Meta

Submission on draft Industry Standards

The draft Online Safety (Relevant Electronic Services – Class 1A and 1B Material) Industry Standard 2024

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

22 January 2023

Executive Summary

Meta shares the Australian Government's objectives to keep Australians safe and protect them from the most harmful types of content online. We invest significantly in the safety and integrity of our services, including policies, technology, industry collaboration and local partnerships to action harmful content as soon as we become aware of it, in many cases, proactively – before Australians are exposed to it. We have around 40,000 people working on safety and security at Meta, and we have invested more than US$20B (~AU$ 29B) in teams and technology to enhance safety and security since 2016.

We also recognise the important role that transparency and accountability play in giving policy makers, regulators and others insights into, and confidence in our investment and commitment to safety and integrity. It is over five years since we first started publishing our Community Standards Enforcement Report[1], and we have steadily increased the number of transparency reports we provide over the years.

Meta welcomes the opportunity to provide a submission on both of the draft Online Safety (Relevant Electronic Services - Class 1A and Class 1B) Industry Standard 2024 and the Online Safety (Designated Internet Services - Class 1A and Class 1B) Industry Standard 2024 (respectively the **RES Standard** and **DIS Standard** - together the **draft Standards**). Our actions to date demonstrate that we share the objectives of the *Online Safety Act 2021 (Cth)* (**OSA**) and of the draft Standards – to ensure that the digital industry is investing in and doing its utmost to keep Australians safe online and allow the digital economy to contribute to Australia's social and economic well-being.

Since the OSA took effect in 2022, we have been working to be a constructive industry contributor to the development and implementation of the six industry codes (**Codes**),[2] which have all been registered by the eSafety Commissioner as of 12 September 2023, as well as responding in detail to Basic Online Safety Expectations (**BOSE**) notices issued to Meta in 2022. We have dedicated internal teams who have spent the last six months coordinating our response to the new obligations under the Codes and the broader online safety regulatory frameworks in Australia and globally.

---

[1] Meta, 'History of Meta's Transparency and Community Standards Enforcement Reports', May 2022, https://about.fb.com/wp-content/uploads/2023/05/10-Years-of-Transparency-Reporting.pdf; Meta, *Community Standards Enforcement Report*, Transparency Centre, https://transparency.fb.com/reports/community-standards-enforcement

[2] These industry codes are: the *Social Media Services Online Safety Code (Class 1A and Class 1B Material)*; *App Distribution Services Code (Class 1A and Class 1B Material)*; *Hosting Services Online Safety Code (Class 1A and Class 1B Material)*; *Internet Carriage Services Online Safety Code (Class 1A and Class 1B Material)*; *Equipment Online Safety Code (Class 1A and Class 1B Material);* and the *Internet Search Engine Services Online Safety Code (Class 1A and Class 1B Material)*, published at https://www.esafety.gov.au/industry/codes/register-online-industry-codes-standards

Our compliance experience to date has highlighted the importance of establishing durable, sustainable regulatory frameworks for online safety. Specifically, as the regulatory environment for online safety develops globally, there is necessarily an investment from global platforms in compliance systems that meet all the requirements of these regulatory frameworks. Consistency, predictability and sustainability of obligations allow platforms to make effective investments in online safety *and* build robust compliance systems that move the needle on user safety. For online safety regulatory frameworks to achieve their goals, it is helpful for obligations to be directed toward platform investment in compliance systems that actively and effectively enhance safety, rather than compliance systems that are bespoke, may expire, or become technologically or practically less effective over time. We appreciate that this is consistent with the intent of the draft Standards – namely, to be outcomes- and risk-based in their approach, recognising that different services have different risk profiles, that compliance measures should be proportionate to the level of risk, and flexible to enable effective implementation given these differences.[3]

Against this background, we welcome the opportunity to participate in this consultation on the draft Standards, not only because we want to help shape regulatory frameworks that hold companies such as Meta, and the digital industry more broadly, to account and to showcase best practice, but also because the draft Standards constitute a parallel framework to the Codes. When finalised, the draft Standards will house a separate and more particularised enforcement regime to the Codes. Within this parallel framework the powers of the Commissioner are expanded to take more immediate action against breaches of Standards, and consequently, as the Standards offer less room for consultation and a compliance approach that cannot be as readily tailored by a regulated service, it is even more important that obligations are clear and correctly scoped. As the scope of services captured by the Standards includes new and evolving technology, the removal of meaningful consultation parameters and remediation process prior to enforcement (eg., a 'Notice to Comply') make it critical the Standards are developed with goals of longevity, durability and compatibility with the existing regulatory framework under the OSA.

Given the significant investment already made by many, including Meta, in the digital industry in safety and integrity, the draft Standards can best be framed to incentivise ongoing investment with the goal of raising the bar, rather than designed for the sake of compliance alone. Whilst the draft Standards are being developed in the wake of the rejection by the eSafety Commissioner of two industry codes covering the same service categories (RES and DIS), many providers will nonetheless be developing safety, integrity and compliance measures across both the Codes

[3] eSafety Commissioner, *Discussion paper: Draft Online Safety (Relevant Electronic Services - 1A and 1B Material) Industry Standard 2024 and Draft Online Safety (Designated Internet Services - Class 1A and 1B Material) Industry Standard 2024*, November 2023, p6, https://www.esafety.gov.au/sites/default/files/2023-11/Discussion-Paper-draft-Online-Safety-Standards-%28Class-1A-and-1B%29.pdf

and the draft Standards, so ensuring the obligations are streamlined and consistent, where appropriate, will assist in achieving this objective.

With this in mind, we have some general comments on the draft Standards that go to firstly, the technical feasibility of the requirements, and secondly, suggestions intended to simplify compliance obligations. We have also included specific comments in response to the discussion questions raised in the Discussion Paper[4] that was released with the draft Standards.

Specifically on the technical feasibility of the draft Standards, we have made suggestions for the Office's consideration with respect to encrypted services, new and emerging services, and Generative AI.

- With respect to encrypted services, we appreciate the intention expressed by eSafety that nothing in the draft Standards are intended to require 'providers to design systematic vulnerabilities or weaknesses into end-to-end-encrypted services'.[5] However, we are concerned that the definition of *technical feasibility* within the draft Standards leaves room for ambiguity on this point and we believe that greater clarity would be helpful. To assist the Office in its considerable task of reviewing all of industry's feedback during this short consultation period, we have suggested some language that may improve the draft Standards to provide greater certainty for industry with respect to preserving the integrity of encrypted services.

- With respect to new and emerging services, the nature of the services can mean that the safety mitigations will be different. This is because many new and emerging services are highly immersive and immediate, which forces the focus on conduct rather than static content. Additionally, many of the technologies used for content governance in a two-dimensional world don't work or exist for three dimensional experiences in the way. Our suggested language with respect to the concept of *technical feasibility* is also intended to address this issue.

---

[4] eSafety Commissioner, *Discussion paper: Draft Online Safety (Relevant Electronic Services - 1A and 1B Material) Industry Standard 2024 and Draft Online Safety (Designated Internet Services - Class 1A and 1B Material) Industry Standard 2024*, November 2023, https://www.esafety.gov.au/sites/default/files/2023-11/Discussion-Paper-draft-Online-Safety-Standards-%28Class-1A-and-1B%29.pdf

[5] eSafety Commissioner, *Discussion paper: Draft Online Safety (Relevant Electronic Services - 1A and 1B Material) Industry Standard 2024 and Draft Online Safety (Designated Internet Services - Class 1A and 1B Material) Industry Standard 2024*, November 2023, p14, https://www.esafety.gov.au/sites/default/files/2023-11/Discussion-Paper-draft-Online-Safety-Standards-%28Class-1A-and-1B%29.pdf; eSafety Commissioner, *Fact sheet, Draft Online Safety (Designated Internet Services – Class 1A and Class 1B Material) Industry Standard 2024*, November 2023, p9, https://www.esafety.gov.au/sites/default/files/2023-11/Fact-sheet-Draft-Online-Safety-%28Designated-Internet-Services%E2%80%93Class-1A-and-Class-1B-Material%29-Industry-Standard-2024.pdf; see also *eSafety Commissioner, Fact Sheet: Draft Online Safety (Relevant Electronic Services – Class 1A and Class 1B Material) Industry Standard 2024*, updated December 2023, p7, https://www.esafety.gov.au/sites/default/files/2023-12/Fact-sheet-Draft-Online-Safety-Relevant-Electronic-Services%E2%80%93Class1A-and-Class1B-Material-Industry-Standard-UPDATEDDEC.pdf

- With respect to the development of regulatory frameworks for Generative AI – whilst the importance of responsible innovation in relation to Generative AI is incontrovertible, we believe that the shape and nature of safety and integrity mitigations with respect to this transformative technology and, consequently, enforceable obligations applied to these measures, requires more detailed consideration than the one month consultation available on the draft Standards allows. More considered engagement will ensure a comprehensive discussion about the different responsibilities and compliance strategies that are possible with respect to Generative AI.

  We recognise the intent of the draft DIS Standard with respect to Generative AI is to reduce evolving risks in a clear and targeted manner, while providing certainty over which services are captured, and, more specifically, to encourage services to proactively consider end-user safety and implement appropriate safeguards, without the need for formal regulation.[6] However, as presently drafted, the draft DIS Standard does not appear to achieve this goal. The provisions in the draft DIS Standard introduce unclear definitions, potentially lead to unintended and adverse consequences, overlook existing safety mitigations that have been deployed, and overlook technical limitations to what actions different stakeholders within the Generative AI ecosystem can do with respect to safety.

  With respect to foundational Generative AI models in particular, we note that recent research conducted by the Stanford Center for Research on Foundation Models cautions against policy proposals that disproportionately damage open foundation model developers and encourages consideration of interventions a choke points downstream of the foundation model layer.[7]

  For this reason, we suggest that consideration of Generative AI governance models be removed from this consultation on the draft Standards and directed in to a broader and more fulsome discussion of the complexities and nuances of this technology within the context of the broader and current other reviews and discussions, both internationally and within Australia, that are occurring with respect to governance frameworks for Generative AI.

To promote the simplification of compliance obligations, our key areas for general comment are focused on:

---

[6] eSafety Commissioner, *Fact sheet, Draft Online Safety (Designated Internet Services – Class 1A and Class 1B Material) Industry Standard 2024*, November 2023, p6, https://www.esafety.gov.au/sites/default/files/2023-11/Fact-sheet-Draft-Online-Safety-%28Designated-Internet-Services%E2%80%93Class-1A-and-Class-1B-Material%29-Industry-Standard-2024.pdf

[7] Standard University Human-Centered Artificial Intelligence, Considerations for Governing Open Foundation Models https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models

- Ensuring consistency and compatibility in definitions used across the Online Safety Framework, including the *Online Safety (Basic Online Safety Expectations) Determination 2022* (Cth) (BOSE Determination) and the six registered industry codes,
- Simplifying the scope of the draft RES Standard,
- Ensuring terms of use obligations are fair, consistent and consumer-friendly,
- Simplifying reporting requirements, and
- Recognising and adjusting obligations to allow for conflicts of law.

Recognising the short timeframe for consultation on the draft Standards, and in an effort to be helpful, we have provided some suggestions for the Office's consideration intended to achieve the goals of providing greater clarity for industry and measures that are both proportionate and flexible.

We appreciate the ongoing commitment of the Office to consulting with industry and building on the extensive work of industry bodies in the development of the draft Standards and look forward to the ongoing engagement. We are ready to assist the Office with any further details with respect to technical feasibility and our compliance work, that may be useful in finalising these draft Standards.

Table of Contents

Meta's commitment to safety & integrity

Recent updates to promote safety & combat harmful content

As noted above, Meta shares the intent of the draft Standards to establish and implement systems, processes and technologies to manage online safety risks and invests significantly in the safety and integrity of our services, and in resources and partnerships to ensure that we are adopting appropriate community safeguards for Australian consumers.

For this reason, we have consistently worked with the Office, the Australian Federal Police, and others, to proactively detect, remove, and report the most extreme forms of harmful content on our services. We acknowledge that Class 1A content (and in some cases Class 1B content) is illegal and harmful. As participants in the digital ecosystem, we recognise that we play a role in improving the online safety of all Australians by effectively managing the risks associated with these types of content in relation to our services. In addition to regulatory compliance, our approach comprises four elements: policies, enforcement of those policies, industry collaboration, and tools.

Policies

We have policies which outline what material is and is not allowed on our services (for example, the Facebook Community Standards[8] and the Instagram Community Guidelines[9] and the WhatsApp Terms of Service[10] and Channels Guidelines[11]).

Our policies are based on feedback from our community, and the advice of experts in fields such as technology, public safety, child safety and human rights. We regularly update our policies based on this feedback to ensure a continued safe experience for our users. Content broadly falling under Class 1A and Class 1B classifications is also violative of our policies.

Enforcement

We use a combination of technology and review teams to detect, review and take action on millions of pieces of content every day on Facebook and Instagram. Our technology proactively detects and removes the vast majority of violating content and accounts before anyone reports it.[12]

To provide transparency about how we enforce our policies across Facebook and Instagram, we regularly publish a Community Standards Enforcement Report[13] to more

[8] Facebook, *Facebook Community Standards*, Transparency Centre, https://www.facebook.com/communitystandards/introduction
[9] Instagram, *Community Guidelines*, Help Centre, https://help.instagram.com/477434105621119
[10] WhatsApp, *WhatsApp Terms of Service*, https://www.whatsapp.com/legal/terms-of-service
[11] WhatsApp, *WhatsApp Channels Guidelines*, https://faq.whatsapp.com/245599461477281
[12] Meta, 'Integrity Reports, Third Quarter 2023', *Transparency Centre*, 15 December 2023, https://transparency.fb.com/en-gb/integrity-reports-q3-2023
[13] Meta, *Community Standards Enforcement Report*, https://transparency.fb.com/reports/community-standards-enforcement

effectively track our progress and demonstrate our continued commitment to making our services safe and inclusive. This quarterly report details how we are doing at preventing and taking action on content that violates our policies. For example, in Quarter 3, 2023, we took action against:

- 16.9 million pieces of content for child sexual exploitation on Facebook, 99 percent of which was proactively detected and actioned before people reported it;[14]
- 750,000 pieces of organised hate content on Facebook, 94.3 per cent of which was proactively detected and actioned , and 129,000 pieces of organised hate content on Instagram, 77.2 percent of which was proactively detected and actioned before anyone reported it;[15]and,
- 9.6 million pieces of hate speech (one of the underlying causes of many forms of terrorism) on Facebook, 94.8 percent of which was proactively detected and actioned, and 7 million pieces of hate speech on Instagram, 96.5 percent of which was proactively detected and actioned before anyone reported it.[16]

On WhatsApp, through a combination of advanced technology and user reports, we ban around 8 million accounts a month[17], including around 300,000 accounts suspected of sharing child exploitative imagery.[18]

We also take broader enforcement measures. For example we have removed 1151 militarised social movements from our platform,[19] and when we become aware of apparent child exploitation, we report it to the National Center for Missing and Exploited Children (NCMEC), in compliance with applicable law.

Additionally, we hire specialists, including in law enforcement and online child safety, to find predatory networks and remove them. These specialists monitor evolving behaviors exhibited by these adversarial networks – such as new coded language – to not only remove them, but to inform the technology we use to proactively find them. Between 2020 and 2023, our teams disrupted 32 abusive networks and removed more than 160,000 accounts associated with those networks.[20]

---

[14] Meta, 'Community Standards Enforcement Report Q3 2023', https://transparency.fb.com/reports/community-standards-enforcement/child-nudity-and-sexual-exploitation/facebook
[15] Meta, 'Community Standards Enforcement Report Q3 2023', https://transparency.fb.com/reports/community-standards-enforcement/dangerous-organizations/facebook
[16] Meta, 'Community Standards Enforcement Report Q3 2023', https://transparency.fb.com/reports/community-standards-enforcement/hate-speech/facebook
[17] WhatsApp, *About WhatsApp and Elections*, https://faq.whatsapp.com/518562649771533/?helpref=uf_share
[18] Australian Institute of Criminology, 'Trends & issues in crime and criminal justice', July 2022, p8, https://www.aic.gov.au/sites/default/files/2022-07/ti653_csam_and_end-to-end_encryption_on_social_media_platforms.pdf
[19] Meta, 'An Update to How WeAddress Movements and Organizations Tied to Violence', *Newsroom*, 17 October 2022, https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence
[20] Meta, 'Our Work To Fight Online Predators', *Newsroom*, 1 December 2023, https://about.fb.com/news/2023/12/combating-online-predators

Industry collaboration

While we have made significant progress as a company in combating harmful content, our work is supported by a multi-faceted and collaborative effort between a range of stakeholders, including companies, governments and regulators, law enforcement, safety experts, and civil society organisations.

Some of our important industry partnerships include:

- **Project Protect** – a partnership between Meta, Google, Microsoft and 15 other tech companies to fight child sexual exploitation and abuse online based on a five-pillar plan focused on: tech innovation; collective action; transparency and accountability; information and knowledge sharing; and independent research.[21]
- **Lantern –** Meta is a founding member of the Lantern program, which enables technology companies to share signals about accounts and behaviors that violate their child safety policies. We provided the Tech Coalition with the technical infrastructure that sits behind the program as well as oversee the technology with them, ensuring it is simple to use and provides our partners with the information they need to track down potential predators on their own platforms.[22]
- **the Global Internet Forum to Counter Terrorism (GIFCT)** – an independent NGO, of which we are a founding member, that aims to prevent terrorists and violent extremists from exploiting digital platforms. The GIFCT's database of shared digital "hashes" (fingerprints) and agreed protocols for responding to a live terrorist incident improve our ability to enforce our policies. As of 2022, the GIFCT Hash Sharing Database contained 2.1 million hashes, making up 370,000 unique and distinct items (representing distinct clusters of hashed content).[23]

Meta is also a founding member of the Digital Trust and Safety Partnership (DTSP), which aims to set out global best practices to mitigate content and conduct-related risks associated with internet services and then verify that companies adhere to those best practices through internal and independent third-party assessments. It is a first-of-its-kind initiative bringing together technology companies of different sizes and business models around a common approach to increasing Trust & Safety across the internet. Members include Meta, Google, Twitter, LinkedIn, Microsoft, Apple, Zoom, Pinterest, Vimeo, Reddit, Shopify, Bitly, and Patreon. DTSP is an independent, non-partisan, industry-led expert group developing best practices to enhance trust and safety online. These best practices are developed with input and collaboration with consumer/user advocates, policymakers, law enforcement, relevant NGOs and various industry-wide experts, looking at the entire product development lifecycle that is broken down into 5 key areas:

---

[21] Tech Coalition, *What we do*, https://www.technologycoalition.org/what-we-do
[22] Meta, 'Introducing Lantern: Protecting Children Online', *Newsroom*, 7 November 2023, https://about.fb.com/news/2023/11/lantern-program-protecting-children-online
[23] GIFCT, GIFCT Transparency Report 2022, https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf

- **Development.** Identify, evaluate, and adjust for content- and conduct-related risks in product development.
- **Governance.** Adopt explainable processes for product governance, including which team is responsible for creating rules, and how rules are evolved.
- **Enforcement.** Conduct enforcement operations to implement product governance.
- **Improvement.** Assess and improve processes associated with content- and conduct-related risks.
- **Transparency.** Ensure that relevant trust & safety policies are published to the public, and report periodically to the public and other stakeholders regarding actions taken.

Last year, the DTSP published a new report entitled *"The Safe Assessments: An Inaugural Evaluation of Trust & Safety Best Practices"*[24] that synthesizes self-assessment submissions from ten DTSP members and provides an anonymized snapshot of industry's posture regarding digital trust and safety.

Tools

We design tools to give people more control over their online experience and help them stay safe. We have developed more than 30 tools to support safe, positive online experiences for teens and their families.[25] For example, we automatically set teens' accounts to private when they join Instagram, we restrict adults from sending private messages to teens who do not follow them, we use age assurance technology to help teens have age-appropriate experiences, and we have parental supervision tools that let parents see who their teen reports or blocks.[26]

We block known child sexual abuse material (**CSAM**) terms and hashtags that use or share violating content. When people search for these terms, we show a warning message. The message, developed with expert input, offers ways to get help from offender diversion organizations and shares information about the consequences of viewing illegal content. We have blocked tens of thousands of terms and hashtags linked to potentially predatory activity. As new terms are added to our system, the terms are restricted across Facebook and Instagram simultaneously.

With respect to violent extremism and dangerous organisations, additional tools include our Redirect Initiative, which helps combat violent extremism and dangerous organizations by redirecting hate and violence-related search terms on Facebook towards resources, education, and outreach groups that can help.[27] In 2019, we extended this program to Australia via a partnership with Exit Australia, a local organisation that helps people leave violent extremism

[24] Digital Trust & Safety Partnership, DTSP Safe Assessments Report, https://dtspartnership.org/dtsp-safe-assessments-report/
[25] Meta Help Center, Our tools, features and resources to help support teens and parents https://www.meta.com/en-gb/help/policies/safety/tools-support-teens-parents/
[26] Instagram, *About Instagram teen privacy and safety settings*, Help Centre, https://help.instagram.com/3237561506542117
[27] Facebook, *Counterspeech*, https://counterspeech.fb.com/en/initiatives/redirect

and terrorism. In 2023, we added Step Together as a partner we work with in the New South Wales region.

Investment in Australian initiatives & partnerships

We note the intent of the draft Standards is to improve online safety for Australians, especially pertaining to the exposure to Class 1A material or Class 1B material through the services. In Australia, we invest in partnerships with local organisations to promote greater awareness of our measures to provide meaningful protections for Australians, and we also invest in research with local experts to inform and help us to further strengthen our policies and tools to promote online safety and combat hate and violent extremism. For example, in 2023 we:

- continued our decade-long support for PROJECT ROCKIT, including their youth-led, peer-based anti-bullying Digital Ambassadors initiative, which has now empowered more than 25,000 young Australians to tackle cyberbullying and online hate;
- ran online education and awareness campaigns, including
    - a sextortion awareness campaign delivered in partnership with Thorn, Kids Helpline and the Australian Centre To counter Child Exploitation to inform young people about the dangers of online sextortion and encourage them to report it and seek support
    - a First Nations creator-led campaign delivered in partnership with ReachOut focusing on the social and emotional wellbeing experience of First Nations people in the lead up to the Voice to Parliament Referendum
    - a youth-driven initiative titled 'Intimate Images unwrapped' delivered in partnership with PROJECT ROCKIT that discusses online intimacy, intimate image sharing, respect, and consent.
- held a Youth Online Safety Workshop that brought together Meta's local online safety and wellbeing partners, as well as youth consultants, to share their perspectives and experiences regarding the most pressing online safety and wellbeing challenges and opportunities;
- supported the following Australia-specific research focusing on online hate and abuse:
    - Islamophobia Register Australia's fourth Islamophobia Report, conducted in partnership with Charles Sturt University and ISRA (the Islamic Sciences and Research Academy);[28]
    - Media Diversity Australia's research investigating the online abuse of diverse journalists and media workers in Australia;[29] and,
    - new research by the Online Hate Prevention Institute investigating hate speech targeting First Nations Australians specifically in the context of the debate around

[28] Assoc. Prof. Derya Iner Dr. Ron Mason Chloe Smith, *Islamophobia in Australia (2014-2021) - IV*, 2023, https://researchoutput.csu.edu.au/ws/portalfiles/portal/313346505/UPDATED_IslamophobiaInAustralia_ReportIV_digital_lowres_spread_update.pdf

[29] Media Diversity Australia, *Research and Resources - Online Safety of Diverse Journalists*, https://www.mediadiversityaustralia.org/online-safety-of-diverse-journalists

the Referendum for an Aboriginal and Torres Strait Islander Voice. This is due to be published in early 2024.

Our work is also informed and guided by the following Australia-specific advisory groups that Meta has convened to ensure that we consider relevant and diverse local perspectives across the development of our policy and program initiatives:

- The Australian Online Safety Advisory Group, which comprises experts from online safety, youth and mental health organisations including PROJECT ROCKIT, Alannah and Madeline Foundation, CyberSafety Solutions, WESNET, and the Butterfly Foundation among others;
- The Australian Combating Hate Speech Advisory Council**,** which comprises a mix of academic, civil society and NGO representatives from a range of Australian communities and provides members with a direct channel to provide us with advice and feedback about how to better identify and combat online hate on our services.

And finally, we are also committed to ongoing dialogue with government, civil society and industry stakeholders about how industry can work collectively to better address online safety. For example, in 2023 we actively participated in WESNET's fifth Technology Safety Summit, the Attorney-General's Department-led National Modern Slavery Conference and the eating disorder sector-led Body Image and Social Media roundtable.

Comments on the draft Standards

The Office has asked a number of discussion questions seeking feedback on particular areas of the draft Standards, and in an effort to be helpful, we have prepared detailed responses to these below. However, there are a number of overarching principles and general comments that we would encourage the Office to consider as part of its framing of the draft Standards that may require feedback and amendment beyond the specific areas that are the focus of the discussion questions.

General comments

Our General Comments can be categorised in two ways: those that address concerns around technical feasibility, and those that address concerns around streamlining compliance obligations so that industry can achieve the objectives of the draft Standards and OSA more broadly – namely: commitment to online safety and compliance.

Concerns relating to technical feasibility

We have concerns that the draft Standards may impose obligations, and thus penalties, on service providers that are not technically feasible, including but not limited to encrypted services, new and emerging services and Generative AI. With respect to encrypted and new and

emerging services, we have suggested some concepts for the Office's consideration that may assist the draft Standards to align more closely with the stated intentions in the supporting documentation of the fact sheets and Discussion Paper. With respect to Generative AI, we suggest further consultation and discussions should be undertaken given both the considerable international and domestic existing review processes underway with respect to safety and Generative AI.

Maintaining the integrity of encrypted services

We recognise that there is general agreement across industry, civil society and within Government about the value of encryption to promote privacy, safety, and security. While there are concerns that have been raised about the ability to promote safety and combat pro-terror content on encrypted services, for Meta, the values of safety, privacy, and security are mutually reinforcing. An independent Human Rights Impact Assessment of Meta's expansion of end-to-end encryption - conducted by NGO Business for Social Responsibility in line with UN Guiding Principles on Business and Human Rights - found, among other areas, that encryption increased the realisation of privacy, freedom of expression, protection against cybercrime threats, physical safety, freedom of belief and religious practices and freedom from state-sponsored surveillance and espionage.[30] In line with these findings, we continue to invest in behavioural analysis and metadata as effective harm prevention rather than undermine encryption, which is what we have done and are continuing to progress.

In crafting the draft Standards, Meta appreciates the acknowledgement by the Office in the Discussion Paper that, in some cases, there will be technical limitations as to what can be done to detect and remove harmful material.[31] Additionally, the Discussion Paper and fact sheets that accompany both draft Standards explain that the Office 'does not expect providers to design systematic vulnerabilities or weaknesses into end-to-end-encrypted services'.[32] However, this certainty is not replicated in the Standards themselves. Whilst the Discussion Paper and facts sheets both suggest that this outcome is achieved by ensuring that the Standards do not require service providers to do anything that is not *technically feasible*, the concept of *technical feasibility* as it has been included within the draft Standards does not have this intended effect.

---

[30] Business for Social Responsibility, *Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption*, 2022, https://www.bsr.org/reports/bsr-meta-human-rights-impact-assessment-e2ee-report.pdf

[31] See eSafety Commissioner, *Discussion paper: Draft Online Safety (Relevant Electronic Services - 1A and 1B Material) Industry Standard 2024 and Draft Online Safety (Designated Internet Services - Class 1A and 1B Material) Industry Standard 2024*, November 2023, pp13-14, https://www.esafety.gov.au/sites/default/files/2023-11/Discussion-Paper-draft-Online-Safety-Standards-%28Class-1A-and-1B%29.pdf

[32] See, for example, *eSafety Commissioner, Fact Sheet: Draft Online Safety (Relevant Electronic Services – Class 1A and Class 1B Material) Industry Standard 2024*, updated December 2023, p7, https://www.esafety.gov.au/sites/default/files/2023-12/Fact-sheet-Draft-Online-Safety-Relevant-Electronic-Services%E2%80%93Class1A-and-Class1B-Material-Industry-Standard-UPDATEDDEC.pdf

Technical feasibility issues with effective disruption and deterrence of CSAM and pro-terror material, however, are not limited solely to encryption. Technical feasibility can also be an issue in new and emerging services for which effective systems and technologies are not developed. For example, for ephemeral, 3D and other new and emerging online experiences like synchronous multiplayer gaming with audio communication, what is a technically feasible and effective safety mitigation looks different.Safety mitigations that are effective and proportionate in a two dimensional and static content environment may not be technically feasible, transferable or appropriate in more immersive environments. The tools maintaining safety, security, and integrity in these new and emerging services are different, and these tools are also new and emerging.

Clarification of the concept of technical feasibility

Given the above discussion of the complexities in relation to encrypted and new and emerging technologies, it is concerning that the definition of *technical feasibility* included in the draft Standards only refers to financial considerations (i.e., the financial cost of taking the relevant action) – as distinct to what may be technically possible or technically effective. Whilst financial considerations may be relevant to considerations of what is within the bounds of feasibility, there can also be technical and practical constraints that may prevent a service provider from taking a particular action. This should include consideration of factors such as:

- the technologies used to deliver the service, and any inherent limitations to those technologies, or inherent effectiveness of existing technologies;
- any technological safeguards used to protect the security and integrity of the underlying service;
- what technologies currently exist (if any) that would enable the service provider to take the relevant action in question and whether those technologies are readily available to the service provider (e.g., whether any necessary licences are readily available on commercial terms); and
- whether what may be “feasible” is in fact practical, effective or appropriate to address online harm.

We encourage the Office to give consideration to explicitly clarifying within the draft Standards that the concept of *technical feasibility* does not require a provider to introduce a systemic vulnerability or weakness that would undermine the integrity of encrypted services or new and emerging services . One example of how this could be achieved, shared in an effort to be helpful given the tight timeframes for this consultation, is to include the following proposed additional language (see below the language included in *italics*, using the language drawn directly from section 8(2) of the BOSE Determination):

“In considering whether it is or is not technically feasible for the provider of a electronic service to take a particular action, the matters to be taken into account include:

(a) the expected financial cost to the provider of taking the action;
(b) whether it is reasonable to expect the provider to incur that cost, having regard to the level of the risk to the online safety of end-users in Australia of not taking the action; and
(c) the current state of technology, including any inherent technical limitations in the communication networks, software and hardware used by the provider to provide the service.

An action will not be taken to be technically feasible where it would involve the provider undertaking steps that could do the following:

(a) implement or build a systematic weakness, or a systematic vulnerability, into a form of encrypted service or other information security measure;
(b) build a new decryption capability in relation to encrypted services; or
(c) render methods of encryption less effective."

This is also consistent with clause 6.1 of the Head Terms for the Codes (which will operate in parallel with the draft Standards) and with section 8(2) of the BOSE Determination.[33] The approach of expressly recognising that obligations with respect to encrypted services do not require providers to implement or build a systemic weakness or a systemic vulnerability or to take other actions that would render methods of encryption less effective is also consistent with another legislative framework imposing obligations on service providers, namely Part 15 of the *Telecommunications Act 1997* (Cth) (Telco Act).[34]

Regulation of Generative AI in the DIS Standard

Another area of concern with respect to the technical feasibility of service providers to comply with obligations under the draft Standards, and more specifically the DIS Standard, relates to Generative AI and in particular open source models.

The global Generative AI regulatory landscape

By way of background, whilst the public debate with respect to this transformative technology is relatively new, the work on developing the technology is not. Just by way of one example, in November 2023, at Meta, we celebrated the ten-year anniversary of Meta's Fundamental AI Research (FAIR). For the past ten years FAIR has produced breakthroughs on many of the hardest problems in AI through open and responsible research – in a broad range of areas

[33] The explanatory statement for the BOSE Determination provides that the purpose of this language is to 'provide additional assurances to service providers that it is not expected that services access encrypted messages'.
[34] Part 15 of the Telco Act requires designated communications providers to provide technical assistance to national security and law enforcement agencies. Section 317ZG expressly provides that any requested assistance must not extend to implementing or building a systemic weakness or a systemic vulnerability into a form of electronic protection, such as by implementing or building a new decryption capability or taking other actions that would render methods of encryption less effective. As explained in the explanatory memorandum for the bill that introduced the industry assistance regime into the Telco Act:

> New section 317ZG ensures that providers cannot be requested or required to systemically weaken their systems of electronic protection under a technical assistance notice or technical capability notice. The limitation is designed to protect the fundamental security of software and devices. It ensures that the products Australians enjoy and rely on cannot be made vulnerable to interference by malicious actors.

including object detection, unsupervised machine translation, and large language models – which in turn have had global, real-world impact.[35] For example, Yale and EPFL's Lab for Intelligent Global Health Technologies used our open source large language model, Llama 2, to build Meditron, the world's best performing open source large language model tailored to the medical field to help guide clinical decision-making. Meta also partnered with New York University on AI research to develop faster MRI scans. And we are partnering with Carnegie Mellon University on a project that is using AI to develop forms of renewable energy storage.[36]

As you know, there has been significant international and Australian discussion and consideration of governance frameworks with respect to Generative AI. To recap – globally, in late 2023, the US Government released an *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*,[37] the Group of Seven (G7) Leaders released a *Statement on the Hiroshima AI Process* in which they, among other things, instructed relevant ministers to accelerate the process toward developing the *Hiroshima AI Process Comprehensive Policy Framework*,[38] and the UK Government hosted the AI Safety Summit resulting in the *Bletchley Declaration* to which Australia is a signatory.[39] These complement existing global frameworks, such as the OECD *Principles on Artificial Intelligence* adopted in May 2019 by OECD member countries.[40] In Australia, governance models for AI have been considered as part of the Department of Industry, Science and Resources' consultations on *Positioning Australia as a leader in digital economy regulation (automated decision making and AI regulation)*[41] and *Safe and Responsible AI in Australia*.[42] And in addition to consideration of Generative AI within the DIS Standard, it is also scheduled for further consideration as part of the review of the BOSE Determination and OSA, and under the new measures proposed by the Australian Government this past week – including an AI Safety Standard and mandatory guardrails to promote the safe design, development and deployment of AI systems.[43]

---

[35] Meta, 'Celebrating 10 years of FAIR: A decade of advancing the state-of-the-art through open research', 30 November 2023, https://ai.meta.com/blog/fair-10-year-anniversary-open-science-meta

[36] *Meta, 'On AI, Progress and Vigilance Can Go Hand in Hand'*, 19 January 2024, https://about.fb.com/news/2024/01/davos-ai-discussions

[37] US National Archives Federal Register, 'Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', Executive Order 14110, 88 FR 75191, 30 October 2023, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[38] G7. 'G7 Leaders' Statement on the Hiroshima AI Process', 30 October 2023, https://g7g20-documents.org/database/document/2023-g7-japan-leaders-leaders-language-g7-leaders-statement-on-the-hiroshima-ai-process

[39] UK Government, 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023', https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[40] *OECD, 'Artificial Intelligence: OECD Principles', https://www.oecd.org/digital/artificial-intelligence*

[41] Department of Industry, Science and Resources, 'Positioning Australia as a leader in digital economy regulation (automated decision making and AI regulation): issues paper', 18 March 2022, https://consult.industry.gov.au/automated-decision-making-ai-regulation-issues-paper

[42] Department of Industry, Science and Resources, 'Safe and responsible AI in Australia consultation - Australian Government's interim response', 17 January 2024, https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf

[43] The Hon Ed Husic MP, 'Action to help ensure AI is safe and responsible', 17 January 2024, https://www.minister.industry.gov.au/ministers/husic/media-releases/action-help-ensure-ai-safe-and-responsible

Australia's peers have acknowledged that generative AI is a nascent technology at its early stages, and that a rush to regulate without taking the necessary time to determine whether there are net new, clear and actionable harms and the most effective way to address those, will risk dramatically curtailing innovation in this area, and the benefits that it can bring to individuals, societies, businesses, and governments. As noted by Export Finance Australia, Goldman Sachs suggest AI adoption could lift global productivity growth by 1.5 percentage points over a 10-year period and drive a 7% (or US$7 trillion) increase in global GDP[44].

This is why countries such as the US, UK, India, Japan and Singapore have not rushed to regulate, and instead are collaborating through international efforts, while also pursuing domestic initiatives to establish pragmatic frameworks (such as the White House Voluntary Commitments) that can provide guidance to all the actors in the AI lifecycle on the responsible and safety development and deployment of this technology.

Meta's approach to safety for Generative AI

We share the concerns of the Australian Government and those of policy makers internationally that it is important that all technology, but particularly technology such as AI, is built and deployed responsibly. This is why we have developed five pillars of responsible AI that inform our work, to ensure that AI is designed and used responsibly.[45] As part of that, we recognise the value of how open source AI can contribute to responsible AI development. These are the principles that have guided our approach to responsible innovation:

- *Privacy and security:* AI can enhance privacy. We are investing in research on privacy-preserving machine learning technology (differential privacy, federated learning, encrypted computation) and teaching people how to use it. By making our models open source, others will be able to advance research in this area too.
- *Fairness and inclusion:* As part of our Massively Multilingual Speech (MMS) project, we've released models[46] for speech-to-text, text-to-speech, and more for 1,100+ languages. Others are now able to build on those models, improving inclusivity.
- *Robustness and safety:* Open sourcing can lead to safer products through an open community that can iteratively improve them. We also leverage red team methodologies to test the robustness of our AI-powered integrity systems against threats. We have been working with external partners to red team our Generative AI models.
- *Transparency and control:* As well as open sourcing our models, we often provide accompanying model cards and weights, which aids transparency and reproducibility.

---

[44] Export Finance Australia, *Australia—AI adoption creates benefits and challenges for businesses*, https://www.exportfinance.gov.au/resources/world-risk-developments/2023/may/australia-ai-adoption-creates-benefits-and-challenges-for-businesses/

[45] Meta, *Our commitment to Responsible AI*, https://ai.meta.com/responsible-ai

[46] Meta AI, Introducing speech-to-text, text-to-speech, and more for 1,100+ languages https://ai.facebook.com/blog/multilingual-model-speech-recognition

- *Accountability and governance:* We have invested in our Privacy Review efforts, developed approaches and tools to improve our understanding and ability to address concerns about our AI systems, and increased transparency and control around our AI products and features.

With respect to safety for on-platform Generative AI features, we have the following mitigations in place to design for safety:

- *Input filters:* When someone interacts with our AI models, anything they say is scanned by classifiers to detect whether they're soliciting content that could violate our policies.
- *Output filters:* We scan the content that our models create to detect potential violations of our policies. Content that is confirmed to violate is regenerated until no violation is present.
- *Social systems:* Content created by Meta's GenAI features, when shared onto existing social surfaces (ex: FB Feed) remains subject to our broader policies, user reporting mechanisms, and existing sets of content classifiers.
- *Lockouts:* People who repeatedly attempt to generate content that violates our policies will lose the ability to interact with our GenAI features for a fixed period of time.

With respect to safety and privacy for our models available for businesses and developers, our Llama 2 research paper[47] outlines Meta's approach to safety and privacy, for example we:

- Analyse for bias and adopt privacy protections in pre-training data.
- Conduct model evaluations against industry safety benchmarks (e.g. truthfulness, toxicity).
- Deploy red teaming, and also:
  - Submitted Llama 2 to DEFCON for nearly 2,500 hackers to stress test.
  - Incentivise researchers to stress test and report vulnerabilities through our bug bounty program.
- Implement our Responsible Use Guide[48], which includes guidance for developers on responsible fine tuning and red teaming.
- Provide a voluntary reporting form[49] that developers can use to report risky content that is generated by Meta's models that can be used to tweak fine-tuning to minimise the ability to replicate this in future.
- Put an Acceptable Use Policy in place that prohibits certain use cases to help ensure that these models are being used fairly and responsibly.[50]
- Work with governments, industries, academia, and civil society on collaboratively developing guardrails. This includes working with the Partnership on AI on Responsible

[47] Hugo Touvron, *et al*, 'Llama 2: Open Foundation and Fine-Tuned Chat Models', 18 July 2023, https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models
[48] Meta, *Responsible Use Guide: your resource for building responsibly*, https://ai.meta.com/llama/responsible-use-guide
[49] Available at http://developers.facebook.com/llama_output_feedback
[50] Meta, 'Llama 2 - Acceptable Use Policy', https://ai.meta.com/llama/use-policy

Practices for Synthetic Media[51] and its draft Guidance for Safe Foundation Model Deployment which is now open for public comment.[52] We are also founding members of AI Alliance along with IBM, Intel, numerous universities and Fast.ai from Australia, with the goal of advancing the safe and responsible use of AI through open innovation by working with the global community to create resources and benchmarks to enable the responsible and safe development and deployment of AI.

We are also developing new software tools for testing and improving robustness, and share them with the AI research and engineering community. We recently launched Purple Llama, an umbrella project featuring open trust and safety tools and evaluations meant to level the playing field for developers to responsibly and safely deploy generative AI models and experiences in accordance with best practices shared in our Responsible Use Guide.[53] This is a major step towards enabling community collaboration and standardizing the development and usage of trust and safety tools for generative AI development.

Our concerns regarding generative AI and the DIS Standard

Against this background, we have concerns that the current inclusion of specific requirements with respect to Generative AI within the DIS Standard are not technically feasible and therefore, are unlikely to achieve their intended purpose. Specifically, some of our concerns are:

- *Unclear and impractical definitions*: The AI-related definitions within the DIS Standard are unclear when we assess them against many of our Generative AI-related products. By way of example – it appears that the DIS Standard intends that the website from which Meta's Llama 2 fine tuned models are available for download meets the definition of a *machine learning model platform service,* but it is not clear that the actual models that are available for download are also intended to meet that definition. By way of further example, it appears – based on the fact sheet for the DIS Standard – that models that are offered to enterprise customers become an *enterprise DIS* and not a *machine learning model platform service*. However, it is unclear what happens if a model, such as Meta's Llama 2 fine tuned and pre-trained models, is offered to both organisations and also individual researchers. Finally, whilst the fact sheet makes it clear that if safeguards are implemented, a Generative AI service will no longer be considered a *high impact generative AI DIS,* this is not sufficiently clear in the drafting of the DIS Standard itself. All of these concepts are further complicated by the fact that many of the international AI governance frameworks, including the Bletchley Declaration, to which Australia is a

[51] Partnership on AI, *Responsible Practices for Synthetic Media - A Framework for Collective Action*, https://syntheticmedia.partnershiponai.org/#landing
[52] Partnership on AI, *draft Guidance for Safe Foundation Model Deployment - A Framework for Collective Action*, https://partnershiponai.org/modeldeployment
[53] Meta, 'Announcing Purple Llama: Towards open trust and safety in the new world of generative AI', 7 December 2023, https://ai.meta.com/blog/purple-llama-open-trust-safety-generative-ai

signatory, refer specifically to safety risks at the frontier of AI, rather than current models and products.[54]

- *Technically infeasible and disproportionate safeguards*: The DIS Standard contains highly specific safeguards, which may not be appropriate for all AI models, may not be technically feasible or proportionate for all stakeholders in the Generative AI ecosystem, and overlooks the safety mitigations already in place. We have outlined the safety mitigations adopted in relation to one of our open source Generative AI models, Llama 2, above, as an example. But the proposed safeguards in the draft DIS Standard include very different and potentially impossible mitigations. For instance, expanding on the example of Llama 2, it is not possible for us to suspend provision of Llama 2 once it has been downloaded nor terminate an account, or to deter, disrupt, detect, report, or remove content from models that have been downloaded. More generally, the proposed safeguards are disproportionate and technically infeasible for upstream stakeholders in the Generative AI ecosystem, who do not provide Generative AI products to end-users. .
- *Upstream stakeholders may have little to no oversight*: It is not always possible for a distributor of a machine learning model to monitor the uses of that model by downstream third party developers nor comply with the obligations imposed by the draft DIS Standard. The draft DIS Standard attempts to recognise this with respect to Enterprise DIS, but this is not clear in relation to providers of services that meet the definition of *machine learning model platform services*. This distinction does not necessarily map to the technical reality of how AI models such as Llama operate and how the AI ecosystem functions. It is also unclear why the Discussion Paper believes that there is a higher risk with respect to open source AI models (i.e. machine learning model platform service) than Enterprise DIS in this regard. Whilst the Fact Sheet states that the limited requirements imposed on an *Enterprise DIS* reflects the "limited visibility and control they have over the downstream uses, and the capability of such provides to build in impactful protections", the same could be said about *machine learning platform services* providers, but this is not reflected in the drafting.

Even if the Generative-AI specific definitions in the DIS Standard were removed, a developer that is providing a Generative AI service will still be a 'designated internet service' and will be required to assess the risk that the service may be used to generate Class 1A and 1B material under section 8 of the draft DIS Standard. Accordingly, once a risk assessment is conducted, these services would be considered Tier 1, 2 or 3 designated services and the remainder of the DIS Standard would apply as relevant. Given this, it is not clear what benefit arises from including specific Generative AI concepts into the DIS Standard, and it would be more beneficial to rely on the general category of 'designated internet service' as a broad and technically neutral

[54] UK Government, 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023', https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

category, which will help to better future-proof the DIS Standard against the fast-changing Generative AI technology and ecosystem.

Open source Generative AI

We appreciate that the draft DIS Standard has been crafted based on a belief that open source generative AI models present a significant risk of producing extremely harmful content on the basis that safeguards can be manipulated or removed. However, we have two concerns with this assumption. Firstly, there is debate about whether there is significant evidence that the marginal risk of open foundation models, relative to closed models or pre-existing technologies, is greater.[55] Secondly, the safeguards that have been proposed are in most cases not possible or appropriate, especially for *Enterprise DIS* and *machine learning platform services* who provide models that do not necessarily have a user interface and are not capable of generating content. The proposed safeguards are also not possible to enforce with respect to downstream uses that are not capable of being controlled once a model has been downloaded by a third party.

Moreover, while we agree that open source is not always the most fitting approach, we do not agree with the characterization that a responsible open approach to releasing generative AI models presents an ex ante high/unacceptable risk. In fact, it is widely accepted that open innovation improves safety outcomes. The current drafting of the DIS Standard will likely have the unintended effect of disincentivising open sourcing of Generative AI models, as it imposes onerous and technically infeasible safeguards on providers. For this reason, the value of open source in promoting safety and addressing other concerns with respect to Generative AI seems to have been overlooked. By democratising access to AI models, toxicity, bias, bugs and vulnerabilities can be identified by an open community, and mitigations found, to iteratively improve them. Additionally, open source creates a bigger, collective security team to combat those bad actors. In fact, it lets outsiders (including academics and researchers with extensive experience) hold companies accountable which, in turn, strengthens public trust.
Meta's approach in this respect has been deliberate – before releasing models for commercial purposes, our approach has been to start by releasing models to researchers to learn and improve before we release the models more broadly. There is a long history of open sourcing making products safer in the cyber security context. Two decades ago, the industry was closed source out of the fear that threat actors would use this technology for adversarial means. Today, it is overwhelmingly open source, and our systems are more secure and our defenses stronger because of this open approach.

The complexity of these issues have been acknowledged by the Office in its recent *Position Statement on Generative AI*,[56] in which the Office acknowledged the complexities of regulatory

[55] Standard University Human-Centered Artificial Intelligence, Considerations for Governing Open Foundation Models https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models
[56] eSafety Commissioner, *Tech Trends Position Statement – Generative AI*, 15 August 2023, https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

approaches to AI and the ongoing global discussions regarding them. These complexities include the question of 'whether the development of AI should occur in open or closed environments' and consideration of the responsibilities of various actors in the generative AI ecosystem - including developers, companies that integrate the models into their applications, and the people who use the applications - in preventing and mitigating harms. The Office noted it was 'crucial' to find the 'right balance' between open and closed systems, so as to both foster innovation and manage the risks posed by AI.[57] As the Office recognises in that paper, open source models help to democratise access to AI, allow researchers to identify errors and biases, and 'mitigat[e] the risk that generative AI is overly concentrated in large tech companies with access to training data and computing power.'[58]

Given these nuances surrounding Generative AI , the ecosystem that develops it and builds on it, and the diversity of use cases and consumer interactions, we suggest that discussions around the governance frameworks and regulatory obligations for Generative AI, including with respect to safety, be channeled into a separate process and benchmarked against international governance developments.

Streamlining compliance obligations

Ensuring consistency and compatibility in definitions used across Online Safety Framework, including the Industry Codes and Standards

We recognise that the Office has made the decision to issue Standards with respect to RES and DIS in an effort to ensure industry is meeting Australian community standards when offering services that fall within the scope of these two Standards. Acknowledging this, it is also important to recognise that the type of content to which both the draft Standards and the Codes will apply, and consequently, the compliance systems that industry will be investing to adhere to them, are the same.

Specifically, the six existing registered Codes and the draft Standards are both intended to deal with certain subcategories of 'class 1 material' as defined in section 106 of the OSA. The definition of 'class 1 material' in turn draws on the classification regime established under the *Classification (Publications, Films and Computer Games) Act 1995* (Cth) and in doing so pulls in many concepts defined under that Act and in subordinate instruments used in the National Classification Scheme. The Codes and the draft Standards have both made attempts to synthesise these definitions in a way that makes them more easily comprehensible so that they can be applied in practice by online service providers operating at scale. However, they do so in

[57] eSafety Commissioner, *Tech Trends Position Statement – Generative AI*, 15 August 2023, p11 https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

[58] eSafety Commissioner, *Tech Trends Position Statement – Generative AI*, 15 August 2023, p11 https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

slightly different ways, meaning that the definitions across the Codes and draft Standards are not the same.

As a consequence, the same piece of content may in theory be treated differently if made available on, for example a social media service (which would be subject to the SMS Industry Code) rather than an instant messaging service (which would be subject to the draft RES Standard). This seems to add compliance obligations for the sake of it to online service providers who may provide a range of different types of online services but will nonetheless need to develop service-wide applicable compliance systems under both the Codes and draft Standards.

To address these concerns, we consider that:

- to the extent possible, definitions used in the draft Standards should be aligned with those used in the registered Codes; and
- to the extent it is not possible to align the definitions, the Office should provide guidance for the Standards once finalised that includes:
    - a detailed comparison between the definitions used in the Codes and those used in the draft Standards with notes to indicate any material differences in scope; and
    - a statement to the effect that where a service provider who provides multiple different online services covered by a combination of Codes and Standards has in good faith designed its compliance systems by reference to the definitions in one Code or Standard, they will not be considered to be in breach of any other Code or Standard simply because the application of a different set of definitions may result in a different action being required in that other Code or Standard. This will support service providers in designing and implementing consistent technological solutions to ensure compliance across the full range of online services they provide.

The draft Standards also require that where a particular action is not technically feasible, such as automatic detection and removal of harmful material, the relevant service provider may be required (in multiple reports) to report to the Office on the extent to which the action is not technically feasible and the alternative actions that the service provider will take to address the underlying compliance obligation (see, for example, section 34 of the draft RES Standard). The information in such reports will likely be highly sensitive and susceptible to abuse by bad actors wishing to avoid techniques used by service providers to detect and prevent harmful activities on their services, in the event it was published or otherwise made available. Such sensitive and business-confidential disclosures could potentially undermine the overall safety and integrity ecosystem of online services. Consequently, consistent with the intent of the Standards and the overall Australian online safety regulatory framework, we suggest that consideration be given to expressly drafting confidentiality protection provisions where commercially sensitive and highly technical detail is required by the Office. This clarity would be welcome, especially given that Part 15 of the OSA is otherwise the only section that deals with confidentiality restraints. Additionally, in order to simplify reporting under the Standards (as we set out below), we

recommended that providers be required to report details of when any actions are not technical feasibility in the annual compliance report only. We expand on this in our section on 'Simplifying reporting requirements' below.

Simplifying the scope of the draft RES Standard

The scope of the draft RES Standard, including definitions of different types of RES, could benefit from simplification, making the obligations unambiguous for industry to develop effective compliance systems for and to support public understanding and fluency in the Standards.

The draft RES Standard uses a series of definitions to distinguish between different types of RES, with different compliance obligations applying in each case. Some types of RES will fall into a 'pre-assessed' category where no further risk assessment is required. However, different compliance obligations may still apply to different types of RES within this pre-assessed category. Other types of RES that do not fall within the pre-assessed category will need to be separately assessed and assigned to one of three different risk tiers, with different compliance obligations applying to each different tier. For example, the way Part 3 operates a particular service could be *both* pre-assessed RES *and* Tier 1, with different obligations applying to pre-assessed or Tier 1. Part 3 also has provisions that exclude pre-assessed and Tier 1 RES from risk assessment obligations, alongside obligations that inherently require a risk assessment. The RES definitions fuse together several regulatory classification and assessment concepts - functional classification (based on the primary purpose or primary function of a product), declaration of size and scale of service (with a higher bar for larger services), and assessment of the risk of exposure to Class 1A and Class 1B content.

Our view is that the complexity inherent in this approach is undesirable and will make the draft RES Standard harder to apply in practice, particularly for service providers that may provide a range of different types of RES that may each be subject to slightly different compliance obligations. It will also make the draft RES Standard less adaptable and durable in the context of new and evolving technology, and increase uncertainty as to how new types of RES, which may not currently be contemplated within the definitions put forward in the draft RES Standard, are to be regulated.

To address these concerns, we consider it would be desirable to look for ways to consolidate the different types of RES so that there is greater consistency in the range of compliance obligations that apply across this section of the online industry. For example, it may be possible to consolidate into two general categories of 'open' or 'closed' RES. We consider that this streamlining of definitions would also make the draft RES Standard easier for the public to understand and for industry to comply in practice.

For those services which require a risk assessment to determine tiering, the complex, inconsistent and incompatible risk assessment obligations across the Online Safety Framework have the unintended impact of an unnecessary compliance burden that does not necessarily

deliver harm reduction or effectively combat the harms in scope. Tiering based on size and scale of service to accrue obligations is a tested framework appropriate to stratify obligations across services. However size and scale are not necessarily indicators of risk of exposure to harmful content. The Part 3 obligations blend these concepts in a way that creates an intense obligation for a compliance system that requires a variety of risk assessments triggered by frequently occurring events (ie changes to service, features or functionality). Risk assessment obligations should be part of a clear risk assessment framework that is compatible with global norms and industry standards and is flexible to accommodate new and emerging technology. A globally standard and appropriate risk assessment framework (for example, the Digital Trust and Safety Partnership risk assessment framework (see above)), removes the likelihood of confusion among industry that is likely to result in inconsistent interpretation and application.

Striking a balance between flexibility and enforceability

We appreciate the acknowledgement in the Discussion Paper about 'the importance of striking a balance between flexibility and enforceability' in the draft Standards.[59] We agree that maintaining this balance is critical because it allows for the draft Standards to be adaptable to changes and innovations both in the types of services being offered, as well as changes and improvements in safety and integrity measures deployed by industry.

For these reasons, we support the approach in the draft Standards of allowing service providers a degree of discretion in determining what 'appropriate actions' to take in relation to the services they provide to fulfil the compliance obligations to which they are subject. However, we are concerned that some requirements are highly prescriptive without a clear indication to industry that these will cause an increase in investment in and accountability for online safety and integrity.

For example, the requirement to implement 'systems, processes **and** technologies that effectively' disrupt and deter CSAM and pro-terror material is highly prescriptive. There is a risk that a service provider who is using, for example, a specific technology to disrupt the sharing of CSAM or pro-terror material will not meet this requirement because they are not also using a process or a system. For example, a service provider may use behavioural signals (a technology) to detect the sharing of CSAM or pro-terror material in an encrypted setting. However, the (perhaps unintended) consequence of the language used in this requirement is that behavioural signals alone may not meet the requirement because they also do not involve a "process" or "system". We suggest that this language be replaced by a requirement to implement 'appropriate systems, processes **and/or** technologies that effectively disrupt and deter CSAM and pro-terror material' (as is similarly required in the registered Codes).

---

[59] eSafety Commissioner, *Discussion paper: Draft Online Safety (Relevant Electronic Services - 1A and 1B Material) Industry Standard 2024 and Draft Online Safety (Designated Internet Services - Class 1A and 1B Material) Industry Standard 2024*, November 2023, p10, https://www.esafety.gov.au/sites/default/files/2023-11/Discussion-Paper-draft-Online-Safety-Standards-%28Class-1A-and-1B%29.pdf

In addition, the above requirement requires service providers to implement systems, processes and technology that **effectively** deter and disrupt end-users from sharing CSAM and pro-terror material. While a service provider should always strive to achieve this outcome, the effectiveness of such safety measures can be hard to predict, especially as the tactics used by bad actors continue to evolve. For this reason, we suggest the inclusion of a technical feasibility exemption to this requirement, given that it is not practically possible for service providers to implement measures which are failproof, and in relation to some technologies, to develop a baseline metric against which to compare efficacy. Alternatively, we suggest that the draft Standards adopt a similar position to the Codes on this issue. The Codes acknowledge simply require service providers to 'take actions' and invest in systems, processes and technologies that 'aim' to disrupt or deter the relevant harmful activity. We suggest that the draft Standards adopt similar language so that the obligation is on service providers to implement systems, processes and technologies that 'are designed to' effectively deter and disrupt end-users from sharing CSAM and pro-teror material.

A further illustration of this is the requirement in section 14(2) of each draft Standard, where the provider of a service is required to include highly prescriptive provisions in the terms of use for a relevant RES or DIS. This includes amongst other things 'imposing an obligation on the account holder of the service to ensure that the service is not used, whether by the account holder, or by an end-user in Australia, to solicit, access, distribute or store Class 1A material or Class 1B material'. On its present wording, it seems that this would require the introduction of a specific contractual term that expressly addresses Class 1A and Class 1B material as defined in the draft Standards, thereby importing into the contract complex concepts originating from the Australian online safety and classification framework. This would carry little meaning for an international audience and also may not be readily understandable to Australian audiences. While a similar requirement was included in the Codes, there was also significantly more leeway for a service provider to take a nuanced approach, especially in the prohibition of Class 1B content (which may not be illegal in all cases). By way of practical example, drug related material used in an appropriate context (for example in anti-drug messaging), may be permitted in the Codes. Under the draft Standard, this would be prohibited with no room for appropriate adaptations.

In the interests of clarity and workability, requirements should be applied expressed flexibly so that it can be applied in a service specific and user-friendly way. For example, in compliance with the provisions and protections under the the Codes (and specifically, the Head Terms), for our Facebook service we impose specific Facebook Community Standards[60] dealing with issues around violence and incitement, dangerous individuals and organisations, child sexual exploitation and restricted goods and services in a way that aligns with what the draft Standards are targeting in the form of Class 1A material and Class 1B material, but is described in different terms. Ideally to protect online harmonisation, compatible parameters for policies should be incorporated in the Standards. The current wording creates ambiguity and risk of

[60] See Facebook, *Facebook Community Standards,* Transparency Centre, https://www.facebook.com/communitystandards/introduction

insensible interpretation that would require platforms having to develop Australia-specific versions of policies, which would impede effective enforcement at a global scale.

Simplifying reporting requirements

We recognise that, as a large company, the decisions we take relating to content on our services can be significant. That is why we have a dedicated Transparency Centre, where we report transparently on a wide range of areas – our community standards enforcement, government and law enforcement requests, content restrictions, widely viewed content – among others.[61]

Reporting obligations create accountability through transparency, but transparency is most effective where it is readily understood and cohesive and goes toward improving the integrity of online safety. Against the backdrop of reporting obligations throughout the Online Safety Framework, reporting obligations in the Standards should be streamlined to minimise duplication and avoid imposing an unreasonable administrative burden on service providers. The Office would bring clarity to the regulatory framework by providing a binding commitment to keep all information confidential and not disclose it to any third party without the service provider's express prior consent in writing, and to be clear on how the enforcement regime will apply to such extensive information production obligations. For example, under the BOSE Determination there is increased clarity around what a compliant response to a BOSE notice is. Such clarity in the Standards or indeed the consequence and utility of such extensive information production is needed.

If implemented in their current form, the draft Standards would impose extensive reporting obligations on affected service providers without clear safety objectives or utility. The requirements to produce risk assessments for services, new products, feature or functionality changes, alongside a 'development plan' per service and annual compliance reporting obligations per service across multiple codes create extensive documentation obligations without a clear nexus toward how this information production furthers the objectives of the Codes and commitments to improving online safety.

Ideally, information production obligations are clearly linked to the objectives of the regulation. For example, an annual report on how service providers meet the requirements of the code is an established accountability and transparency measure for platform regulation, as are implementing internal processes for safety by design assessments for new products, features or functionality. It is unclear how a forward-looking development plan or ongoing requirements to produce risk assessments connect to these more established transparency and accountability mechanisms.

For example, information that may need to be produced under the draft RES standard would include:

[61] Meta, *Transparency Centre*, https://transparency.fb.com/en-gb

- Copies of risk assessments for services, products of services and features and functionalities of those products; and
- An annual online safety development program; and
- Reports on the technical feasibility of complying with certain compliance obligations, and alternative actions taken (to the extent applicable) (section 34); and
- Notices regarding new service features or functions that may affect the risk profile of the relevant service (section 35); and
- Reports on activities and investments undertaken to implement relevant online safety development programs, and the outcomes of those activities and investments in terms of enhancing online safety for end-users in Australia (section 36); and
- An annual compliance report for each relevant RES setting out details of: the average number of monthly active users for each month; details of the most recent risk assessment (presumably only where one is required); a description of the service's functionalities and features; details of the steps taken to comply with the RES Standard; an explanation of why those steps were appropriate; a statement of the extent to which it was not technically feasible to detect or remove harmful material from the service and why; the amount of harmful material that was removed from the service; details of how the removed material was detected and identified; and the number of complaints made about the service provider's compliance with the RES Standard (section 37).

This is on top of existing reporting obligations under the OSA, including reporting on compliance with the BOSE Determination under Pt 4 of the OSA.

Cumulatively, these information production requirements impose significant administrative obligations that do not have a clear utility for service providers discharging their safety obligations under the Standards. The information in many cases will be duplicated across reports, risk assessments, development programs, and notices, or overlap with information that service providers already make publicly available (as we do through our quarterly Community Standards Enforcement Report[62]). The problem will be exacerbated if the Office requires reporting to be provided outside a regular cycle, in response to specific requests by the Office. In this regard, we note that the draft Standards give the Office significant discretion to determine both the form of the reports and the time within which they must be delivered. This may result in unrealistic information production requirements being imposed, as well as significant cross-over with reporting requirements under the broader Online Safety Act framework, without a clear framework under which all information produced is assessed, handled and published.

In our view, the reporting regime should be simplified and harmonised across the Online Safety Act framework, with a clear objective for different types of assessments and information

[62] See Meta, *Community Standards Enforcement Report*, Transparency Centre, https://transparency.fb.com/reports/community-standards-enforcement

production. For example, service Tiering based on size and scale may be best used to determine which obligations appropriately apply to large, medium or small services. Risk assessments for risk of exposure to Class 1A and Class 1B content may be best dealt with either in a separate risk assessment framework that is specifically for new products or new features / functionality, or more simply in Safety by Design obligations. Annual reporting on compliance may be best coupled with information about investments and activities to improve online safety ('development plan') given such investments and activities will be responding to the obligations under the Standards. And for all the above, information is produced at appropriate set intervals, with additional information only required out of cycle in very limited circumstances, noting that the Office also has investigative powers under Part 14 of the OSA that can be used to obtain more information if required to facilitate closer consideration of complaints or other matters arising under the OSA.

Finally, we have concerns about the volume and detailed nature of prospective technical information that may be required to fulfil reporting obligations as currently contemplated under the draft Standards, where such information is likely to contain sensitive and proprietary business information not specifically connected with online safety, and the disclosure of which may in fact be detrimental to online safety if accessed by bad actors. Across the reporting regime it would be beneficial to have clarity on the confidentiality or otherwise of such information required to be produced.

Recognising and adjusting obligations to allow for conflicts of law

The draft Standards do not explicitly state that providers are not obliged to undertake actions that will result in a breach of any applicable Australian or (in relation to foreign end-users) foreign privacy law or of any other Australian law or regulation by which the service provider is bound.

For example, it is an offence in Australia and numerous other jurisdictions to share child sexual exploitation material, including in the US where companies such as Meta are domiciled. Under both draft Standards, a service provider is required, as soon as practicable, to report any identified child sexual exploitation material on its service to an enforcement authority. In the absence of a clear statutory intent addressing conflict of law, 'enforcement authority' must be read as an *Australian* enforcement authority. Consequently, this places service providers at risk of breaching US law, specifically 18 U.S. Code § 2251 and 2252, in order to comply with the obligation in the draft Standards, which only permits service providers to share such material in limited circumstances, being a report to the National Centre for Missing and Exploited Children (NCMEC) in compliance with U.S. Code § 2258A. This could be remedied either by a clear conflict of laws carve out or in the alternative, the definition of 'enforcement authority' should be amended to mean 'an Australian, foreign or international' authority.

The draft Standards do not include any provisions to deal with conflicts of law. Whilst we acknowledge that to some degree, there are already rules dealing with conflicts between

delegated and primary legislation in Australia, explicit recognition of this in the draft Standards would sure that there is a clear signal to industry that there is no intention to create a potential conflict of laws or 'Catch-22' situation whereby a service provider could be forced to choose between either breaching an Industry Standard or breaching another law by which they are bound simply because of the impossibility of complying with both.This approach would be consistent with other Australian laws including:

- the Head Terms for the registered Industry Codes expressly include a provision to the effect that relevant service providers are not obliged to undertake actions that will result in a breach of any applicable Australian or (in relation to foreign end-users) foreign privacy law or of any other Australian law or regulation by which the service provider is bound.
- in relation to the industry assistance scheme, subsection 317ZB(5) of the Telco Act provides that, if a designated communications provider is required to do an act or thing in a foreign country, it is a defence if compliance with the requirement would contravene a law of the foreign country;[63]
- section 13D of the *Privacy Act 1988* (Cth) provides that an act or practice done or engaged in outside Australia is not an interference with privacy under that Act if it is required by an applicable law of a foreign country.;[64] and
- outside Australia, section 124(2(j)) of the recently passed *Online Safety Act 2023* (UK) provides that in considering whether it is necessary and proportionate to require service providers to use an accredited technology to identify and remove harmful content, Ofcom must consider 'the level of risk of the use of the specified technology resulting in a breach of any statutory provision or rule of law concerning privacy that is relevant to the use or operation of the service (including, but not limited to, any such provision or rule concerning the processing of personal data)'.[65]

  In all of the above examples, these laws expressly identify and seek to resolve the risk of conflicts arising with other laws. In our view, it is appropriate and necessary for the draft Standards to do the same.

[63] See *Telecommunications Act 1997* (Cth), https://www.legislation.gov.au/C2004A05145/latest/text/2
[64] See *Privacy Act 1988* (Cth), https://www.legislation.gov.au/Details/C2023C00347
[65] See *Online Safety Act 2023* (UK), https://www.legislation.gov.uk/ukpga/2023/50/enacted

Responses to Discussion Questions

Discussion Paper question	Summary of Meta's response
BOTH STANDARDS (DIS AND RES)	
1. Are the requirements for risk assessment in the draft Standards targeted at the right services and at the right points in a service's development journey? Are the risk factors appropriate?	The scope of the draft RES Standard, including definitions of different types of RES, could benefit from simplification, making the obligations unambiguous for industry to develop effective compliance systems and to support public fluency in the Standards. Please see the section *Simplifying the scope of the draft RES Standard* in our *General Comments* section above. The requirements for risk assessments and the risk factors in the draft Standards appear to fuse several concepts together creating a difficult assessment framework. Service 'Tiering' based on size and scale is a helpful and tested stratification approach to determine which obligations appropriately apply to large, medium or small services. It would be helpful for the option to automatically self-select as Tier 1 was consistent for all services across Standards. This allows services to make the decision to subscribe to the highest level of obligations, without requiring an additional administrative burden of justifying why. But size of a service is separate from the 'risk profile' of a service, and separate from an assessment of an inherent 'risks'. While the size of a service is helpful in determining the breadth of applicable obligations, the size of a service does not easily or automatically equate to risk of exposure to harmful content. Risk assessments for risk of exposure to Class 1A and Class 1B content may be best dealt with either in a separate / specific risk assessment framework targeted toward analysis of exposure risk specifically for new products or new features / functionality. However, these risk assessments may achieve the objective of ensuring new products and features have appropriate safety

	mechanisms more simply in Safety by Design obligations.
2. Do the obligations on each relevant electronic service and designated internet service category appropriately reflect the above considerations [to minimum compliance measures set out on pages 10-11]? Are other considerations relevant?	We welcome the approach taken to try to achieve the minimum compliance measures outlined on pages 10-11 of the Discussion Paper. The considerations are relevant, and we support these as guiding principles for the obligation framework, however the practical implementation of many of the obligations as contemplated in the draft Standards do not clearly meet the objective of these considerations. Please see our comments in the section *Concerns relating to technical feasibility* in our *General Comments* above.
DRAFT RES STANDARD	
3. Is the test in section 5 workable? Is further guidance required to assist providers to determine whether this standard, or another code or standard, applies to a particular online service?	The section 5 test could benefit from compatibility with the drafting of the same test under the Head Terms of the Codes whereby it is made clear a service can only fall under one Code or Standard and that the operator of the service has the discretion to determine which Code or Standard definition fits the primary purpose of the service. Using a consistent "primary purpose" test is preferable as services are rarely singular and usually multifunctional - containing elements of functionality that could be 'closely aligned' with multiple Codes / Standards. A clear direction that only a singular Code or Standard based on "primary purpose" would help to create clear scoping.
4. Is the technical feasibility exception in the obligation to detect and remove known child sexual abuse material and pro-terror material appropriate? How effective will this obligation be with this exception?	We welcome the acknowledgement of the technical limitations that can exist with respect to industry's ability to detect and remove the most harmful types of content. However, as outlined in the section *Concerns relating to technical feasibility* in our *General Comments* above, we have concerns about how "technical feasibility" is defined in the draft Standards, as it only refers to financial considerations without taking into account what is technically possible or technically effective. We have provided suggestions as to how this concept can be improved to avoid introducing systemic vulnerabilities or weaknesses

	into end-to-end-encrypted services.
5. Are there other examples of systems, processes and technologies that can detect, flag and/or remove known child sexual abuse material and known pro-terror material at scale, which should be highlighted in the Standards or accompanying guidance?	Apart from our concerns about the definition of "technical feasibility" noted above, we welcome the approach adopted in the draft Standards and accompanying guidance to avoid mandating a particular approach, system, process, or technology, so the service provider has the flexibility to implement systems and deploy technology that is appropriate to the characteristics of its service including the scale of the service. This allows service providers to focus on achieving the end goal of minimizing Class 1A/1B material in the most effective way they can.
6. Are there any limitations which would prevent certain service providers from deploying systems, processes and technologies to disrupt and deter child sexual abuse material and pro-terror material on relevant electronic services? If there are limitations, how might these be overcome? Is it appropriate for this requirement to apply to gaming services with communication functionality?	In our integrity work with respect to Meta's services, we have seen that our and industry's efforts are forcing threat actors to rapidly evolve their tactics in attempts to evade detection. These changes are likely an attempt by threat actors to ensure that any one service has only limited visibility into the entire operation. When bad actors count on us to work in silos while they target people far and wide across the internet, we need to work together as an industry to protect people. For this and for the reasons outlined in the section *Concerns relating to technical feasibility* in our *General Comments* above, we have concerns as to whether these obligations are appropriate for all the services contemplated as being in scope under the Standards. This includes for new and emerging technologies such as gaming services, especially as they become more immersive and 3D.
7. Are there other examples of systems, processes and technologies that can disrupt and deter the use of a relevant electronic service to solicit, generate, distribute or access child sexual abuse material and pro-terror material, which should be highlighted in the guidance?	We welcome the approach adopted in the draft Standards and accompanying guidance to avoid mandating a particular approach or technology, so the service provider has the flexibility to implement systems and deploy technology that is appropriate to the characteristics of its service including the scale, features and nature of the service.

8. Do you agree with the monthly active user threshold for the investment obligation? Are there other appropriate thresholds that should be considered to ensure the obligation is proportionate to the size and reach of the relevant electronic service?	We do not have a view on this.
9. Are the end-user reporting requirements workable for the relevant service providers? Are there practical barriers to implementation?	We support the requirement for appropriate end user reporting to be included in relevant services under the draft Standards. With respect to the commentary about the impact of end-to-end encryption on user reporting, we note – with respect to Meta's services – that when a user reports an account on encrypted services (e.g. on WhatsApp or encrypted Messenger), we receive a limited number of unencrypted messages from the user's device,[66] which allows us to review and take action against violating accounts. It is important that service providers are given flexibility in the design of end user reporting flows to ensure they are appropriate for the relevant service and feature and can be adjusted depending on changes in use of the service.
10. Should the requirement on certain relevant electronic services to respond to reports of class 1A and class 1B material on their service be limited to a requirement to take 'appropriate action'?	We support this language given the variety of measures and technical limitations that can be taken in response to the sharing of violating content.
11. What are your views on the likely compliance costs and, in particular, the impact of compliance costs on potential new entrants?	We support the principle of greater transparency and accountability through reporting, which is why Meta has been steadily increasing the scope and frequency of our transparency reporting over 10 years. We do have concerns about the extensive information production, ongoing risk assessments and reporting obligations that would be imposed and include suggestions for more simplified reporting under the draft Standards as outlined in our General Comments in *Simplifying the reporting requirements*. The type of

[66] WhatsApp, 'How to block and report contacts', Help Centre, https://faq.whatsapp.com/1142481766359885/?helpref=search&cms_platform=android; Messenger, 'How do I report an end-to end encrypted chat in Messenger?', Help Centre, https://www.facebook.com/help/messenger-app/498828660322839/?cms_platform=iphone-app&help

	information product contemplated in the drafts carries a significant investment that may not be proportionate to the impact these obligations have on improving safety and meeting the objectives of the Standards.
12. Is there any additional information eSafety should consider in determining the Relevant Electronic Services Standard?	Please see our *General Comments* above.
DRAFT DIS STANDARD	
13. Are the categories in Table 2 [of the defined categories of DIS on page 20 of the Discussion Paper] sufficiently clear for designated internet service providers to identify which category they fall within and therefore what obligations apply? What are the benefits and/or challenges of the categories as they are currently proposed?	The definitions of *High impact generative AI DIS*, *Machine learning model platform services* and *Enterprise DIS* are not sufficiently clear. Please see our comments in the *Regulation of Generative AI in the DIS Standard* section. Having separate Generative AI specific definitions may be redundant when Generative AI services will still be regulated under the broad category of 'designated internet service' to the same effect.
14. Are the section 9(5) matters in the draft Standard appropriate and sufficiently clear to help designated internet service providers accurately self-assess which tier their service falls within?	The Risk Assessment provisions of the DIS would benefit from greater clarity and synchronicity with other content regulation risk assessment frameworks, for example the four step approach proposed under the UK online safety legislation. Overall the Risk Assessment obligations here appear disconnected from similar obligations under the Codes by introducing new criteria for assessments, but without guidance on the weighting or otherwise of these criteria. The proposed methodology could provide clearer guidance on how to weigh risk against mitigations. For example, of the 11 factors listed 6 are factors that would negate a finding of risk (eg Terms of Use for the Service typically include prohibitions on illegal content, safety by design guidance, predominant functionality of the service) while others add weight to assumed risk without a clear or necessary connection to actual risk (eg communications functionality, the extent to which material on the service will be available to end-users of the Service in Australia). In particular, this factor on availability of material to end-users in Australia, is an

	example of a factor that may be of little utility in determining risk as all services would contain material available to end-users in Australia.
15. eSafety is seeking to place requirements on service providers that are best-placed to prevent the use of generative AI features to create and disseminate class 1A and class 1B material. Does the proposal achieve this?	No, they do not achieve this. Please see our comments in the *Regulation of Generative AI in the DIS Standard* section. Providers of user-facing Generative AI services are best-placed to prevent the use of generative AI features to create and disseminate class 1A and 1B material. The draft DIS Standard imposes obligations on *machine learning model platform services* and *enterprise DISs*, which are disproportionate to the limited control and oversight of these services.
16. Do the draft definitions for a high impact generative AI designated internet service and machine learning model platform service capture the right services? Are there types of providers that should not be included or should be excluded?	The draft definitions do not capture the right services. Please see our comments in the *Regulation of Generative AI in the DIS Standard* section. In particular, the drafting in the DIS Standard should make it clear that if providers of Generative AI services implement appropriate safeguards, the Generative AI service will no longer be considered a *high impact Generative AI DIS*.
17. The high impact generative AI designated internet services category only captures models that meet a high impact threshold. It must be reasonably foreseeable that a service can be used to generate synthetic high impact material that would be classified as X18+ or RC. Is this threshold: (a) appropriate for differentiating high impact and therefore high risk models? (b) sufficiently clear to enable service providers to assess whether or not they meet the definition? eSafety welcomes views on	It is not clear why a non-AI DIS qualifies as a "high impact DIS" when it has the "sole or predominant purpose" of making high impact material available - yet for a "high impact GenAI DIS" it is a much lower standard – merely that it is "reasonably foreseeable that the service could be used to generate synthetic high impact material". The threshold of what constitutes 'reasonably foreseeable' is unclear, and we recommend adopting the approach set out in the Fact Sheet, which states that "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". Please also see our comments in the *Regulation of Generative AI in the DIS Standard* section above.

alternative thresholds which may be more suitable.	
18. In relation to high impact generative AI designated internet services, do the proposed obligations (in particular, the section 21 obligation to 'detect and remove' and the section 23 obligation to 'disrupt and deter' child sexual abuse material and pro-terror material) provide appropriate safeguards? Are there specific challenges to deploying these measures in a generative AI context?	No, these do not provide appropriate safeguards. They overlook existing and appropriate safeguards that have been deployed within industry, including by Meta. Please see our comments in the *Regulation of Generative AI in the DIS Standard* section.
19. In relation to machine learning model platform services, do the proposed obligations (in particular, the section 23 obligation to 'disrupt and deter') provide appropriate safeguards? Are there specific challenges to deploying these measures?	No, these do not provide appropriate safeguards. They overlook existing and appropriate safeguards that have been deployed within industry, including by Meta. Most of the proposed obligations are not technically feasible with respect to *machine learning model platform service* providers, particularly for model distributors that have no control or oversight over downstream developers and their use cases. Please see our comments in the *Regulation of Generative AI in the DIS Standard* section.
20. In relation to relevant enterprise providers, do the proposed obligations (in particular, the section 23 obligation to 'disrupt and deter') provide appropriate safeguards? Are there specific challenges to deploying these measures?	No, these do not provide appropriate safeguards. They overlook existing and appropriate safeguards that have been deployed within industry, including by Meta. Most of the proposed obligations are not technically feasible with respect to *enterprise DIS* providers, and it is more appropriate to impose the majority of the obligations on the downstream account holders who provide the Generative AI services to end users. Please see our comments in the *Regulation of Generative AI in the DIS Standard* section.
21. Do sections 16 to 19 effectively reflect the considerations on minimum compliance measures outlined on pages 10-11 of this discussion paper?	No, we are concerned about the lack of nuance with respect to the differing roles and responsibilities of providers and developers within the ecosystem. Please see our comments in the *Regulation of Generative AI in the DIS Standard* section.

22. Do the obligations for detecting and removing child sexual abuse material and pro-terror material effectively reflect the considerations on minimum compliance measures outlined on pages 10-11 of this discussion paper?	No, we have concerns with respect to technical feasibility with respect to Generative AI products and services. No, these do not provide appropriate safeguards. They overlook existing and appropriate safeguards that have been deployed within industry, including by Meta. And they are not technically feasible with respect to many Generative AI products. Please see our comments in the *Regulation of Generative AI in the DIS Standard* section.
23. Is the technical feasibility exception in the obligation to detect and remove known child sexual abuse material and pro-terror material appropriate? How effective will this obligation be with this exception?	No, we have concerns that these are not appropriate. Please see our comments in the section *Concerns relating to technical feasibility* in our *General Comments* above.
24. Do you agree with this monthly active user threshold, or are there other thresholds which can be deployed to ensure this obligation is proportionate?	We have no comments on this.
25. What are your views on the likely compliance costs for service providers and, in particular, the impact of compliance costs on potential new entrants?	For online safety regulatory frameworks to achieve their goals, it is helpful for obligations to be directed toward platform investment in compliance systems that actively and effectively enhance safety, rather than compliance systems that are bespoke, may expire, or become technologically or practically less effective over time. Please see our comments above in *Simplifying the reporting requirements* for additional comments about concerns that these reporting requirements are focused more on compliance than increased investment in online safety.
26. Is there any additional information eSafety should consider in determining the Designated Internet Services Standard?	Please see our *General Comments*.