Microsoft Response to the Public Consultation on Australia's Draft Online Safety Industry Standards (Class 1A and 1B)

Microsoft welcomes the opportunity to provide input to the eSafety Commissioner's public consultation on the draft Designated Internet Services Standard ("DIS Standard) and Relevant Electronic Services Standard ("RES Standard") for class 1A and class 1B material (collectively, "Industry Standards"). We recognise that regulatory measures have an important role to play in creating safer online services and ecosystems and welcome the opportunity to ensure that the proposed Industry Standards will support practical, proportionate, and effective protections for Australians, alongside the extant Industry Codes.

In this submission, we provide comment on three thematic issues. We are available to discuss these, as well as the detail of the proposed obligations with the eSafety Commissioner and team, as required. This submission addresses:

1. Maintaining a risk-based and rights-protective approach to addressing class 1A and 1B content.
2. Supporting compliance and a proportionate approach to enforcement.
3. Effectively tailoring obligations for services leveraging generative AI.

Maintaining a risk-based and rights-protective approach to addressing class 1A and 1B 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.

While drafting the initial Industry Codes for Relevant Electronic Services (RES) and Designated Internet Services (DIS), industry associations proposed a number of sub-categories of RES and DIS services, in order to help break down what are otherwise extremely broad categories. We appreciate that the draft Industry Standards have maintained many of the additional definitions proposed in those drafts, to help ensure that the obligations are appropriately tailored to the services in question.

The criticality of a risk-based approach has always been at the centre of the Industry Codes and Standards drafting process: eSafety's 2021 Position paper on "Development of industry codes under the Online Safety Act" outlined its position that the Codes would "adopt an outcomes- and risk-based regulatory approach"[1]; aiming to avoid applying a one-size-fits-all model. In keeping with this, the Industry Standards require in-scope services to complete a risk assessment and to ensure that risk assessments are carried out when material changes are proposed to a service. The Standards also appropriately limit this requirement, for example, by excluding its application to enterprise RES and DIS.

[1] eSafety Commissioner, "Development of industry codes under the Online Safety Act: Position Paper", Sep 2021, page 46.

We welcome the clarity on the matters to be addressed in a risk assessment, as well as the flexibility to leverage guidance such as the Digital Trust and Safety Partnership's Safe Framework. Efforts are currently underway through international standards processes to encode some of these requirements in a way that supports global harmonisation and interoperability – we recommend the Commissioner consider including in the "matters to be taken into account" the extent to which a risk assessment aligns to a recognised international standard (e.g., from the International Standards Organization) or whether a service's risk assessment has met the requirements of another international safety regime (e.g., a potential mutual recognition arrangement among members of the Global Online Safety Regulators Network). This will enable the further development of a global approach, supporting effective and consistent safety information-gathering across jurisdictions.

We are concerned, however, that the draft Industry Standards' approach to risk assessment deviates from a risk-proportional approach in some ways. We recommend additional focus on analysing the actual risks of specific harms on each individual service and, importantly, the efficacy of mitigations already in place to address those risks, before attaching Tier 1 obligations. This tailored focus could help drive meaningful review and improvement of mitigation measures, as well as to assess potential human rights impacts.

The breadth of the proposed approach is also demonstrated by the broad range of risks to be addressed in the draft Industry Standards, including risk that Australians will "solicit, generate, distribute, get access to or be exposed to class 1A material or class 1B material through the services". This is broader than the objectives eSafety established in its 2021 Position Paper to guide the Codes, capturing a much wider range of potential content and conduct that may vary across services. Enabling services to take more tailored and risk-proportional approaches to assessment would in turn help to ensure individual services are effectively tailoring their safety interventions to address identified challenges.

To support that risk-based and rights-respecting approach, we also recommend including additional considerations to the guidance provided on what constitutes "appropriate action" in section 12 of both Industry Standards. As written, the draft Standards enable an assessment of risk-proportionality to be undertaken only in response to a decision with respect to specific material (section 12(b)(3)). We suggest amending the Industry Standards to clarify that when deciding whether an action is suitable for meeting a Standard's requirements, the following factors should also be considered: (1) how much the action in response to the industry standard matches the evaluated risks of class 1A and 1B material on that service; and (2) the significance of upholding fundamental human rights online, such as the right to freedom of expression and the right to avoid arbitrary or unlawful interference with privacy.

We do have concerns, for example, that without such guardrails, there is a risk of over-removal of certain content, with implications for the human rights of Australians. For example, the requirement that end-user managed hosting services use technologies to detect and remove pro-terror material may risk being disproportionate to the harm, which may rest (for example) on pro-terror material being shared, glorified, or being used to promote a terrorist ideology. Material may be held in cloud storage for a range of reasons, including research or journalism, with no intent to cause harm. For example, analysis of pro-terror messages for themes and strategies can inform the design of pro-social interventions to counteract radicalisation. Unlike child sexual abuse material, not all pro-terror material will be harmful per se – often, intent and/or contextual determinations are required as part of a moderation process.

Supporting compliance and a proportionate approach to enforcement.

As noted above, we have welcomed eSafey's emphasis on establishing a risk- and outcomes-based approach. The Discussion Paper accompanying the draft Industry Standards reiterates this, stating that the requirements are intended to provide flexibility in how they are achieved. Embracing flexibility of this kind can support an informed conversation between the regulator and in-scope services on the ways in which tailored safety interventions can address identified safety risks – we look forward to continuing to engage constructive dialogue with the eSafety Commissioner and her team, including as we evolve our approach to meet new threats and new regulatory requirements.

Recognising the need for regulated services to comply with a variety of Australian legislative regimes, we recommend incorporating into the draft Industry Standards the content included in the Head Terms to the Industry Codes, including the "Limitations and lawful conduct" provisions in Section 6. Inclusion of the Head Terms in their entirety to the draft Industry Standards could clarify the application of the secondary regulation created under the Online Safety Act 2021 – including for users, who may struggle to understand the protections in the coming dual Codes/Standards regime.

Across the draft Industry Standards, we also recommend providing additional guidance to help clarify the way in which the eSafety Commissioner will engage with in-scope services and providers on compliance. For instance, we recommend building into the draft Standards objective considerations for how the Commissioner determines the effectiveness of safety measures. For example, section 22 of the draft RES Standard requires certain services to implement systems, processes, and technologies that "effectively" deter and disrupt end-users of a service from creating, offering, soliciting, accessing, distributing, or otherwise making available or storing child sexual abuse material or pro-terror material. Given the potential breadth of in-scope content and conduct, especially some of the uncertainties around what it might mean, for example, to create, offer or solicit pro-terror content (with its contextual nature) and the challenges of measuring deterrence, there is considerable uncertainty around how to judge effective compliance. Adding objective criteria on which "effectiveness" may be measured would provide needed clarity for regulated services. Equally, the new requirements in the draft Industry Standards for "development programs" would benefit from objective criteria for measuring "investment and development activities" and "support."

Finally, we have some continuing concerns about the proportionality of measures that rely upon actions of a third-party non-governmental organisation, given these may result in non-compliance that a service is unable to remedy. For example, it is worth clarifying that the Global Internet Forum to Counter Terrorism (GIFCT) does not maintain a database of pro-terror material that it has "verified": it provides a central repository for content that has been hashed by a participating member company on the basis that it violates relevant company policies and falls within the GIFCT taxonomy.

Effectively tailoring obligations for services leveraging generative AI.

Over the last year, advances in artificial intelligence (AI) have transformed the technology landscape, creating significant opportunities but also highlighting the need for governance frameworks and guardrails to support responsible innovation. 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.[2]

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 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, those interventions will differ, depending on the layer of the stack. We provide below at figure 1 an illustration of the AI technology stack.

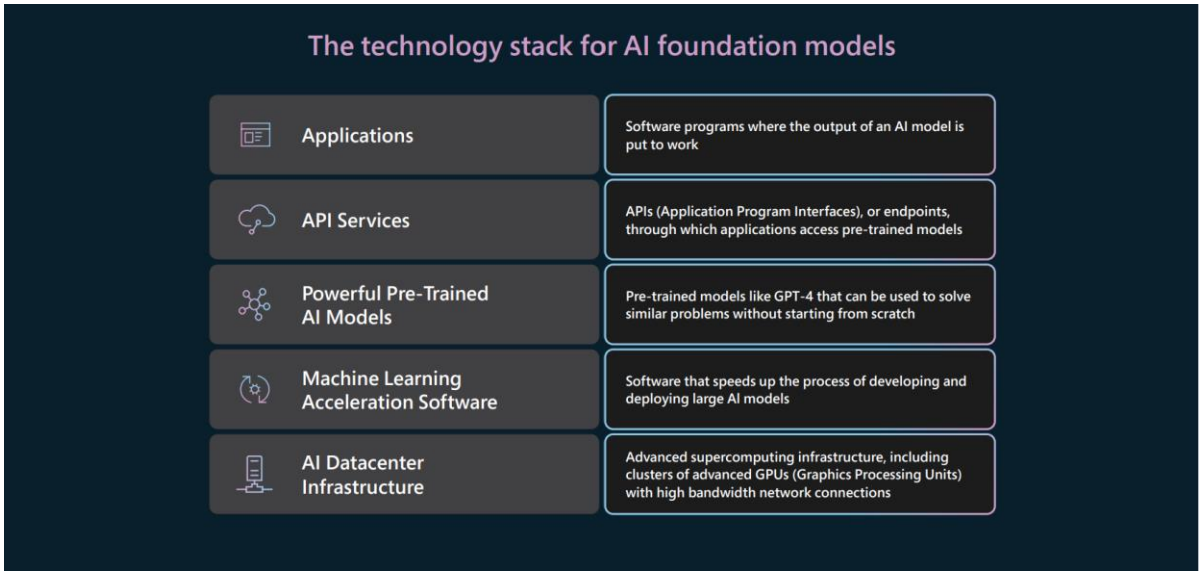

Figure 1: The technology stack for AI foundation models

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. For example, the AI safety mitigations that should be implemented by a datacentre infrastructure provider, are dramatically different from safety mitigations that should be implemented by an application provider providing a service that interacts with a large number of end users. As another example, a foundation model provider offering pre-trained AI models, on which others can build applications, should test the AI model when building it and share information about the capabilities and limitations of that model. Microsoft provides Transparency Notes for its AI services.[3] An organisation building an application on those pre-trained models should then use this information to build an application that performs safely, securely and responsibly. That organisation should also determine what additional measures are required so that the application performs appropriately for its intended use case, with greater safeguards required for higher-risk use cases. For instance, if the intended use of an

[2] Microsoft, "Governing AI: A Blueprint for the Future", 2023, available at Governing AI: A Blueprint for the Future (microsoft.com).
[3] For Azure OpenAI, for example: Transparency Note for Azure OpenAI.

AI application was to detect hate speech so that such speech could be moderated on a social media service, imposing a safety intervention within a foundation model that blocks hate speech would frustrate this purpose.

We support the extension of content and conduct safety laws and regulations at the application layer, where the protection of end users is most capable of being effectuated and the intended use of the service by end users is understood in context. We therefore support the inclusion of AI applications in the Industry Standards, as it reflects the objective of the Online Safety Act to address the safety of technology's end users. This approach has already been implemented through the Industry Codes process, where associations worked with the Commissioner to update the draft Code for search engine services to appropriately reflect additional mitigations that may be needed where AI features are integrated into an existing service at the application layer. However, the treatment of foundation models within the draft Industry Standards is misaligned with this objective and may actually reduce the effectiveness of AI safety systems. Dedicated responsible AI governance models and best practices addressing foundation models are separately emerging that will also support application developers to meet their regulatory requirements, enabling all parties to effectively meet their shared responsibilities. Microsoft looks forward to continuing to engage with the Australian government's wider work on Safe and Responsible AI, including through the forthcoming expert advisory group and on reforms to existing laws.

Supporting definitional clarity

At the outset, we recommend changes to help support clarity in the scope and definitions in the draft DIS Standard. The draft DIS Standard introduces a new category of a "high impact DIS", which uses machine-learning models to enable an end-user to produce material and for which it is reasonably foreseeable that the service could be used to generate synthetic high impact material. We understand from discussions between industry and the eSafety team that this category is not intended to capture generative AI "chat" services that have in place Responsible AI safety systems to prevent the generation of harmful content.

However, this is not clear from the face of the draft DIS standard and may benefit from further qualification. The draft DIS Standard may also benefit from a new section clarifying that the obligations therein are limited to services or features delivering material to end users at the application layer, not the model or API layers. This aligns with the definition of "designated internet service" at section 14 of the Online Safety Act 2021, which focuses on services allowing end-users to access material or services delivering material to persons having equipment appropriate to receiving that material by means of an internet carriage service.

Distinguishing AI services

We support the provisions in the draft DIS Standard requiring risk assessments but note that the content and format of such an assessment may differ depending on whether they relate to an AI feature, whether that AI feature is incorporated into another category of designated internet service, or whether the primary focus of the service is the provision of AI functionality. In the latter case, we recommend the draft DIS Standard alternatively allow an AI feature or service to provide its impact assessment as a means through which to assess and enable the mitigation of potential AI risks. Throughout the draft DIS Standard, we also recommend considering the nature of such services, which do not host content in an ongoing way in the same way as a website or service focused on enabling the "posting" or sharing

of user-generated content. This means that, for example, the requirement at section 17 to "remove" child sexual exploitation material or pro-terror material (also at section 22) may not make sense in the context of AI features or AI applications. It also may create challenges in effectively implementing the requirements at section 18 with respect to class 1B material. As has previously been discussed with industry, class 1B material is also likely to be highly contextual – particularly given the need to consider specific material in the context of the Australian classification framework. For the provider of a high impact generative DIS, it may be highly challenging if not impossible to make such an assessment for each individual user prompt and output, including to determine whether a user has intentionally sought to violate the applicable terms of service.

Reflecting the technology stack and the roles of developers versus deployers

As outlined above, we support a regulatory architecture that takes into account the unique roles played by both the developers and deployers of AI foundation models and applications. But as it pertains to online content harms, as with other established regulatory domains, we see it as critical to distinguish between these different roles. Equally, as with digital safety regulation, requirements should focus on advancing responsible AI outcomes for high-risk systems.

While we firmly support the Standards' intent to address the potential risks of AI-generated child sexual abuse or exploitation material, we are concerned that the specific requirements in section 23 risk blurring distinct roles and responsibilities, especially given the breadth of the requirements to not only prevent generation but to also effectively deter and disrupt a wide range of potential conduct. First, we are concerned that the provider of an enterprise DIS (whether an AI service such as a developer of foundation models for third party enterprises' use in its own applications, an enterprise service with AI features, or otherwise) will have limited technical and legal abilities to deter or disrupt the potential activities of an application's end users. As discussed in the development of the Industry Codes, this reflects the lack of direct relationship between enterprise service providers and end users. In addition to the lack of privity of contract, enterprise service providers act as data processors, not data controllers. Meeting the requirements of section 23(2) would require an enterprise service provider to have significant levels of access to customer data, which most enterprises – and in particular government entities – are unlikely or even legally unable to agree to.

Secondly, the minimum requirements for high impact generative AI DIS incorporate activities for both developers and deployers. We recommend limiting this list to activities for deployers, keeping the scope of the draft DIS Standard to the application layer. This would enable additional obligations for AI to be appropriately developed by the Australian Government in due course, while also ensuring meaningful protections are available in the short-term for Australians. It would also avoid a situation in which the provider of an application service has responsibility for activities that may sit with a separate actor. eSafety may therefore wish to remove or modify the requirements at section 23(3)(b), (c) and (g).

Preserving the ability to use AI for safety

Finally, we note that some of the proposed requirements in the draft DIS Standard may have the unintended consequences of limiting the effectiveness of AI safety systems and preventing the development of innovative AI safety solutions. To ensure that AI models and safety systems (such as classifiers) can be trained to detect and flag such content requires that the AI is exposed to such content and evaluation processes are put in place to measure and mitigate risks. Entirely "clean" training data may reduce the effectiveness of such tools and reduce the likelihood they operate with precision and

nuance. One of the most promising elements of AI tooling for content moderation is advanced AI's ability to assess context – without training data that supports such nuanced assessment, we risk losing the benefits of such innovation.

Conclusion

We appreciate the opportunity to provide comment on the draft Industry Standards and would welcome the opportunity to discuss the contents of this submission, as needed.