Microsoft submission on **Online Safety (Basic Online Safety Expectations) Amendment Determination 2023**

Submission to the Department of Infrastructure, Transport, Regional Development, Communications, and the Arts

14 February 2024

Introduction

Microsoft welcomes the opportunity to respond to the Department of Infrastructure, Transport, Regional Development, Communications and the Arts' (the **Department**) exposure draft of the *Amending the Online Safety (Basic Online Safety Expectations) Determination 2023 – Consultation Paper* (the **consultation paper**).

At Microsoft, digital safety is core to how we design experiences for people to create, connect and share knowledge online. In Australia, the Basic Online Safety Expectations (**BOSE**) Determination, made pursuant to the *Online Safety Act 2021* (Cth) (the **OSA**), plays an important role in outlining measures that can support such safety by design approaches and enhance the online safety of Australians.

We offer the following feedback on the proposed amendments in the consultation paper:

1. We recommend a holistic review of whether online harms are being effectively addressed under the OSA, including:
   a. reviewing the BOSE regime as part of the broader independent review of the OSA due to commence shortly;
   b. clarifying and improving consistency of definitions under the BOSE regime with the broader OSA, e.g. in relation to 'harmful material', and
   c. examining the broad and expanding reach of the BOSE regime to ensure the BOSE remain able to accommodate the breadth and diversity of in-scope services.

2. It is not clear that additional BOSE to regulate generative AI technologies is necessary as emerging technologies are already captured within the existing, technology-neutral drafting of the BOSE Determination. Moreover, online harms of the type regulated under the BOSE Determination are best addressed by interventions at the application layer of the AI technology stack. There are inherent limitations in regulating online harms at the model layer of the AI stack where end-users do not interact with the technology.

3. Transparency is a critical part of accountability and building trust. It also needs to be balanced with competing priorities such as managing confidential business information and protecting safety systems and processes from exploitation. In particular, the BOSE regime may benefit from additional guardrails on regulatory discretion to request, disclose, and publish confidential information.

4. We urge the Department to reconsider the addition of duplicative requirements under the draft Determination, such as the annual reporting requirement. These requirements overlap significantly with existing reporting requirements under the Online Safety Codes and Standards as well as other legislation including the misinformation and disinformation code and telecommunications laws.

1. Microsoft's approach to digital safety

Microsoft recognises the unique role that technology companies play in helping make the internet safer. Service providers should seek to design and operate their services responsibly, while anticipating and reducing digital safety risks unique to each of their services. Microsoft has a long-standing commitment to digital safety, as well as a history of working closely with Governments, industry, civil society organisations, and academia to reduce the presence of illegal and harmful online content.

At Microsoft, we recognise that we have a responsibility to protect our users from illegal and harmful content, particularly our youngest users. In doing so, we equally must balance our commitments and obligations to respect human rights, including privacy, freedom of expression, and access to information. As a company with a diverse range of products and services, we aim to strike this balance through a risk-proportional approach: that is, by tailoring our safety interventions depending on the nature and characteristics of a service and the type of harm that we are seeking to address.

2. General commentary on the Proposed Amendments

This section of our submission covers matters that are common across the BOSE scheme that are highlighted or expanded upon in the consultation paper.

2.1 The role of transparency

Transparency is a key aspect of how the technology sector can build trust and help users understand the safety measures in place on a service. Transparency measures also play an important role in accountability and in helping to understand a particular service's safety outcomes. We publish a range of corporate transparency reports, including a bi-yearly voluntary Digital Safety Content Report. Recognising the unique features of gaming, we also now publish a regular Xbox transparency report.

It is important that transparency is meaningful and provides data points and context that enable governments, civil society, and end-users to contextualise, understand, and advance online safety outcomes. Different audiences are likely to have different information needs. Transparency also needs to be balanced with competing priorities such as managing confidential business information and protecting safety systems and processes from exploitation by bad actors. We encourage thoughtfulness by the Department as the proposed amendments are considered, to ensure the BOSE powers can be effectively deployed to meet transparency needs while maintaining the efficacy of safety measures and the confidentiality of business information.

2.2 A holistic approach to reform

Microsoft welcomes the Department's announcement of a broader independent review of the OSA in early to mid-2024 and note that this broader review will necessitate further consideration of the BOSE regime. In light of this, it would be logical to review the BOSE Determination following the broader OSA review commencing in only a few months' time.

This would avoid duplicative and potentially parallel review processes of the same regime, especially during a time when many online service providers and civil society stakeholders are also contributing to other in-depth consultations, as well as implementing complex compliance processes under the eight Online Safety Codes and/or Standards. It would also ensure that the efficacy of the BOSE regime can be holistically examined in the context of the numerous other regulatory tools and transparency measures in Australia's online safety framework, including those set out in the industry codes and standards.

Additionally, the proposed amendments to the BOSE Determination are not time-sensitive. As repeatedly noted in the consultation paper, many of the new areas explored in the proposed amendments are already capable of being covered under the existing BOSE Determination, as the drafting is sufficiently broad and neutral to extend to a wide range of existing and emerging harms.

3.3 Unlawful or harmful material definitions

As with the existing BOSE Determination, the draft Determination frequently refers to 'unlawful or harmful material' when discussing content-based harms and the expectations on service providers to prevent and respond to such harms. In explanatory materials associated with the current Determination, and in eSafety's latest regulatory guidance, 'harmful' material or activity is that which may not be unlawful but:

- **is covered within the scope of the Act**, such as: cyberbullying material targeted at an Australian child, adult cyber abuse material, a non-consensual intimate image of a person (however, this is likely to also be unlawful under criminal laws), non-illegal class 1 material, material promoting abhorrent violent conduct, X18+ material or R18+ material; or

- **'should' fall under a provider's terms of use, policies, or standards of conduct**: this is described in the current explanatory memorandum to the BOSE as including material that is not necessarily unlawful or explicit referenced in the Act. Examples include hate against a person or group of people on the basis of a protected or vulnerable attribute; promotion of suicide and self-harm content, such as pro-anorexia content; volumetric attacks; and promotion of dangerous viral activities that have the potential to result in real injury or death. The boundaries of this category of materials are not clear.

The concept of 'harmful' has been described in the consultation paper as being useful to capture emerging and developing forms of harm, as well as material that is 'distressing.' Although this broad definition creates future flexibility, it results in considerable ambiguity for providers in interpreting the BOSE, as the Determination rests on a definition of an indeterminate category of material. This ambiguity creates significant uncertainty not only for service providers but also the regulator and the public.

As such, we recommend that 'harmful material' is limited to the material within the scope of the OSA. This also provides another reason to review the BOSE regime as a part of the upcoming OSA review. Doing so would not only improve consistency of definitions but could also allow a full consideration of whether online harms are being effectively addressed under the OSA and, if not, what amendments can be made to achieve the desired outcomes.

3.4 Expanding reach of the BOSE regime

Unlike the Online Safety Codes, which have a tiered risk approach, the BOSE Determination applies to all services falling within the broad categories of social media service (SMS), relevant electronic service (RES), and designated internet service (DIS), regardless of size, risk profile, or user thresholds. As such, the BOSE Determination applies broadly across the economy. To date, this breadth has been tempered by the enforcement decisions of the eSafety Commissioner, who has only issued enforcement notices to a small number of service providers, Microsoft included. However, the BOSE nonetheless exists as a standing set of expectations for a huge number of businesses globally. As the draft Determination significantly extends the BOSE regime, it risks amplifying existing regulatory uncertainty (e.g., to the meaning of 'harmful' material).

Further, there are aspects of the draft Determination, particularly those relating to generative AI, that may also adversely impact providers and customers of enterprise services by creating expectations or uncertainties that do not take into consideration size or risk profile of enterprise services, nor the well-established legal, contractual and technical frameworks that address the privacy needs of enterprises, relative to their service providers. As such, we urge the Government to consider any new regulatory expectations regarding generative AI services as a part of the Australian Government's wider work on Safe and Responsible AI, including through the forthcoming expert advisory group and on reforms to existing laws recently announced by the Department of Innovation, Science and Resources.

Finally, given the broad and expanding reach of the BOSE regime, it is important that the BOSE can accommodate the diversity of in-scope services and recognise that some of the requirements may not be appropriate for all relevant services. For example, the expectation that users will have granular control over the types of content they can receive is not relevant or practical in all settings, as there are many websites and apps captured under the definition of DIS where such controls would not be appropriate or practicable.

Specific commentary on the draft Determination

4.1 Generative AI capabilities (section 8A)

4.1.1 Express references to generative AI not necessary

As the consultation paper itself notes, the existing BOSE Determination already captures generative AI through its technology-neutral drafting. This is further evidenced by eSafety's current approach, which has included requests for transparency with respect to recommendation algorithms. In addition, the existing Internet Search Engine Services Code also already contains provisions with transparency and accountability obligations that apply expressly to materials generated by artificial intelligence. Similarly, the draft Industry Standards on which the eSafety Commissioner's Office is currently consulting also contain extensive additional obligations relating to generative AI.

In light of this, there does not appear to be a need for additional specific references in the BOSE to generative AI services. To the extent there are perceived gaps, we recommend these are considered holistically in the upcoming OSA review and the wider context of the Government's efforts to advance Safe and Responsible AI. If specific references are included, we offer the following comments.

4.1.2 Targeting the appropriate layer of the AI tech stack

Microsoft is committed to developing and deploying AI in a safe and responsible way. We recognise, however, that the guardrails for AI should not be left to technology companies alone. To support this work, we have offered our thoughts on a blueprint for the governance of AI, while recognising that every part of that will require discussion and deeper development.[1]

A key principle of our approach is the need to develop a regulatory architecture that maps to the technology architecture for AI. We believe that while the risks must be considered at different layers of

[1] Microsoft, "Governing AI: A Blueprint for the Future", 2023, available at Governing AI: A Blueprint for the Future".

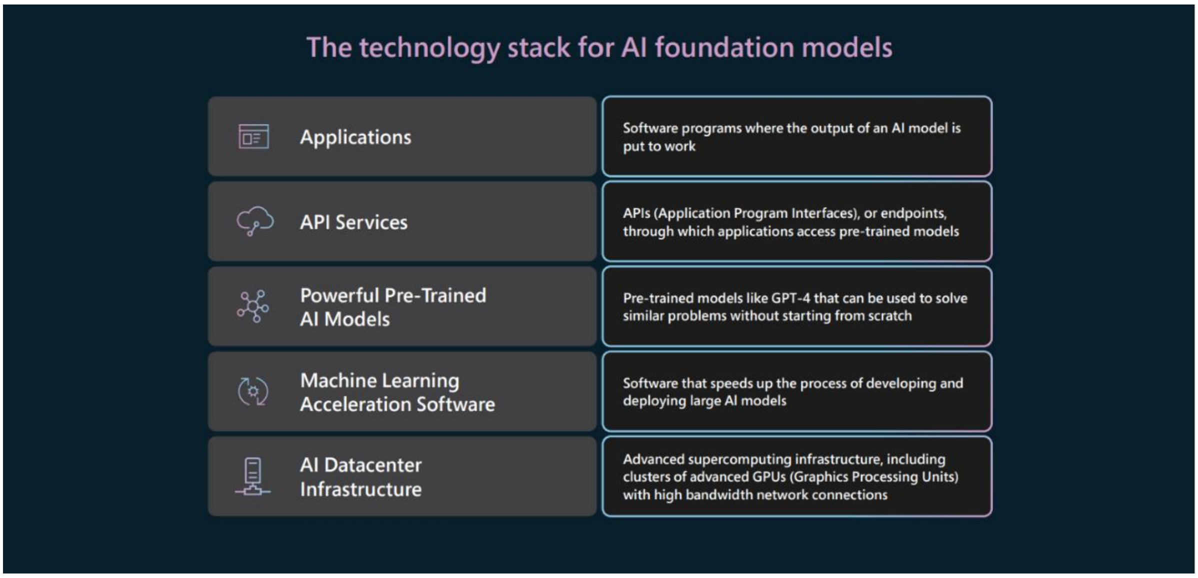
the AI technology "stack," their mitigations operate through interventions within those layers which, when taken together, best address online harms. A challenge for such a framework is that, while AI safety is a set of cumulative efforts, those efforts are often operationalised by different actors, who may separately be responsible for individual parts of the stack and license these to others. And while cumulative, the nature of AI safety interventions will differ depending on the layer of the stack. We provide below at Figure 1 below an illustration of the AI technology stack.

As such, specific obligations with respect to AI safety systems across multiple layers of the stack may not be capable of being effectuated by a single layer's owner. Each layer's safety works in conjunction with the others to deliver AI safety to the end user at the application layer.

Similarly, the safety interventions at each layer of the stack should be calibrated to address the nature of risk at that layer. Online harms of the nature regulated under the OSA and the BOSE Determination are best addressed by interventions at the application layer, where the protection of end users is most capable of being implemented. It also enables the intended use of the service by end users to be understood in context.

We therefore recommend clarifying language is added to any additional expectations relating to generative AI services to make clear that they apply only to the application layer of SMS, RES and DIS services. We further recommend shifting from the language of "capabilities" to focus on "services" and "features", which will enable more precise understanding of scope and greater regulatory certainty for services providers.

Figure 1: The technology stack for AI foundation models



The technology stack for AI foundation models

Applications	Software programs where the output of an AI model is put to work
API Services	APIs (Application Program Interfaces), or endpoints, through which applications access pre-trained models
Powerful Pre-Trained AI Models	Pre-trained models like GPT-4 that can be used to solve similar problems without starting from scratch
Machine Learning Acceleration Software	Software that speeds up the process of developing and deploying large AI models
AI Datacenter Infrastructure	Advanced supercomputing infrastructure, including clusters of advanced GPUs (Graphics Processing Units) with high bandwidth network connections

4.1.3 Training materials

We also recommend considering AI in the context of the wider review and Government efforts to support additional expert evaluation of measures appropriate at each layer. For example, in the draft Determination, an example of a reasonable step that providers could take is to ensure that 'training materials for generative AI capabilities and models do not contain unlawful or harmful material'.[2] This measure extends beyond the application layer and present several challenges:

1. First, although it is reasonable to expect providers to take steps to ensure that generative AI models are not trained on datasets including known CSAM, it is not practicable to expect

[2] At Section 8A(3)(c).

providers to exclude all material that may fall within the ambiguous category of 'unlawful or harmful' material. Online harms are often dependent on context and conduct. Therefore, it is not always possible to objectively assess whether some types of material can be considered harmful in isolation. For instance, certain visual media may be harmful to those with photosensitive epilepsy but will not be considered harmful by others. Or an image of a wounded animal may be of professional interest to a veterinary professional, yet acutely upsetting and harmful to a child.

A similar logic also applies for categories of unlawful material, given that most such material is not illegal in all circumstances. For instance, while certain material relating to acts of terror may be illegal to publish in some circumstances, it can be legal to share in other contexts (such as for academic, artistic or legal purposes). For example, public debates reflect different perspectives on what footage and commentary from the current conflict in the Middle East should be considered pro-terror or journalistic reporting.

2. Second, limiting training data to exclude potentially "harmful" content is likely to limit the utility of large language models. The benefits of these models arise from the very breadth and volume of material on which they have been trained. An expectation that the very broad category of potentially harmful content is excluded from training datasets is both unworkable and undesirable. AI models also require training on diverse material in order to understand the types of material to avoid generating, and the types of prompts, including prompt attachments, to ignore or handle differently.

   Training data sanitised of any material that could be construed as harmful in any context would also undermine the efficacy of safety mitigations, reducing the likelihood for them to operate with precision and nuance. One of the most promising opportunities for AI is to provide sophisticated technological tools to advance online safety and support content moderation. Without training data of sufficient diversity and scope to enable such nuanced assessment, we risk losing the benefits of this innovation. That is, given the breadth of material potentially captured in the scope of 'harmful' material, excluding such material from training data could have the inadvertent effect of making generative AI less capable of identifying and preventing the creation of more high-risk material.

These challenges above illustrate the inherent limitations, arising from lack of context, in establishing expectations at the model layer of the AI stack, where end-users do not interact with the technology. As a result, we underscore our recommendation that the Department consider safety requirements for AI models through some of the wider processes currently underway in Australia on Safe and Responsible AI. This will ensure that model requirements can be considered holistically, supporting safety, equity, security and innovation outcomes.

4.2 Recommender systems (proposed section 8B)

Neither the current nor draft Determination includes a definition for 'recommender systems'. However, the consultation paper does refer to an eSafety Tech Trends Position Paper on recommender systems, which defines them as follows:

> *Recommender systems, also known as content curation systems, are the systems that prioritise content or make personalised content suggestions to users of online services. A key component of a system is its recommender algorithm, the set of computing instructions that determines what a user will be served based on many factors. This is done by applying machine learning techniques to the data held by online services, to identify user attributes and patterns and make recommendations to achieve particular goals.*

The focus of that Paper is primarily on recommender systems used in SMS and other services leveraging recommendations of user-generated content. The lack of a definition in the draft Determination could widen the interpretation of their meaning to include a broader range of "algorithms", used across a hugely diverse variety of services. We therefore recommend that the term is clearly defined to clarify the intended application of the draft Determination.

This uncertainty and complexity also risks being compounded by the uncertainty and breadth of 'unlawful or harmful material', as discussed above. In particular, it is not clear how the expectation on diverse service providers to deprioritise recommendations of contextually harmful content would work in practice.

4.3 Appropriate age assurance mechanisms

Microsoft takes seriously our responsibility to keep users safe on our services and agrees with the importance of protecting younger users from age-inappropriate content. We further note that, in late 2023, the Minister for Communications noted in her response to eSafety's Age Verification Roadmap that "technological developments in this space are still new and evolving." As a result, the Minister advised that the question of appropriate age assurance mechanisms should be dealt with through the Class 2 Industry Codes, which will be developed later this year.

As such, we recommend deferring this amendment until after the Class 2 Industry Codes process has enabled a robust exploration of suitable age assurance interventions, including balancing the overlapping considerations of child rights, online safety, privacy, freedom of expression, and cybersecurity.

4.4 Supporting investment through appropriate confidentiality safeguards

The draft Determination also refers to reasonable steps regarding resourcing and investment.[3] We recommend adding clarifying language to this requirement to acknowledge and encourage the many and varied types of investments that providers make in online safety, which may include participation in research, pilot projects, and collaborations.

This example also illustrates the need to balance the benefits of public transparency with the need to support confidential business information. Certain categories of financial or business investment information may not be appropriate for public reporting under the BOSE regime. It may also lead to inaccurate or confusing comparisons between companies that operate in very different ways. As such, we urge the Department to consider implementing some effective boundaries on how confidential business information can be requested and published by the regulator. The Department may wish to look to existing guardrails for other Australian regulators' powers to request, disclose, and publish confidential business information.[4]

[3] At sections 6(3)(f) and (h).

[4] E.g. ASIC Regulatory Guide 103, Confidentiality and release of information; ACCC Guidelines on section 95ZK claims in price inquiries, etc.

4.5 Hate speech

4.5.1 Lack of coordinated approach to addressing hate speech

The proposed paragraph 6(3)(i) creates a new expectation for providers to detect and address hate speech, which is defined by new subsection 6(4). As outlined above, we recommend limiting the application of the BOSE to categories of harm that Parliament has agreed should be addressed through the OSA. We further note that the Minister has expressed that a key objective of the upcoming OSA review will be to ensure the Act is amended to cover hate speech, and therefore recommend against introducing this amendment shortly before a wider review of hate speech regulation has occurred.

Additionally, all Australian jurisdictions address hate speech to varying degrees through measures built around concepts of anti-discrimination, vilification, and incitement, implemented through a variety of standalone statutes, as well as through criminal and telecommunications laws. Considering this through the legislative review will enable this wider context to be factored in, as well as how measures to address this harm may vary across services.

Finally, the framing of the proposed amendment is to "**detect** and address" hate speech (emphasis added). Given the breadth of the services in scope for the BOSE, including private communications services, such an expectation could disproportionately impact other rights including privacy and free expression. If the draft Determination is amended to include hate speech, we suggest it does not extend beyond an expectation to "address" hate speech, reflecting the need to assess potential hate speech in context and to adapt safety interventions to the nature of the service.

4.6 Transparency

4.6.1 Additional annual reporting requirement (proposed section 18A)

We urge the Department to reconsider the proposed addition of at least annual reporting requirements under the draft Determination. This amendment significantly overlaps with the existing reporting requirements under the OSA, as well as other legislation. For example, many services providers already have reporting obligations under the eight Class 1 Industry Codes and Standards and obligations to respond to periodic and non-period notices under the BOSE, along with other additional reporting requirements such as those under the misinformation and disinformation code and telecommunications regulations.

Accordingly, we recommend that any additional transparency reporting requirements should be revised to avoid duplication with the existing reporting processes and obligations under the OSA.

4.7 Enforcing terms of use

Proposed subsections 14(1A) and (2) include expectations that services will take proactive steps to detect breaches of terms of use, policies and procedures. However, technology-driven proactive detection is not a silver bullet: it requires nuance and benefits from human reviewers being included in the review loop to ensure fair and equitable outcomes. Leveraging such tools may also not be appropriate for all harm types, on all services, given the need to balance competing human rights to privacy, self-determination, and freedom of expression.

We therefore recommend clarifying this proposed expectation to require in-scope services to prioritise and tailor detection strategies depending on the potential impact of violative material and the type of service in question. While two of the proposals are caveated with 'reasonable steps', it may be helpful to include additional qualifiers to assist with providers' assessment of whether proactive detection is necessary and appropriate, including a consideration of the risk profile of the service and the likely impact of proactive detection on other competing rights.

Conclusion

Microsoft thanks the Department for the opportunity to provide our feedback on the consultation paper and proposed amendments to the BOSE Determination. Overall, we would like to reiterate our view that this review of the BOSE regime may be better considered within the broader review of the OSA to ensure better alignment and coherence across this important regulatory regime and to align with other cross-government policy processes.

We would welcome the opportunity to discuss any of our feedback further with the Department and look forward to continuing to engage in this reform process.