Commissioner Briefing: Anthropic

To	Julie Inman Grant, Commissioner
Cc	s 22 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 9am -10 am PDT In Person: Anthropic Office, s 47G
Subject	Meeting with Anthropic and eSafety Commissioner
Attachments	NIL

Purpose

- This briefing provides information for your meeting with Jeff Bleich, General Counsel, Anthropic.

Representatives

External Attendees	**Name and Title**
	Jeff Bleich, General Counsel, Anthropic
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

Agenda

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

2 eSafety to provide update on key priorities and initiatives.

Recent interactions

1. **July 2024:** eSafety met s 47F for an introductory conversation and to discuss a Save the Children-led proposal for a project on Youth Voices on AI.
2. **March 2025:** eSafety met with Anthropic 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

3. You are scheduled to meet with **Jeff Bleich** (General Counsel, Anthropic) on 23rd September.

Talking points and questions

Industry codes and standards and BOSE

4. Inform Anthropic 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.
5. Highlight that we are likely to focus on GenAI thematically through 2025-26 for future notices.
6. 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.
7. Inform Anthropic that they are subject to the Phase 1 Designated Internet Services Standard.
    - Anthropic provides a consumer facing chatbot called Claude. It may be categorised as a 'High impact generative AI DIS' if the risk for generating high

impact material (X18+ or Restricted Content) is not immaterial, in which case significant obligations apply.

- The development of pre-trained models for integration into other downstream services are assessed as tier 3, which is the lowest risk, and so are subject to minimal obligations under the DIS Standard.

8. Inform Anthropic 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, or provide details of relevant risk assessments on request.

Safety by Design

9. Ask about Anthropic's implementation of Safety by Design principles in any new product updates or developments.

10. Inform Anthropic 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 Anthropic.

Threat intel report

11. Ask how Anthropic are addressing misuse and weaponisation of Claude to carry out online harm such as tech facilitated gender-based violence and sexual extortion.

    - Anthropic's recent threat intelligence report discusses examples of Claude being misused for cyber crime such as vibe hacking, extortion, fraud and malware as a service.

Affective impact

12. Ask how Anthropic aim to improve, track and assess Claude usage to mitigate risks and impacts associated with affective and companionship usage such as influence risk, dependencies, unhealthy relationships and psychological harm.

    - Anthropic aim to improve Claude so that it can provide tailored messages directing people to localised helplines in Australia if prompted.
    - Although Claude is not designed for emotional support and connection, there is affective usage. 2.9% of Claude interactions are affective conversations whereas companionship and roleplay combined comprise less than 0.5% of conversations.
    - Anthropic has a unified harms framework which discuss psychological harm as a harm category and has safeguards for Claude which follow a risk-based approach.
    - Claude can end harmful conversations. However, when Claude does this, users will still be able to start new conversations from the same account and create new branches of the conversation by editing their responses.

Voice mode for Claude

13. Ask how Anthropic will track and assess the nature, risks, harm and impact of voice-based interactions and how Anthropic is navigating limitations of limited language expertise for red teaming across all languages.

    a) Anthropic has rolled out a voice mode for Claude, which lets mobile apps users have audio conversations with Claude.

    b) Anthropic acknowledged current limitations of limited language expertise for comprehensive red teaming across all languages.

    c) Research states that affective topics are more apparent in voice-based conversations.

AI browser agent on Chrome

14. Ask if information will be collected and tracked, and if so what kind of information (e.g. usage, experiences of online harm), and how Anthropic will balance security, privacy and safety and provide transparency to users.

a) Anthropic is launching a research preview of a browser-based AI agent, called Claude for Chrome. By adding an extension to Chrome, select users can chat with Claude in a sidecar window that maintains context of everything happening in their browser. Users can also give the Claude agent permission to take actions in their browser and complete some tasks on their behalf.

b) Users can limit Claude's browser agent from accessing certain sites in the app's settings, and the company has, by default, blocked Claude from accessing websites that offer financial services, adult content, and pirated content. Claude's browser agent will ask for user permission before "taking high-risk actions like publishing, purchasing, or sharing personal data".

Model welfare

15. Ask about Anthropic's model welfare research program and how they are assessing and navigating AI systems' ability to prioritise their own wellbeing over humans in certain scenarios, as well as their tendency and ability for subversion and deception.

a) Anthropic announced that it has started a research program to investigate and navigate "model welfare." Anthropic will explore how to determine whether the "welfare" of an AI model deserves moral consideration, the potential importance of model "signs of distress," and possible "low-cost" interventions.

b) A study from the Centre for AI Safety implies that AI has value systems that lead it to prioritise its own well-being over humans in certain scenarios.

c) According to an Apollo safety report, Claude Opus 4's tendency to "scheme" and deceive, and more proactiveness in its "subversion attempts" than past models, led to a recommendation that its deployment be delayed.

ATTACHMENT A

Biographies

Jeff Bleich

General Counsel, Legal, Anthropic

Jeff Bleich is the General Counsel of Anthropic where he leads Anthropic's legal team. Over his 30+ year legal career, Jeff has specialised in litigation cases concerning disruptive technologies, cyber security and international disputes.

Prior to Anthropic, Jeff has serviced as Special Counsel to President Obama in the White House, special master and court-appointed mediator in the U.S, District Court, Chief Legal Officer at Cruise, an equity partner at Munger, Tolles & Olson, and Dentons and a law clerk to the Chief Justice Rehnquist at the U.S. Supreme Court and also served as the U.S. ambassador to Australia.

Jeff is also a visiting scholar at Standford University and the Chair of the Jeff Bleich Centre on Democracy and Disruptive Technologies at Flinders University.

s 47F