Meta

# Submission on the draft Online Safety (Basic Online Safety Expectations) Amendment Determination 2023

**February 2024**

## Executive summary

Meta appreciates the opportunity to contribute to the Department of Infrastructure, Transport, Regional Development, Communications and the Arts' (**Department**) consultation on the amendments to the Online Safety (Basic Online Safety Expectations) Determination 2022 (**BOSE**) in the draft Online Safety (Basic Online Safety Expectations) Amendment Determination 2023 (**BOSE Amendments**).

Meta supports the intent of the BOSE and its role within the *Online Safety Act 2021* (OSA) framework – namely, to increase the transparency and accountability of providers, thereby helping to incentivise and improve safety standards. We recognise that the intent of the BOSE is to provide the highest degree of flexibility to determine the most appropriate method of achieving the expectations.[1]

Meta has worked to ensure that we are responsive to the BOSE. We have submitted two formal responses to BOSE notices issued under s.56(2) of the OSA, one in relation to Facebook and Instagram and one in relation to WhatsApp, as well as responses to several formal follow-up questions from the Commissioner. Each notice contained over 30 questions seeking detailed information and data about Meta's products. Meta acknowledges the importance of transparency and worked to ensure that it provided responses to these questions within the rubric required by the regulator. This transparency under the BOSE comes in addition to our existing online safety investments and broader transparency and accountability.

Meta invests in industry-leading approaches to protect our users and build confidence in the integrity of our services with substantial investments made in recent years. At a global level, we now have around 40,000 people working on safety and security at Meta. We've invested over $20 billion since 2016 on safety and security, including around $5 billion in the last year alone.

We continue to introduce new features to help users manage their experience. These tools are informed by our consultations with industry, experts and civil society organisations. Our tools aim to discourage harmful behaviour, help users control their experience, and guide users to authoritative information and safety resources.[2] Since the BOSE Determination came into force in 2022, for example, we have announced additional tools and resources to give teens more age-appropriate experiences on our apps, such as placing teens into the most restrictive content control setting on Instagram and

[1] See, Online Safety (Basic Online Safety Expectations) Determination 2022 [Explanatory Statement] https://www.legislation.gov.au/F2022L00062/asmade/text/explanatory-statement

[2] See, for example, Meta, Safety Center, which provides information on how we are working to keep users safe on our platforms, https://about.meta.com/actions/safety

Facebook,[3] and stricter default message settings, meaning teens under 16 (and under 18 in certain countries) are not able to receive messages from anyone they do not follow or are not already connected to, providing more protection against potential scammers.[4] We have now developed more than 30 tools and resources to support teens and their parents.

We also continue to invest in cross-industry partnerships to improve online safety. For example, in February 2023, Meta and the National Center for Missing and Exploited Children (NCMEC) launched 'Take It Down', a new NCMEC portal created with support from Meta, designed to proactively prevent young people's intimate images from spreading online.[5] We have worked to expand 'Take It Down' to more countries and languages, allowing millions more teens to take back control of their intimate imagery.

Furthermore, we regularly provide updates on the efficacy of our work enforcing our policies, including on child safety. For example, we automatically disable accounts if they exhibit a certain number of the 60+ signals we monitor for potentially suspicious behaviour from adults.[6] We identified and removed more than 90,000 accounts from 1 August 2023 to 31 December 31 2023 as a result of this method.

We note that the BOSE Amendments are being considered at a time of considerable and ongoing online safety and digital reforms. At present draft industry standards for Designated Internet Services (DIS) and Relevant Electronic Services (RES) in relation to class 1A and class 1B material are under consideration following public consultation, a review of the OSA will shortly commence, and privacy reforms including the introduction of a Children's Online Privacy Code are being actively considered.

Against this backdrop of considerable and ongoing reform, it can become challenging for industry to invest in and design compliance systems to meet online safety expectations that are seemingly in a constant state of flux. To take one example – transparency reporting, in 2021, the regulatory impact statement for the Online Safety Act assumed that the regulatory burden of complying with the BOSE would be modest, and even for large businesses the assumption was that there would be '1 transparency report per year on average, and additional effort to uplift online safety practices, with 2 staff members

[3] Meta, 'New Protections to Give Teens More Age-Appropriate Experiences on Our Apps', Newsroom, 9 January 2024, https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps
[4] Meta, 'Introducing Stricter Message Settings for Teens on Instagram and Facebook', Newsroom, 25 January 2024, https://about.fb.com/news/2024/01/introducing-stricter-message-settings-for-teens-on-instagram-and-facebook
[5] Meta, 'New Updates to Help Prevent the Spread of Young People's Intimate Images Online', 27 February 2023, https://about.fb.com/news/2023/02/helping-prevent-the-spread-of-young-peoples-intimate-images-online
[6] Meta, 'Our Work to Help Provide Young People with Safe, Positive Experiences', Newsroom, 31 January 2024, https://about.fb.com/news/2024/01/our-work-to-help-provide-young-people-with-safe-positive-experiences/

taking 22.5 hours to produce'.[7] This has not held true – with the industry codes and draft industry standards requiring, at a minimum, 10 reports and the proposed BOSE Amendments now requiring even more transparency reporting.

We recognise the Government's concern to expand the BOSE to include expectations of industry to meet the best interests of the child, take action to combat hate speech and to ensure responsible innovation with respect to Generative AI. At Meta, we support the intent of ensuring that industry is investing appropriately in addressing each of these issues appropriately. However, we are concerned that in working to cover these the BOSE is straying from its intended purpose and becoming less flexible and more prescriptive by referencing specific types of content (namely hate speech) and specific technologies, such as recommender systems and generative artificial intelligence. Additionally, there may be duplicative requirements both within the existing reforms being considered within the OSA framework, the upcoming OSA reform and other proposed reforms such as the introduction of a Children's Online Privacy Code as part of the Privacy Act review, modelled on the UK's Age Appropriate Design Code.[8]

Moreover, by proposing the expansion of the BOSE to cover hate speech, we are concerned that without a review of the scope of the OSA, the proposed hate speech measures may result in the regulation of online hate speech being piecemeal and not meaningfully addressing the type of harmful content that can be experienced by individuals and vulnerable communities.

Given this, we suggest that the BOSE Amendments should be considered as part of the broader OSA review that will be commenced shortly. We look forward to continuing to engage constructively in the future discussions about further amendments to Australia's online safety regulatory framework.

[7] See Explanatory Memorandum, Online Safety Bill 2021 (Cth), p49 https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r6680_ems_3499aa77-c5e0-451e-9b1f-01339b8ad871/upload_pdf/JC001336%20Clean4.pdf;fileType=application%2Fpdf

[8] See Media Release, *Albanese government to strengthen privacy protection* https://ministers.ag.gov.au/media-centre/albanese-government-strengthen-privacy-protections-28-09-2023

# Table of contents

## Benefit of OSA Review & the BOSE Amendments

Meta supports the intent of the BOSE and has worked steadily to meet the transparency and accountability expected of digital platforms under them. We encourage the Government to consider several factors as part of finalising the BOSE Amendments such as the potential duplication with existing review processes and recently concluded online safety industry codes; and, the efficacy of the additional reporting under the BOSE Amendments.

On the issue of duplication, the recently announced OSA Review specifically seeks to consider many aspects that are covered in the BOSE Amendments. These include Generative AI, recommender systems, the best interests of the child standard, online hate and information gathering and information disclosure powers.

Additionally, some of these same aspects are already covered or proposed to be covered in the industry codes and the draft Standards. At present, the BOSE utilise terms that are consistent with the existing OSA framework; specifically, the BOSE apply to "social media services", "relevant electronic service of any kind" and "a designated internet service of any kind." However, the BOSE Amendments propose to make specific obligations with respect to recommender systems and Generative AI, despite these already being covered by the existing service descriptions of the existing BOSE. For example, recommender systems generally form part of social media services, which are covered by the industry codes and Generative AI is proposed to be covered by the draft DIS Standards. It is not clear why the BOSE Amendments propose specific and additional requirements on subsets of services that are already covered under the OSA, Codes and draft Standards.

With respect to recommender systems, the rationale for singling these out for special mention in the BOSE Amendments is unclear, when ranking or search systems are not. Additionally, given there are already obligations on GenAI under the draft DIS Standard, it is not clear why this technology merits further and additional regulatory expectations in the BOSE.

As well as creating uncertainty, constant and ad hoc amendments to the regulatory framework including the BOSE Amendment will divert time and resources away from the real work of developing and improving methods of addressing online safety risks effectively.

These are potential areas for overlap when considered against the other and additional reviews that are outlined in the Consultation Paper.[9]

We encourage the Government to consider clearer direction and predictability for industry by ensuring that any existing concerns be identified and addressed at a high level via the review of the Online Safety Act before deciding what changes, if any, need to be flowed down into the BOSE or reserved to be addressed at an operational level via industry codes or standards.

We also welcome further consideration of the value of the additional transparency reporting requirements outlined in the BOSE Amendments. We note that companies such as Meta already provide significant transparency measures on a regular basis.[10] The BOSE Amendments suggest that there should be additional reporting, specific to Australia, at regular intervals of between one to 12 months. As currently proposed, this requirement overlaps with some of the existing reporting requirements under Subdivisions 3A and 3B of Part 4 of the Online Safety Act which already provide the Commissioner with the powers to require service providers to report on their compliance with the BOSE and which may take the form of periodic report notices, periodic report determinations, non-periodic report notices and non-periodic report determinations. The Commissioner has already exercised these powers and has issued notices to Meta and other providers which required substantial resources and time to respond to considering the level of detail required, and in the format requested.

The proposed reporting requirements are that a service will publish regular (between one and 12 months) transparency reports to include: the service's enforcement of its terms of use, policies and procedures; the safety tools and processed deployed by the service and their effectiveness; metrics on the prevalence of harms, report and complaints, and the service's responsiveness, and; the number of active end-users of the service in Australia (including children) each month during the reporting period.

These proposed reporting requirements will create overlap with reporting obligations under the Class 1 industry codes currently in operation, and with the industry standards currently in development by eSafety, and potentially with the future Class 2 industry codes as well. For example, it is a requirement under the Social Media Services Online Safety Code (SMS Code) that a service provider takes enforcement action against end-users for breaches of their terms of use regarding Class 1 and Class 1 material. It is a further requirement that Tier 1 service providers (and Tier 2 providers at the written request of the Commissioner) submit annual reports to include the steps that the

[9] Department of Infrastructure, Transport, Regional Development, Communications and the Arts, *Amending the Online Safety (Basic Online Safety Expectations) Determination 2022 - Consultation paper*, November 2023, https://www.infrastructure.gov.au/sites/default/files/documents/amending-the-online-safety-basic-online-safety-expectations-determination-2022-consultation-paper-november2023.pdf

[10] For example, Meta publishes regular reports to give our community visibility into how we enforce our policies, respond to data requests and protect intellectual property: https://transparency.fb.com/reports/

provider has taken to comply with minimal compliance measures.[11] The annual report under the SMS Code will therefore include details of enforcement action taken by the service. This is duplicative of the proposed requirement in the BOSE Amendments for services to publish reports with the service's enforcement of its terms of use.

In addition to being duplicative, these reporting requirements seem to have moved beyond the Government's original intention for reporting requirements under the Online Safety Act, as set out in the Discussion Paper for Online Safety Legislative Reform that preceded the introduction of the Act, which specified:

> *To minimise the burden on social media services, a single reporting framework would be established. This would, to the fullest extent possible, integrate the reporting requirements of the proposed basic online safety expectations, the transparency recommendation of the Taskforce to Combat Terrorist and Extreme Violent Material Online, the OECD's voluntary transparency reporting protocol (when completed), and the UK's draft transparency reporting template, developed as part of the UK Government's Online Harms White Paper process. It is not expected that companies would have multiple separate transparency reporting obligations, as this would be duplicative and onerous.*[12] (Emphasis added).

Consistent with this, the regulatory impact statement for the Online Safety Act assumed that the regulatory burden of complying with the BOSE would be modest, and even for large businesses the assumption was that there would be '1 transparency report per year on average, and additional effort to uplift online safety practices, with 2 staff members taking 22.5 hours to produce'.[13] This has not been the case in practice, with service providers already grappling with several layers of detailed reporting, which requires significant resources to address in excess of the assumptions made in the regulatory impact statement. The additional reporting obligations contemplated under the BOSE Amendments will aggravate this issue, with no clear objective outcome of what would be achieved in requiring regular transparency reports in addition to the existing reporting obligations under the online safety framework

Indeed, there is a real risk of the reporting obligations becoming a distraction in themselves and drawing focus away from the real work of addressing safety harms in practice. In its current form, the BOSE Amendment obligations would impose extensive reporting obligations without an online safety objective distinct from the online safety objectives which the reporting obligations in the Codes and Standards are seeking to meet.

---

[11] Social Media Services Online Safety Code (Class 1A and Class 1B Material) Minimum Compliance Measures 3, 12, 32 & 33.
[12] See Department of Communications and the Arts, *Online Safety Legislative Reform - Discussion Paper*, December 2019, p23 https://www.infrastructure.gov.au/sites/default/files/consultation/pdf/online-safety-legislation-reform-discussion-paper.pdf

[13] See Explanatory Memorandum, Online Safety Bill 2021 (Cth), p49 https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r6680_ems_3499aa77-c5e0-451e-9b1f-01339b8ad871/upload_pdf/JC001336%20Clean4.pdf;fileType=application%2Fpdf

Consistent with our suggestion that the BOSE Amendments be considered as part of the broader OSA Review, we encourage the Government to consider how to streamline the reporting obligations that currently apply under the OSA Framework with reporting obligations clearly linked to the expectations under the BOSE, without duplication, and with transparency for industry as to how the production of reports furthers the objectives of the expectations.

## BOSE Amendments Potentially Losing Their Flexibility

We are concerned that the BOSE Amendments move the BOSE away from its originally stated intention to be a flexible regulatory instrument that sets broad expectations that the online services industry should strive to meet in order to keep Australians safe online. As the Explanatory Memorandum to the *Online Safety Bill 2021* (**Explanatory Memorandum**) outlined:

> *The basic online safety expectations will be a set of expectations that the Australian Government expect service providers to meet in order to uphold the safety of Australian end-users on their services, but also allow them flexibility in the method of achieving these expectations.*
>
> *...*
>
> *Service providers are best placed to identify these emerging forms of harmful end-user conduct or material, and so the flexibility of this regime means that providers can choose the best way to address them on their service in the most responsive way.*[14] (Emphasis added)

While we support the purpose and objectives of the BOSE as a regulatory instrument to set baseline standards for online safety, we are concerned that by including a focus on specific types of technology – such as Generative AI and recommender systems – and specific categories of content types such as individually-directed hate speech, the proposed amendments are moving away from the original purpose of the BOSE.

## Specific comments on BOSE Amendments

In an effort to assist the Government in ensuring that the BOSE remain fit for purpose as Australians' use of technology and recent innovations in technology evolve and change, we share some specific comments on aspects of the BOSE Amendments.

[14] See Explanatory Memorandum, Online Safety Bill 2021 (Cth), p91 https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r6680_ems_3499aa77-c5e0-451e-9b1f-01339b8ad871/upload_pdf/JC001336%20Clean4.pdf;fileType=application%2Fpdf

## Align 'best interests of the child' internationally & with proposed Australian privacy reforms

Meta supports the inclusion of a requirement that service providers consider the best interests of the child. However, we suggest that clause 6(2A) be amended to align with the UK Information Commissioner's Office Age Appropriate Design Code (**UK Code**) to ensure consistency across international regimes.

Aligning a 'best interests of the child' standard in the BOSE Amendments with the UK Code would mean minor adjustments to how it is currently proposed in the BOSE Amendments.

For example, clause 6(2A) in the BOSE Amendments sets an expectation that:

> *The provider of the service will take reasonable steps to ensure that the best interests of the child are a primary consideration in the design and operation of any service that is used by, or accessible to, children.*

This is modelled after Article 3 of the United Convention of the Rights of the Child. It is also closely aligned with Article 1 of the UK Code, which provides:

> *The best interests of the child should be a primary consideration when you design and develop online services likely to be accessed by a child.*[15]

Key differences are that the BOSE Amendments would apply to any service that is 'accessible' to a child, whereas the equivalent requirement under the UK Code would be limited to services that are 'likely to be accessed' by a child. We see this as a subtle but important distinction.

The UK Code notes that it is not intended to 'cover *all* services that children could *possibly access*'[16], which is what the current proposed wording under the BOSE Amendments would capture. That is, clause 6(2A) would capture any services that could theoretically be accessed by children, even if they are not designed for or aimed specifically at children or likely to be used by children at all. We suggest that the more limited scope in the UK Code is more appropriate because industry can most effectively direct their resources in making the interests of children a primary consideration only for services that children are likely to ever access.

This is consistent with the Australian Government's proposed privacy reforms. In its response to the Privacy Act Review report, the Government agreed to develop a Children's Online Privacy Code that would align with international approaches including

[15] See UK Information Commissioner's Office, *Age-Appropriate Design: a Code of Practice for Online Services*, p7 https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services-2-1.pdf

[16] See UK Information Commissioner's Office, *Age-Appropriate Design: a Code of Practice for Online Services*, p17 https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services-2-1.pdf

the scope of the UK Code. This proposed code would apply to online services that are 'likely to be accessed by children'.[17]

We also note that the standard of 'best interests of the child' and 'duty of care' have also been included as factors for consideration in the OSA Review[18], which again points to the utility of having these considerations folded into the broader OSA Review.

## Ensuring consistency in combating online hate speech

Meta shares the Government's intent to ensure that people who use online services are not subjected to hate speech. We have long-standing policies that prohibit hate speech and have steadily increased our investment in proactive detection technology over the years such as that, for example, in Q3, 2023, we proactively detected and actioned 94.8 percent of hate speech content on Facebook and 96.5 percent of hate speech content on Instagram, before people reported it[19] Whilst we have always removed hate speech when becoming aware of it, the increasing use of AI to identify hate speech has meant that we are able to action it more often before people are exposed to it.

Our progress is due in large part to our recent AI advances in a few areas:

- *Lingual understanding:* the ability to build machine learning classifiers that can analyse the same concept in multiple languages - and learning in one language can improve its performance in others. This is particularly useful for languages that are less common on the internet.
- *Whole post understanding or WPIE:* the ability to look at a post in its entirety, whether it's images, video and text, and look for various policy violations simultaneously instead of having to run multiple different classifiers.

We also use artificial intelligence to prioritise content that needs reviewing, after considering several different factors:

- *Virality:* Content that is potentially violating that's being quickly shared will be given greater weight than content that is getting no shares or views.
- *Severity:* Content that's related to real-world harm such as suicide and self-injury or child exploitation will be prioritised over less harmful types of content such as spam.
- *Likelihood of violating:* Content that has signals which indicate that it may be similar to other content that violated our policies will be prioritised over content which does not appear to have violated our policies previously.

---

[17] Attorney-General's Department, *Government response to the Privacy Act Review Report*, 28 September 2023, pp13, 30 https://www.ag.gov.au/sites/default/files/2023-09/government-response-privacy-act-review-report.PDF

[18] See Department of Infrastructure, Transport, Regional Development, Communications and the Arts, *Terms of Reference – Statutory Review of the* Online Safety Act 2021, February 2024, https://www.infrastructure.gov.au/sites/default/files/documents/tor-statutory-review-online-safety-act-2021-8Feb.pdf

[19] Meta, Community Standards Enforcement Report, Q3, 2023, https://transparency.fb.com/reports/community-standards-enforcement/hate-speech/facebook

Prioritising content in this way, regardless of when it was shared on our services or whether it was reported by a user or detected by our technology, allows us to get to the highest severity content first.

Whilst we share the same objective of the BOSE Amendment proposals with respect to hate speech, we have concerns about whether and how it can be fully included in the BOSE as presently envisaged.

The OSA, the BOSE and the industry codes all cover the same types of harmful content (in varying degrees): cyberbullying material targeted at an Australian child; cyber-abuse material targeting an Australian adult; image-based abuse material; Class 1 and Class 2 content; and material depicting, promoting, inciting or instructing in abhorrent violent conduct. These concepts are clearly defined in the OSA itself and underpin the entire OSA framework. Whilst it may be possible that some forms of hate speech fall within the category of cyberbullying targeted at a child or cyber-abuse material directed at an adult, many forms of hate speech will not fall within these categories because it is frequently targeted at a group rather than an individual. We recognise that the infrastructure that has been set up to support the implementation of the OSA may be helpful in supporting oversight of a broader category of online harms including hate speech. However, to ensure that the protection of vulnerable groups in Australian society is not uneven, we suggest that this may more properly be considered by the OSA Review.[20]

If the requirements with respect to hate speech are retained within the BOSE Amendments, we suggest that the current definition of hate speech is too broad and may go beyond its original intent. Service providers such as Meta have built considerable nuance and protections balancing removal of hate speech against public interest and debate.[21] The definition of "hate speech" proposed would capture *anything* that breaches our terms of service, whereas we suggest it would be more appropriate (and more reflective of the amendments' intent) for the requirements to apply only to hateful content *that* breaches our terms of service.

---

[20] The Terms of Reference for the review of the *Online Safety Act 2021* expressly includes consideration of '[w]hether additional arrangements are warranted to address online harms not explicitly captured under the existing statutory schemes, including[] online hate': see Department of Infrastructure, Transport, Regional Development, Communications and the Arts, *Terms of Reference – Statutory Review of the* Online Safety Act 2021, February 2024, https://www.infrastructure.gov.au/sites/default/files/documents/tor-statutory-review-online-safety-act-2021-8Feb.pdf

[21] Please see Meta, *Facebook Community Standards - Hate Speech*, https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/; see also, Meta, 'Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?', Newsroom, 27 June 2017, https://about.fb.com/news/2017/06/hard-questions-hate-speech/

## Align Generative AI requirements with BOSE purpose & other governance frameworks

We appreciate that Generative AI has been the focus of considerable debate over the past 18 months. However,  AI is not new. 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 including object detection, unsupervised machine translation, and large language models – which in turn have had global, real-world impact.[22] Additionally, we have widely deployed AI on our services to address many of the harms that the OSA Framework seeks to regulate.

That said, we very much appreciate the concern that any new technologies should be designed with a 'safety by design' approach and consistent with the principles of responsible innovation.[23]

Given this, it is understandable that there is consideration given as to if, and if so, how best, to update the BOSE to ensure that it is effective within the context of Generative AI.

At present, the issue of Generative AI has been the subject of much international debate and governance discussions, including the Bletchley Declaration to which Australia is a signatory. Additionally, Generative AI is expressly mentioned for regulation in the draft DIS Standard and for further consideration as part of the OSA Review.

For this reason, it continues to make sense that any inclusion of Generative AI be considered as part of the broader OSA Review. If there is a desire to persist with inclusion of Generative AI within the BOSE Amendments in advance of the conclusions of the OSA Review, we suggest that consideration be given to ensure consistency of terminology with the draft DIS Standard and the international frameworks such as the Bletchley Declaration and also appropriate flexibility in keeping with the general approach of the BOSE.

With respect to terminology, at present, clause 8A in the BOSE Amendments imposes specific requirements on generative artificial intelligence and refers both to 'generative artificial intelligence' and then more generically to 'artificial intelligence' with neither term being defined. We suggest that any references to 'artificial intelligence' should be

[22] 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
[23] See Meta, *Meta's five pillars of responsible AI that inform our work*, https://ai.meta.com/responsible-ai

amended to 'generative artificial intelligence' to ensure consistency and avoid extending the scope of any new expectations beyond what was intended, and then made consistent with the terms used in the draft DIS Standard and the Bletchley Declaration.

Given the approach of the BOSE is to allow industry to meet basic online safety expectations but have flexibility on how best to achieve this, we also suggest that the requirements in the BOSE Amendments to take reasonable steps as they relate to Generative AI are clarified. At present, the BOSE Amendments contain a requirement to detect and prevent prompts that may be used to manipulate generative AI into producing unlawful or harmful material. Specifically, we suggest that clause 8A(3)(d) is adjusted to make it less prescriptive and to recognise the limitations of what service providers can achieve at different levels of the Generative AI ecosystem.[24]

At present, clause 8A(3)(d) in the BOSE Amendments provides that reasonable safety measures for generative AI could include:

> *ensuring that generative artificial intelligence capabilities can detect and prevent prompts that generate unlawful or harmful material.*

The BOSE was intended to be a regulatory instrument with flexibility to cover online safety issues that are still evolving but without strict 'black letter' requirements which necessitate strict compliance, with such requirements sitting in the Online Safety Act itself and in the underlying Codes and Standards. In this way, the BOSE was intended to be future-proof.

Whilst Meta agrees in principle that steps should be taken to ensure that generative artificial intelligence services cannot be misused to generate unlawful or harmful material, and that prompt controls are an important way to do this, at the same time, it may be the case that it is useful for testing or creation of bespoke models that they can accept a wide range of prompts. Additionally, it is important to bear in mind the limitations of what is technically feasible, and to recognise that, despite the best efforts of service providers, ordinary language prompts may in some circumstances be used in a deliberate way to manipulate generative AI services so as to produce unintended outputs.

This has been specifically acknowledged by the eSafety Commissioner in its Discussion Paper on the Draft Online Safety (Relevant Electronic Services – Class 1A and 1B Material) Industry Standard 2024 and Draft Online Safety (Designated Internet Services – Class 1A and 1B Material) Industry Standard 2024 released in November 2023:

> *Importantly, eSafety is not proposing that a designated internet service with generative AI features needs to completely rule out the possibility of high impact material ever being generated on its service. Given the nature of generative AI*

[24] See e.g., Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E. Ho, Arvind Narayanan, Percy Liang, *Considerations for Governing Open Foundation Models* (Dec 2023) https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models

*models, there may be risks that an end-user could, with sufficient effort, manipulate the model producing harmful material despite safeguards being built in.*[25]

At present, we are concerned that Clause 8A(3)(d), as drafted, is a strict 'black letter' requirement which necessitates strict compliance (which was not the intention of the BOSE) and which will require service providers to *ensure* that prompts that generate unlawful or harmful material can be detected and prevented. Taking into account the concerns and technical limitations outlined above, this sets an unrealistically high bar for service providers. If clause 8A(3)(d) is to be retained, we consider that the current reference to 'ensuring' should be replaced with more flexible and future-proof language such as 'to the extent reasonably practicable, ensuring' or 'taking reasonable steps to ensure'.

## Clarifying expectations with respect to recommender systems

We appreciate the concern about the role of AI in ranking and recommending content. This is why we have provided transparency about how our ranking and recommendation systems work. For example, in 2021, we published the Content Distribution Guidelines to share more detail on the types of content that we demote in Facebook Feed.[26] While the Community Standards make it clear what content is removed from our services because we don't allow it, the Content Distribution Guidelines make it clear what content receives reduced distribution on Feed because it is problematic or low quality.

The changes we make, particularly ones focused on limiting the spread of problematic content, are based on extensive feedback from our global community and external experts. There are three principal reasons why we might reduce the distribution of content:

- **Responding to people's direct feedback.** We listen to people's feedback about what they like and don't like seeing on Facebook and make changes to Feed in response.
- **Incentivising creators to invest in high-quality and accurate content.** We want people to have interesting new material to engage with in the long term, so we're working to set incentives that encourage the creation of these types of content.
- **Fostering a safer community.** Some content may be problematic for our community, regardless of the intent. We'll make this content more difficult for people to encounter.

[25] 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, p24, https://www.esafety.gov.au/sites/default/files/2023-11/Discussion-Paper-draft-Online-Safety-Standards-%28Class-1A-and-1B%29.pdf

[26] Meta, 'Types of content we demote', *Transparency Centre,* 20 December 2021, https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/

Across our apps, we make personalised recommendations to help users discover new communities and content we think they are likely to be interested in. Some examples of our recommendations experiences include Pages You May Like, "suggested for you" posts in Feed, People You May Know or Groups You Should Join.

It is important that we have high standards for what we recommend. This helps ensure we don't recommend potentially sensitive content to those who don't explicitly indicate that they wish to see it. As noted above, our Recommendations Guidelines set a higher bar than our Community Standards, and content may be removed from recommendations even if it does not violate our Community Standards.

To help people better understand our approach to recommendations, in August 2020, we published a set of Recommendation Guidelines, which outline the types of content that may not be eligible for recommendations.[27] In developing these guidelines, we consulted 50 leading experts specialising in recommendation systems, expression, safety and digital rights. Recommendation Guidelines are available for both Facebook[28] and Instagram.[29]

Given the large and growing volume of content that is shared online, algorithmically organised content is an important feature of ensuring that people continue to see and engage with the most relevant content to them.

For example, one of the ways that people connect with friends, family and other accounts that they follow is via a "Feed".

Historically, these feeds showed content in chronological order. However, as more people started using our services, more content was shared and it was impossible for people to see all of the content that was shared, much less the content that they cared about. Instagram, for example, launched in 2010 with a chronological feed but by 2016, people were missing 70 per cent of all their posts in Feed, including almost half of posts from their close connections. So we developed and introduced a Feed that ranked posts based on what people cared about most.[30]

---

[27] G Rosen, 'Recommendation guidelines', *Meta Newsroom,* 31 August 2020, https://about.fb.com/news/2020/08/recommendation-guidelines/
[28] Facebook, 'What are recommendations on Facebook?', Help Centre, https://www.facebook.com/help/1257205004624246
[29] Instagram, 'What are recommendations on Instagram?', Help Centre, https://help.instagram.com/313829416281232
[30] See e.g., A Mosseri, 'Instagram Ranking Explained, Instagram Blog, 31 May 20238 June 2021, https://about.instagram.com/blog/announcements/instagram-ranking-explained/ https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works

We provide this personalised experience via AI. Our ranking algorithms use thousands of signals to rank posts for each person's Feed with this goal in mind.[31] Our ranking system personalises the content for over a billion people and aims to show each of them content we hope is most valuable to them, every time they come to Facebook or Instagram.

The goal is to make sure people see what they will find most meaningful — not to keep people glued to their smartphone for hours on end.

That said, ranking and recommendation systems are not new and considerable work has been underway for many years to provide transparency and controls to consumers to adjust settings.

With the OSA Review proposing to consider recommender systems, we suggest that the OSA review is also a suitable vehicle to consider both ranking and recommender systems. However, if the Government wishes to pursue the BOSE Amendments with a specific focus on recommender systems, we suggest minor adjustments are made to the proposed requirements to be consistent with the existing industry practices, pending a more fulsome review of these under the OSA Review.

It is not clear how the requirement in clause 8B(3)(c) that users should be able to make complaints or enquiries about the operation of recommender systems can work. This would be impracticable as it would first require users to have a clear understanding of what recommender systems are, which ones impact their online user experience and how they work, which would be different for every platform they use. Without a foundational understanding of how recommender systems work, this requirement risks creating a floodgate of superfluous or vexatious user complaints and enquiries which platforms would need to expend resources to respond to, without any increase in safety and well-being. It would be more effective and meaningful to empower users by giving them greater transparency into how the recommender systems work and user controls that allows them to customise their online experience. The EU Digital Services Act, for example, requires platforms to set out in their terms and conditions the main parameters used in their recommender systems, including any available options for users to modify or influence said parameters.

At Meta, we provide a number of Transparency tools, with links through to the ability to adjust these settings. These include:

[31] A Lada, M Wang, 'How does News Feed predict what you want to see?', Meta Newsroom, 26 January 2021, https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/

- *Why Am I Seeing this post?* - helps users to better understand and more easily control what they see from friends, Pages and Groups in their News Feed. Users are able to tap on posts and ads in News Feed, get context on why they are appearing (such as how their past interactions impact the ranking of posts in their News Feed), and take action to further personalise what they see.[32]
- *Why Am I seeing this Ad?* - provides users with context on their ads, to help them to understand how factors like basic demographic details, interests and website visits contribute to the ads in their News Feed. We are continually improving our transparency offerings to reflect feedback we receive. In 2023, we updated this tool to provide users with clear information about the machine learning models that help determine the ads they see on Facebook and Instagram Feed.[33]
- *Ad Preferences* - allows users to adjust the ads they see while on Facebook and gives them the ability to update their ad settings to control information we can use to show their ads.[34]
- *Control what you see on Facebook and Instagram* - helps users to learn more about and control what kind of posts they may see on Facebook and Instagram, including who they see posts from.[35]
- *Content recommendation controls* - our content recommendation controls - known as "Sensitive Content Control" on Instagram and "Reduce" on Facebook – make it more difficult for people to come across potentially sensitive content or accounts in places like Search and Explore.[36]

We have worked closely with external privacy experts and policy stakeholders from around the world to get input on what transparency changes they want to see.

Within the Transparency Center, we share 25 system cards for Facebook and Instagram that explain how the AI systems in our products work.[37] They give information about how our AI systems rank content, some of the predictions each system makes to determine

[32] Facebook, 'What influences the order of posts in your Facebook Feed', Help Center, https://www.facebook.com/help/520348825116417; Meta, 'Why Am I Seeing This? We Have an Answer for You', Newsroom, 31 March 2019, https://about.fb.com/news/2019/03/why-am-i-seeing-this

[33] Facebook, 'How does Facebook decide which ads to show me?', Help Center, https://www.facebook.com/help/562973647153813/?helpref=uf_share; Meta, 'Increasing Our Ads Transparency', Newsroom, https://about.fb.com/news/2023/02/increasing-our-ads-transparency, Newsroom, 14 February 2023

[34] Facebook, 'Your Ad preferences and how you can adjust them on Facebook', Help Center, https://www.facebook.com/helIncreasing Our Ads Transparency | Metap/247395082112892

[35] Facebook, 'Control what you see in Feed on Facebook', Help Center, https://www.facebook.com/help/1913802218945435/?helpref=uf_share; Instagram, 'How Instagram Feed Works', Help Center, https://help.instagram.com/1986234648360433

[36] Meta, 'Introducing Sensitive Content Control', Newsroom, 20 July 2021, https://about.fb.com/news/2021/07/introducing-sensitive-content-control; Facebook, 'Manage how content ranks in your Feed using Reduce', Help Center, https://www.facebook.com/help/543114717778091

[37] Meta, 'Our approach to explaining ranking', Transparency Center, updated 31 December 2023, https://transparency.fb.com/features/explaining-ranking

what content might be most relevant to users, as well as the controls users can use to help customise their experience.

Given this, it is not clear what more can be done by a service provider to respond to enquiries or complaints. If the BOSE Amendments require adjustments now before the OSA Review is completed, we suggest that this requirement be amended to make clear that service providers must provide users with guidelines, transparency and tools to manage the settings of recommender systems.

## Streamlining information sharing and reporting obligations

We appreciate the focus of the BOSE on encouraging the transparency and accountability of industry, but suggest that before further reporting requirements are introduced into the BOSE, a holistic review is undertaken as part of the OSA Review of the existing reporting obligations on service providers under the industry codes and draft Industry Standards.

We note that there are existing reporting obligations under the BOSE, which have been utilised effectively by the eSafety Commissioner, with detailed findings published on the Commissioner's website.[38] In our view, additional mandatory reporting obligations are not required at this point in time and will add significant administrative overhead cost of complying with these requirements).

We suggest that any reform efforts as they relate to reporting by industry of online safety efforts should be focussed on simplifying and streamlining reporting requirements, rather than adding new requirements for the sake of compliance. We note that this is not a question of not wanting to be transparent – we already voluntarily provide detailed information about the operation of our services, including in relation to enforcement of our community standards[39] - but is rather aimed at ensuring the reporting burden remains balanced and proportionate for all service providers. As noted earlier in our submission, at a minimum, the industry codes and draft industry standards require at least 10 reports and the BOSE Amendments now proposing even more transparency reporting. Given the extent of existing transparency and reporting measures, we suggest that the additional transparency and reporting requirements under clauses 18A and 20(5) in the BOSE Amendments should be removed.

We also suggest that further discussions need to be had with industry in relation to information sharing given the obligations under applicable laws and relevant terms of service.

---

[38] See eSafety Commissioner, Responses to transparency notices, https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices

[39] See Meta, Transparency reports, Transparency Center, https://transparency.fb.com/reports/