eSafetyCommissioner

OFFICIAL

Commissioner Briefing: OpenAI

To	Julie Inman Grant, Commissioner
Cc	s 47E(c), s 47F, General Manager, Corporate and Strategy s 22 Executive Manager, Strategy, Engagement and Research s 22 Executive Manager Industry Compliance and Enforcement s 22 Manager, Industry Insights and Enablement s 22 Manager, Industry Supervision
From	s 22 Assistant Manager Industry, Insights and Enablement
Cleared by	s 22 Manager, Industry Insights and Enablement
Meeting Details	Tuesday 23rd September 2025 11am - 12 pm PDT In Person: 1515 3rd St. San Francisco, CA 94158
Subject	Meeting with s 47F, Head of Policy, APAC
Attachments	NIL

Purpose

- This briefing provides information for your meeting with s 47F, Head of Legal, APAC, OpenAI.

Representatives

External Attendees	Name and Title
	s 47F, Head of Policy, APAC OpenAI
eSafety Attendees	Julie Inman Grant, Commissioner s 47E(c), s 47F, GM Corporate & Strategy s 22 Executive Manager, Industry, Compliance and Enforcement
Note Taker	s 22, Executive Manager, Industry, Compliance and Enforcement

Commented [s 22]: Additional Open AI attendees TBC

Agenda

1 OpenAI to provide overview and updates on current online safety initiatives.

2 eSafety to provide update on key priorities and initiatives.

Recent interactions

1. **September 2024:** eSafety met with OpenAI's Child Safety Technical Program Manager, s 47F, and Head of Policy for Asia Pacific, s 47F. Notes from that meeting are here.
2. **Feb 2025:** eSafety met with OpenAI to discuss Safety by Design, approach to user safety and safety challenges as part of the Industry Insights and Enablement Team's generative AI engagement. Notes from that meeting are here.

Background

- You are scheduled to meet with s 47F (Head of Policy, APAC) and other Open AI representatives (tbc) on 23rd September.

Talking points and questions

Industry codes and standards

3. Inform OpenAI that eSafety are empowered through the Act to give providers transparency notices which compel information relating to how providers are implementing the BOSE and publish summaries of this information to publicly hold providers to account for the safety of their services.
4. Highlight that we are likely to focus on GenAI thematically through 2025-26 for future notices.
5. The BOSE provides specific expectations of providers regarding the use of generative artificial intelligence capabilities. These include:
    - Reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of generative artificial intelligence capabilities on the service;
    - Reasonable steps to proactively minimise the extent to which generative artificial intelligence capabilities may be used to produce material or facilitate activity that is unlawful or harmful.
6. Inform OpenAI that, as previously discussed, ChatGPT is subject to the Phase 1 DIS Standard.
    - ChatGPT is a text and image generative AI service that may be categorised as a 'high impact generative AI DIS' if the risk of generating high impact material (X18+ or Restricted Content) is not immaterial.

  - The same category applies for any consumer facing generative AI product, e.g. Sora.
  - Note however that the development of pre-trained models for integration into other downstream services are assessed in the DIS Standard as tier 3, which is the lowest risk, and so are subject to minimal obligations.

7. Inform OpenAI that the eSafety Commissioner registered 9 industry-drafted codes in relation to age-restricted material (the Phase 2 Codes). These Codes will begin to come into effect from 27 December 2025. The Designated Internet Services (DIS) Code comes into effect on 9 March 2026. Under the DIS Code:
   - High-risk generative AI DIS services that are capable of generating sexually explicit material, violence instruction material or self-harm material must implement appropriate age assurance measures to either a) prevent users under 18 from generating this material or b) prevent children from accessing the service.
   - Services with a moderate risk can also introduce appropriate age assurance measures, or else they must implement robust systems that prevent this material from being generated. Services must regularly review, test and adjust these systems to reduce the unintentional use of models to generate this material.
     - s 47E(d)
   - Both moderate and high-risk generative AI DIS have additional compliance measures under the Code, and must:
     - have and enforce clear terms and conditions relating to age-restricted material
     - provide tools for users to report, flag or make complaints about age-restricted material
     - have sufficient personnel to oversee the safety of the service
     - provide easily accessible and clear safety information for end-users
     - update eSafety about relevant changes to the functionality of the service
     - report to eSafety on Code compliance when requested.
   - Additionally, we are aware that OpenAI's product Sora has a community page that lets users share generated material. This will require the service to follow further compliance measures in the Phase 2 DIS Code:

  - Services that have high-risk of exposing children to online pornography or self-harm material must implement age assurance measures before letting users access this material.
  - As the Sora usage policy prevents the generation of online pornography or self-harm material, they must implement and improve systems that can detect and automatically action this material.

APAC market expansion

8. Ask about OpenAI's plans for their new Sydney office and efforts to support their users within Australia.
   a) In July 2025, OpenAI released a blueprint outlining opportunity areas for AI to deliver benefits to the Australian economy and society, and highlights how OpenAI can support this. The blueprint covers challenges and policy suggestions across productivity, education, government services and AI infrastructure.

Safety by Design

9. Ask about OpenAI's implementation of Safety by Design principles in any new product updates or developments.
   - OpenAI will introduce the following measures:
     - Parental controls that give parents options to gain more insight into, and shape, how their teens use ChatGPT.
     - Referring people to real-world resources and expanding crisis interventions
     - Escalating risk of physical harm to human review
     - Safeguards for long conversations
     - Enabling connections to trusted contacts
     - Refining how content is blocked.
10. Inform OpenAI about Industry Insights & Enablement team's upcoming AI SbD engagement, which aims to explore affective usage and psychological risks and impacts of generative AI usage with a range of services, including Open AI.
11. Ask how OpenAI supports the safety of users in relation to self-harm cases, and if there are foreseeable cases in which safety would be prioritised over privacy, e.g. where it may be reasonably necessary to contact law enforcement.
   - s 47G

Safe Completions Training

12. Ask how OpenAI will assess user's 'apparent' intent and navigate the possibility that users may be able to manipulate prompts for questions which are considered dual use.

- OpenAI has introduced Safe Completions training, which centres safety training on the safety of a model's output, rather than determining a refusal boundary according to the user's input. This teaches the model to give the most helpful answer where possible while staying within safety boundaries. This results in the model partially answering a user's question or only answering at a high level. If the model needs to refuse, GPT-5 is trained to transparently state why it is refusing, as well as to provide safe alternatives.
- s 47G
-

OpenAI & Anthropic Safety Testing Collaboration

13. Ask about future collaboration and focus areas for future safety testing.
    - OpenAI and Anthropic collaborated on a joint evaluation in which they ran internal safety and misalignment evaluations on each other's publicly released models.

Affective Impact

14. Ask how OpenAI aim to improve, track and assess ChatGPT usage to mitigate risks and impacts associated to affective and companionship usage such as influence risk, dependencies, stigmatisation, unhealthy relationships and psychological harm.
    - OpenAI's GPT-5 includes improvements related to supporting users with health-related questions. OpenAI states that it is more like an active thought partner, proactively flagging potential concerns and asking questions to give more helpful answers.
    - GPT-5 has had improvements in areas like avoiding unhealthy levels of emotional reliance, reducing sycophancy, and reducing the prevalence of non-ideal model responses in mental health emergencies by more than 25% compared to 4o.
    - Researchers at Stanford University have observed that therapy chatbots powered by large language models can sometimes stigmatize people with mental health conditions or respond in ways that are inappropriate or could be harmful.
    - OpenAI has convened a council of experts in youth development, mental health, and human-computer interaction referred to as the 'Expert Council on wellbeing and AI.' This council will work in tandem with OpenAI's Global Physician Network, a broader pool of more than 250 physicians to better measure capabilities of AI systems for health.

ATTACHMENT A

Biographies

s 47F [redacted] Head of Policy – Asia Pacific	s 47F [redacted]