a risk assessment is required under this clause in relation to a social media service that includes a messaging feature, the messaging feature will not be considered as part of the risk assessment and the risk assessment will only apply in relation to the other features of the service.

4.3 Methodology used for risk assessment and documentation

If a risk assessment is required under this Code, the provider of the relevant social media service must:

(a) be able to reasonably demonstrate that the provider's risk assessment methodology is based on reasonable criteria which must at a minimum include criteria relating to the functionality, purpose and scale of the social media service (including whether the service enables end-users in Australia to post or share material and any generative AI features of the service) or AI companion chatbot feature (as applicable) and, to the extent reasonably relevant, the additional requirements set out in clause 5 and any other criteria that are reasonably relevant for the purposes of determining the risk profile of the social media service under this Code;

(b) formulate in writing a plan and methodology for carrying out the risk assessment that ensures that each risk factor is accurately accessed;

(c) carry out the risk assessment in accordance with the plan and methodology prepared under clause 4.3(a), and by persons with the relevant skills, experience and expertise; and

(d) as soon as practicable after determining the risk profile of a social media service or AI companion chatbot feature (as applicable), the provider of the service must record in writing:

(i) details of the determination; and

(ii) details of the conduct of any related risk assessment,

sufficient to demonstrate that they were made or carried out in accordance with this Code. The record must include the reasons for the results of the assessment and the determination.

The service provider may carry out a single risk assessment covering each relevant category of material at once, provided that a separate risk profile is assessed for each category.

4.4 Certain categories of social media service are not required to undertake a risk assessment

A provider of a social media service that meets the following requirements is deemed to have a Tier 3 risk profile under this Code for online pornography, self-harm material and high-impact violence material without any further risk assessment being required:

(a) a social media service with the purpose of enabling 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; and

(b) a social media service that does not enable Australian end-users to do any of the following:

(i) create a list of other end-users with whom an individual shares a connection within the system; or

(ii) view and navigate a list of other end-user's individual connections; or

(iii) construct a public or semi-public profile within a bounded system created by the service.

4.5 Changes to risk profile of a social media service

If a provider of a social media service:

(a) makes a change to the service such that it would no longer be exempt from carrying out a risk assessment under clause 4.1 and / or 4.2 (as applicable); or

(b) has previously carried out a risk assessment, but makes a change to its service that would result in the service falling within a higher risk tier,

it must carry out a risk assessment in accordance with clause 4.1 and / or 4.2 (as applicable) as soon as practicable and in any case no later than 6 months after the relevant change takes effect.

5 Risk assessment: requirements

(a) This clause 5 applies where a provider of a social media service is required to undertake a risk assessment under in accordance with clause 4.1 and / or 4.2 (as applicable).

(b) A provider of a social media service must take into account the following matters when undertaking a risk assessment of an AI companion chatbot feature of the service in accordance with clause 4.1 or the other features of the service in accordance with clause 4.2, so far as they are relevant to the service:

(i) in the case of an AI companion chatbot feature:

(A) whether any generative AI restricted category of material is permitted on the feature and if so, the likely portion of that content as compared with other types of content;

(B) the likelihood that the feature may be used to directly expose Australian children to any generative AI restricted category of material;

(C) the likelihood that an Australian child will use the feature to access any generative AI restricted category of material;

(ii) other than in the case of an AI companion chatbot feature:

(A) the terms or arrangements under which the provider acquires any content to be made available on the service;

(B) the likelihood that the service may be used to directly expose an Australian child to online pornography, self-harm material and high-impact violence material (as applicable);

(C) the likelihood that an Australian child will use the service to access online pornography, self-harm material and high-impact violence material (as applicable);

(D) the functionality and features of the service, including any generative AI functionality or features (other than any AI companion chatbot feature);

Note: A service that provides an integrated chat or messaging function should be regarded as higher risk than a service without those features.

(iii) in relation to the social media service as a whole (including any AI companion chatbot feature):

(A) the terms of use for the service;

(B) the ages of end-users and likely end-users of the service;

(C) the likelihood that a significant number of Australian children will access the service;

(D) the number of Australian end-users that are monthly active account holders;

Note: A service with a large number of Australian end-users that are monthly active account holders should be regarded as higher risk than a service with fewer such account holders.

(E) the number of Australian children that are monthly active account holders;

Note: A service with a large number of Australian children that are monthly active account holders should be regarded as higher risk than a service with fewer such account holders.

(F) the primary purpose of the service;

Note: A service.with the primary purpose of enabling general social interaction should be regarded as higher risk than a service with the primary purpose of enabling social interaction within a limited user group (such as a particular school, neighbourhood or enterprise) or for a limited purpose (such as to enable users to post reviews of products and services or for a limited commercial or public purpose such as the crowdfunding of commercial or charitable activities or social causes or to start an online petition for social change).

(G) a forward-looking analysis of:

(aa) likely changes to the operating environment for the service including likely changes in the functionality or purpose of, or the scale of, the service; and

(ab) the impact of those changes on the ability of the service provider to meet the online safety objectives that apply under this Code;

(H) safety by design guidance and tools published or made available by a relevant government agency or a foreign or international body;

Note: Examples of relevant agencies and bodies are eSafety and the Digital Trust & Safety Partnership.

(I) relevant international laws and regulations applicable to the service that address online safety risks and harms similar to those addressed in this Code; and

(J) where applicable, design features and controls deployed to mitigate relevant risks.

Note: Without limiting this clause 5(b), circumstances in which a matter will not be relevant to a service include where it is not relevant to the risk level of the service in the circumstances, relates to a topic that is irrelevant to the particular service due to its nature, or requires consideration of information that is not available for the service.

6 Approach to measures and guidance for social media services

(a) The tables in sections 7, 8, 9 and 10 below contain mandatory compliance measures for providers of social media services under this Code, as follows:

(i) the table in section 7 sets out compliance measures that apply to the extent that online pornography, self-harm material or high-impact violence material is allowed to be posted on the social media service under the applicable terms of use, but do not apply to any messaging feature or AI companion chatbot feature;

(ii) the table in section 8 sets out compliance measures that apply to the extent online pornography, self-harm material or high-impact violence material is not allowed to be posted on a social media service under the applicable terms of use where the service has a Tier 1 or Tier 2 risk profile for online-pornography, self-harm material or high-impact violence material, but do not apply to any messaging feature or AI companion chatbot feature;

(iii) the table in section 9 sets out compliance measures that apply to all social media services that allow online pornography, self-harm material or high-impact violence material and to other social media services with a Tier 1 or Tier 2 risk profile for online pornography, self-harm material or high-impact violence material, but do not apply to any messaging feature; and

(iv) the table in section 10 sets out compliance measures that apply to AI companion chatbot features. These measures apply in addition to compliance measures in the table in section 9 (where applicable). To the extent there is any overlap in the measures in the tables in sections 9 and 10, a single action by the service provider may be sufficient to satisfy both measures.

(b) The tables also include guidance on the implementation of some measures. This guidance is not intended to be binding on providers but to guide them on the way in which they may choose to implement a measure.

(c) Certain compliance measures only apply to certain categories of material (as specified in the column titled 'Material') or to service providers who meet a certain risk profile in relation to a designated category of material (as specified in the column titled 'Risk Tier'), as specified in the relevant table.

Note: A service provider may have a different risk profile in respect of different categories of material. For example, an AI companion chatbot feature may have a Tier 1 risk profile for online pornography but a Tier 2 or Tier 3 risk profile for all other generative AI restricted categories. In that case, the Tier 1 compliance measures for the AI companion chatbot feature will only apply in relation to online pornography.

7 Compliance measures where online pornography, self-harm material or high-impact violence material is allowed

The compliance measures in this table apply to the extent online pornography, self-harm material or high-impact violence material is allowed to be posted on a social media service under the applicable terms of use, but do not apply to any messaging feature or any AI companion chatbot feature. Each measure applies to the type of class 1C and class 2 material specified in the 'Material' column for that measure.

No.	Material	Compliance measure
7.1	online pornography self-harm material	**Age assurance measures** A service provider must, where technically feasible and reasonably practicable, implement: (a) appropriate age assurance measures; and (b) access control measures, before providing access to online pornography and/or self-harm material. A service provider must also take appropriate steps to test and monitor the effectiveness of its age assurance and access control measures over time.
7.2	online pornography self-harm material high-impact violence material	**Safety tools** Except where the primary purpose of the service is to provide access to online pornography, self-harm material and/or high-impact violence material, a service provider must provide appropriate safety tools to Australian end-users which may limit their access or exposure to online pornography, self-harm material and/or high-impact violence material on the service and are appropriate for the service. The service provider must ensure that such safety tools are defaulted to an appropriate setting for Australian child end-users in relation to high-impact violence material, but can otherwise provide such tools to Australian end-users on an opt-in basis. Appropriate safety tools may include solutions for: (a) implementing age-gates, either on the entire service or on identified areas of services where an end-user is most likely to access or be exposed to online pornography, self-harm material and/or high-impact violence material on the service; (b) filtering online pornography, self-harm material and high-impact violence material, including by downlisting, deprioritising or quarantining; (c) blocking online pornography, self-harm material and high-impact violence material; (d) blurring online pornography, self-harm material and high-impact violence material; (e) halting autoplay of online pornography, self-harm material and high-impact violence material;

No.	Material	Compliance measure
		(f) placing interstitial notices on online pornography, self-harm material and high-impact violence material so that users can click through to view if they wish; (g) ensuring that recommender systems, algorithms, and other choice architecture, do not promote online pornography or self-harm material to child end-users; (h) ensuring compatibility with third-party filtering software or tools which may be installed on devices, or provided by internet carriage services. ***Guidance**:* *Appropriate safety tools may vary depending on the type of service, the typical demographic of users on the service, the type of material allowed on the service, and technical and other limitations that may apply.*
7.3	online pornography self-harm material high-impact violence material	**Publishing information about tools and settings** To the extent relevant, a service provider must publish clear and accessible information to Australian end-users about the tools and settings available to limit their access or exposure to online pornography, self-harm material and high-impact violence material in their news and discovery feed.
7.4	online pornography self-harm material high-impact violence material	**Annual reporting to eSafety on Code compliance** A service provider must submit to eSafety a Code report which includes the following information: (a) the steps that the provider has taken to comply with the compliance measures under this Code; and (b) an explanation as to why these steps are appropriate. The first Code report must be submitted by the provider of the social media service to eSafety 12 months after this Code comes into effect. The provider of the social media service must submit subsequent Code reports to eSafety annually. A report under this compliance measure may be combined with any report that the service provider is obliged to provide under any other compliance measure.

8 Compliance measures where online pornography, self-harm material or high-impact violence material is not allowed

The compliance measures in this table apply to the extent online pornography, self-harm material and high-impact violence material is not allowed to be posted on a social media service under the applicable terms of use where the service has a Tier 1 or Tier 2 risk profile for online pornography, self-harm material or high-impact violence material, but do not apply to any messaging feature or AI companion chatbot feature. Each measure applies to services in the risk tier specified in the 'Risk Tier' column and to the type of class 1C and class 2 material specified in the 'Material' column for that measure.

No.	Risk Tier	Material	Compliance measure
8.1	Tier 1 or Tier 2 for online pornography	online pornography	**Use of systems, processes and/or technologies to detect and remove online pornography** A service provider must implement systems, processes and/or technologies designed to detect, flag and/or remove online pornography from the service, for example, through the use of key word searches, hashing, machine learning, artificial intelligence, or other technology designed to identify text, videos and images that may, depending on the context, be online pornography and/or other safety technologies or systems or processes that limit users' exposure to such material on the service. A service provider must also take appropriate steps to continuously improve these systems, processes and/or technologies. ***Guidance**:* *In implementing this measure, service providers should carefully consider the appropriateness of systems, processes and/or technological tools for their services. These may include, but are not limited to, systems, processes and/or tools that scan for hashed materials and/or behavioural or text signals and/or patterns that signal or are associated with online pornography. Providers should consider the appropriateness of different options and the capability of the provider to use those options accurately, including the need for systems and processes that, where appropriate, prioritise materials detected for human review. The rights and expectations of legitimate users of social media services are also important factors for providers to consider when considering the type of approach that is appropriate for a particular service.*
8.2	Tier 1 or Tier 2 for self-harm material	self-harm material	**Use of systems, processes and/or technologies to detect and remove self-harm material** A service provider must implement systems, processes and/or technologies designed to detect, flag and/or remove self-harm material from the service, for example, through the use of key word searches, hashing, machine learning, artificial intelligence, or other technology designed to identify text, videos and images that

No.	Risk Tier	Material	Compliance measure
			may, depending on the context, be self-harm material and/or other safety technologies or systems or processes that limit users' exposure to such material on the service. A service provider must also take appropriate steps to continuously improve these systems, processes and/or technologies. ***Guidance**:* *In implementing this measure, service providers should carefully consider the appropriateness of systems, processes and/or technological tools for their services. These may include, but are not limited to, systems, processes and/or tools that scan for hashed materials and/or behavioural or text signals and/or patterns that signal or are associated with self-harm material. Providers should consider the appropriateness of different options and the capability of the provider to use those options accurately, including the need for systems and processes that, where appropriate, prioritise materials detected for human review. The rights and expectations of legitimate users of social media services are also important factors for providers to consider when considering the type of approach that is appropriate for a particular service.*
8.3	Tier 1 or Tier 2 for high-impact violence material	high-impact violence material	**Use of systems, processes and/or technologies to detect and remove high-impact violence material** A service provider must implement systems, processes and/or technologies designed to detect, flag and/or remove high-impact violence material from the service, for example, through the use of key word searches, hashing, machine learning, artificial intelligence, or other technology designed to identify text, videos and images that may, depending on the context, be high-impact violence material and/or other safety technologies or systems or processes that limit users' exposure to such material on the service. A service provider must also take appropriate steps to continuously improve these systems, processes and/or technologies. ***Guidance**:* *In implementing this measure, service providers should carefully consider the appropriateness of systems, processes and/or technological tools for their services. These may include, but are not limited to, systems, processes and/or tools that scan for hashed materials and/or behavioural or text signals and/or patterns that signal or are associated with high-impact violence material. Providers should consider the appropriateness of different options and the capability of the provider to use those options accurately, including the need for systems and processes that, where appropriate, prioritise materials detected for human review. The rights and expectations of legitimate users of social media services are also important factors for*

No.	Risk Tier	Material	Compliance measure
			providers to consider when considering the type of approach that is appropriate for a particular service.
8.4	Tier 1 or Tier 2 for online pornography, self-harm material or high-impact violence material	online pornography self-harm material high-impact violence material	**Reporting to eSafety on Code compliance** Where eSafety issues a written request to a service provider to submit a Code report, the provider named in such request must submit to eSafety a Code report which includes the following information: (a) details of any risk assessment it is required to undertake pursuant to this Code in relation to online pornography, self-harm material or high-impact violence material (as applicable); (b) the steps that the provider has taken to comply with the compliance measures under this Code; and (c) an explanation as to why these steps are appropriate. A service provider that has received such a request from eSafety is required to submit a Code report within 2 months of receiving the request, but for the first request no earlier than 12 months after this Code comes into effect. A service provider will not be required to submit a Code report to eSafety more than once in any 12-month period. A report under this compliance measure may be combined with any report that the service provider is obliged to provide under any other compliance measure.

9 Other supporting compliance measures

The compliance measures in this table apply to all social media services that allow online-pornography, self-harm material or high-impact violence material and to other social media services with a Tier 1 or Tier 2 risk profile for online-pornography, self-harm material or high-impact violence material, but do not apply to any messaging feature. Each measure applies to the type of class 1C or class 2 material specified in the 'Material' column for that measure.

No.	Material	Compliance measure
9.1	online pornography self-harm material high-impact violence material simulated gambling material	**Terms and conditions relating to class 1C and class 2 material** A service provider must have, and enforce, clear actions, policies or terms and conditions relating to online pornography, self-harm material, high-impact violence material and simulated gambling material, which will include, to the extent applicable, terms and conditions dealing with the types of online pornography, self-harm material, high-impact violence material and simulated gambling material that are allowed or not allowed on the social media service. In implementing this measure, the service provider must: (a) use simple, plain, and straightforward language; (b) to the extent practicable, be clear about the type of any material that is prohibited; and (c) communicate such terms and conditions, standards and/or policies to all personnel that are directly involved in their enforcement. Relevant policies and actions must be implemented according to a graduated, risk-based approach. This approach may be different for different types of material.
9.2	All	**Trust and safety function** A service provider must have, or have access to, sufficient personnel to oversee the safety of the service. Such personnel must have the skills, experience and qualifications needed to ensure that the provider complies with the requirements of this Code at all times.
9.3	online pornography self-harm material high-impact violence material simulated gambling material	**Reporting mechanisms** A service provider must provide tools which enable Australian end-users to report class 1C and class 2 material which they consider may be contrary to the social media service's terms and conditions, and must where appropriate ensure that these reports are evaluated and actioned. Such reporting mechanisms must: (a) be easily accessible and easy to use; (b) be accompanied by clear instructions on how to use them;

No.	Material	Compliance measure
		(c) ensure that the identity of the reporter is not disclosed to the reported end-user or account holder (i.e., the individual who has been reported should not be able to see the person who reported them), without the reporter's express consent, except as required by law.
9.4	online pornography self-harm material high-impact violence material simulated gambling material	**On-platform reporting tools** A service provider must ensure that the reporting tools referred to in compliance measure 9.3 for class 1C and class 2 material are available and accessible to Australian end-users on the interface of the social media service. ***Guidance**:* *In implementing these measures, providers of a social media service should ensure that reporting tools are integrated within the functionality of the social media service in a manner that is visible and accessible at the point the Australian end-user accesses materials posted by other end-users.*
9.5	online pornography self-harm material high-impact violence material simulated gambling material	**Complaints tools** A service provider must provide tools which enable Australian end-users to make a complaint about: (a) the provider's handling of reports about class 1C or class 2 material; or (b) any other aspect of the provider's compliance with this Code. Such complaints tools must: (a) be easily accessible and simple to use; and (b) be accompanied by plain language instructions on how to use them.
9.6	online pornography self-harm material high-impact violence material simulated gambling material	**Appropriate steps for informing Australian end-users about actions taken on reports and complaints** A service provider must take appropriate steps to acknowledge a report referred to in compliance measure 9.3 or complaint referred to in compliance measure 9.5 and must ensure that an Australian end-user who makes such a report or complaint is informed in a reasonably timely manner of the outcome of the report or the complaint, and of any review mechanisms that are available, or is otherwise able to access information about the status of the report or the complaint. ***Guidance**:* *The way a service provider implements this measure and the timeliness of the actions required under this measure will depend on the type of material reported, the likelihood of harm that it poses to Australian end-users, the source of the report and the risk profile of the provider of the social media service.*
9.7	online pornography	**Training for personnel responding to reports and complaints**

No.	Material	Compliance measure
	self-harm material high-impact violence material simulated gambling material	A service provider must ensure that personnel responding to reports referred to in compliance measure 9.3 or complaints referred to in compliance measure 9.5 are trained in the social media service's policies and procedures for dealing with such reports and complaints.
9.8	online pornography self-harm material high-impact violence material simulated gambling material	**Reviews of compliance of personnel with systems and processes** A service provider must review the effectiveness of its reporting systems and processes to ensure reports and complaints are assessed and actioned (if necessary) within reasonably expeditious timeframes, based on the level of harm the material poses to Australian children. Such review must occur at least annually. ***Guidance:*** *This could include review and analysis of data collected for the year (eg responses and outcomes) as well as submitting test reports via the contact mechanism to review handling and response.*
9.9	online pornography self-harm material high-impact violence material simulated gambling material	**Timely referral of unresolved complaints to eSafety** A service provider must promptly refer to eSafety complaints from Australian end-users concerning a material non-compliance with this Code by the service provider, where the service provider is unable to resolve the complaint within a reasonable timeframe.
9.10	online pornography self-harm material high-impact violence material simulated gambling material	**Updates to eSafety about relevant changes to technology** A service provider must take reasonable steps to ensure eSafety receives updates regarding significant changes to the functionality of their services that are likely to have a material positive or negative effect on the access or exposure to, distribution of, or online storage of online pornography, self-harm material, high-impact violence material or simulated gambling material by an Australian child. A service provider may choose to provide this information in an annual report to eSafety under this Code. In implementing this measure, a service provider is not required to disclose information to eSafety that is confidential. ***Guidance****:* *Changes that have a material negative effect should, ideally be communicated before a public announcement of the relevant changes.*

No.	Material	Compliance measure
9.11	All	**Engagement** A service provider must appropriately engage with safety and community organisations (such as civil society groups, public interest groups and representatives of marginalised communities), academics and government to gather information to help inform measures taken for the purposes of protecting or preventing children from accessing or being exposed to class 1C and class 2 material. A service provider must consider information obtained through such engagement. ***Guidance**:* *Engagement may occur within and/or outside Australia as relevant to the issue under consideration.* *Engagement may occur regularly in the course of ongoing relationships with organisations, academics or government, during development of new service features or in other appropriate circumstances.*
9.12	All	**Information for Australian end-users about the role and functions of eSafety, including how to make a complaint to eSafety** A service provider must publish clear information that is accessible to Australian end-users which explains the role and functions of eSafety, including how to make a complaint to eSafety.
9.13	All	**Location on service that is dedicated to providing online safety information** A service provider must establish a location on or via the service that is dedicated to providing online safety information, that: (a) contains information required under this Code; (b) includes information about how Australian end-users can contact third party services that may provide counselling and support; and (c) is accessible to Australian end-users. ***Guidance**:* *A provider could raise Australian end-users' awareness about the availability of safety information on its platform in relation to its services, through interstitial mechanisms such as account notifications, on-platform advertising campaigns or pop-up notices when material is being posted or viewed by Australian end-users. Providers could also contribute to off-platform campaigns targeted at the general public, Australian end-users or specific sections of the community such as teachers, parents and carers, older users or vulnerable groups. A provider could also contribute to an off-platform campaign by providing financial assistance, advertising collateral, expert advisers, or other support services.*

No.	Material	Compliance measure
9.14	online pornography self-harm material high-impact violence material simulated gambling material	**Information about how services deal with risk of harm** A service provider must publish clear and accessible information that explains the actions they take to reduce the risk of harm to Australian child end-users from online pornography, self-harm material, high-impact violence material and simulated gambling material on its service.

10 Compliance measures for AI companion chatbot features

The compliance measures in this table apply to all AI companion chatbot features included as part of a social media service. An AI companion chatbot feature that meets the risk tier specified in the "Risk Tier" column in respect of a generative AI restricted category of material, must comply with that measure for the relevant generative AI restricted category of material. An AI companion chatbot feature may have a different risk profile for each generative AI restricted category of material. For example, an AI companion chatbot feature may have a Tier 1 risk profile for online pornography but a Tier 2 or Tier 3 risk profile for all other generative AI restricted categories. In that case, the Tier 1 compliance measures for the AI companion chatbot feature will only apply in relation to online pornography.

No.	Risk Tier	Compliance measure
10.1	Tier 1	**Age assurance measures** A service provider must, where technically feasible and reasonably practicable, implement: (a) appropriate age assurance measures; and (b) access control measures, before providing access to the feature or being able to generate the generative AI restricted category of material. A service provider must also take appropriate steps to test and monitor the effectiveness of its age assurance and access control measures over time. Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 risk profile.
10.2	Tier 2	**Safety by design defaults – generative AI restricted categories of material** A service provider must either: (a) implement the age assurance and access controls measures outlined in measure 10.1 above before providing access to the feature, or being able to generate a generative AI restricted category of material; or (b) comply with the following: (i) implement systems, processes and/or technologies that prevent the feature from being used to generate outputs that contain a generative AI restricted category of material; (ii) regularly review and test models on the potential risk that model is used to generate a generative AI restricted category of material; and (iii) promptly following review and/or testing, adjust models and deploy mitigations with the aim of reducing the misuse and unintentional use of models to generate a generative AI restricted category of material.

No.	Risk Tier	Compliance measure
		Guidance: *A requirement to put in place systems, processes, and/or technologies to prevent the feature from being used to generate outputs that contain a generative AI restricted category of material should take account of the fact that not all AI companion chatbot feature providers will always have sufficient visibility and control of their models—if a provider lacks that visibility or control of certain aspects so that it cannot deploy all mitigations, it will have to rely on other systems, processes and technologies that are available.* Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 2 risk profile.
10.3	Tier 1 and Tier 2	**Terms and conditions** A service provider must have, and enforce, clear actions, policies or terms and conditions relating to the generation of generative AI restricted categories of material, which will include, to the extent applicable, terms and conditions dealing with whether any type of generative AI restricted category of material is permitted to be generated using the feature. In implementing this measure, the service provider should: (a) use simple, plain, and straightforward language; (b) to the extent practicable, be clear about the type of any material that is prohibited; and (c) communicate such terms and conditions, standards and/or policies to all personnel that are directly involved in their enforcement. Relevant policies and actions should be implemented according to a graduated, risk-based approach. This approach may be different for different types of material. Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 or Tier 2 risk profile.
10.4	Tier 1 and Tier 2	**Reporting mechanisms** A service provider must provide tools which enable Australian end-users to report a generative AI restricted category of material generated using the feature which they consider may be contrary to the social media service's terms and conditions, and must where appropriate ensure that these reports are evaluated and actioned. Such reporting mechanisms must: (a) be easily accessible and easy to use; (b) be accompanied by clear instructions on how to use them; and

No.	Risk Tier	Compliance measure
		(c) ensure that the identity of a complainant is not disclosed to the reported end-user or account holder (i.e., the individual who has been reported should not be able to see the person who reported them), without the reporter's express consent, except as required by law. Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 or Tier 2 risk profile.
10.5	Tier 1	**On-platform reporting tools** A service provider must ensure that the reporting tools referred to in measure 7.5 above are available and accessible to Australian end-users on-the interface of the social media service. ***Guidance:*** *In implementing these measures, the provider should ensure that reporting tools are integrated within the functionality of the social media service in a manner that is visible and accessible at the point the Australian end-user generates materials.* Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 risk profile.
10.6	Tier 1 and Tier 2	**Information about how services deal with generative AI restricted categories of material** A service provider must publish clear and accessible information that explains the actions they take to reduce the risk of harm to Australian children caused by the generation of a generative AI restricted category of material using the feature. Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 or Tier 2 risk profile.
10.7	Tier 1 and Tier 2	**Updates to eSafety about relevant changes to technology** A service provider must take reasonable steps to ensure eSafety receives updates regarding significant changes to the functionality of their AI companion chatbot feature that are likely to have a material positive or negative effect on the risk of generation of a generative AI restricted category of material by an Australian child. A service provider may choose to provide this information in an annual report to eSafety under this Code. In implementing this measure, a provider is not required to disclose information to eSafety that is confidential. ***Guidance****:* *Changes that have a material negative effect should, ideally be communicated before a public announcement of the relevant changes.*

No.	Risk Tier	Compliance measure
		Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 or Tier 2 risk profile.
10.8	Tier 1	**Complaints tools** The service provider must provide tools which enable Australian end-users to make a complaint about: (a) the provider's handling of reports about a generative AI restricted category of material that is generated via the AI companion chatbot feature, except where the sole or predominant purpose of the feature is to provide access to the generative AI restricted category of material; and (b) any other aspect of the provider's compliance with this Code. Such complaints tools must: (a) be easily accessible and simple to use; (b) be accompanied by plain language instructions on how to use them. Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 risk profile.
10.9	Tier 1 and Tier 2	**Training for personnel responding to reports and complaints** A service provider must ensure that personnel responding to reports referred to in compliance measure 10.4 or complaints referred to in compliance measure 10.8 are trained in the social media service's policies and procedures for dealing with such reports and complaints. Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 risk profile.
10.10	Tier 1	**Significant changes to services** A service provider must, before it makes a material change to the AI companion chatbot feature that is likely to significantly increase the risk of enabling an Australian child to generate the generative AI restricted category of material: (a) carry out an assessment of the kinds of features and settings that could reasonably be incorporated into the feature to minimise that risk; and (b) where appropriate, apply features and settings so identified to help to mitigate that risk. Note: A service provider will need to comply with this measure for the generative AI restricted category of material for which it has a Tier 1 risk profile.