Amending the Online Safety (Basic Online Safety Expectations) Determination 2022

Google Submission

26 February 2024

Executive Summary

Google welcomes the opportunity to contribute to the Government's consideration of amendments to the *Online Safety (Basic Online Safety Expectation) Determination* (**BOSE Determination**).

Protecting people from harmful and illegal content and making our products safer for everyone is core to the work of many different teams across Google and YouTube. When it comes to the information and content on our platforms, we take our responsibility to safeguard the people and businesses using our products seriously, and to do so with clear, transparent policies and processes. We also support reasonable and consistent regulation and liability protections that allow us to effectively combat harmful content while protecting the core benefits of online environments, including the exercise of the right of free expression, ability to find a wide range of viewpoints, access useful information and connect with one another.

However, we are unclear the reason for amending the BOSE Determination prior to the completion of the statutory review of the *Online Safety Act 2021* (**OSA Review**). We are also concerned with what is becoming an increasingly complicated, duplicative and confusing regulatory regime for online safety within Australia. And we suggest that amendment of the BOSE Determination prior to the conclusion of the OSA Review only risks further exacerbating these concerns.

The OSA Review (which is scheduled to be completed in October 2024), will include consideration of not only the effectiveness of the BOSE Determination, but also many of the subject matters that the amendments are seeking to address, including, consideration of whether additional protections are warranted in respect of "online hate", specific types of technology such as "recommender systems" and "generative AI" and in ensuring that industry acts in the "best interests of the child". The OSA Review is also occurring in the context of other broader government initiatives which overlap many of the proposed amendments, including a review into the responsible and safe use of AI, a review of the *Privacy Act 1988* (which includes consideration of a Children's Online Privacy Code modelled on the UK Age Appropriate Design Code) and a review of the National Classification Scheme.

Each of the above processes has the potential to supersede any proposed amendments, and more concerningly, result in a confusing, duplicative and inconsistent approach to the regulation of online safety more broadly across all service providers. We strongly recommend that the Government defer consideration of any amendments to the BOSE Determination until after the outcome of the OSA Review. We also urge the Government to ensure the OSA Review properly takes into account a broader range of government initiatives in this area.

Noting our broader concerns with the timing of any amendments ahead of completion of the OSA Review, we otherwise make the following specific recommendations in respect of the proposed amendments:

- **Definition of "unlawful or harmful" material or activity**. We recommend that the Government clearly and consistently define what is meant by "unlawful or harmful" material or activity within the context of the BOSE Determination. It is imperative that service providers know what the obligation or expectation is that they have to meet, and which regulatory regime may apply. It is also difficult to provide feedback as to the reasonableness or feasibility of many of the proposed amendments without understanding what unlawful or harmful material it is seeking to address.

- **Generative AI and Recommender Systems**. We recommend that the proposed expectations directed only at the use of "generative AI" and "recommender systems" by a service should not be included (as the potential harms already addressed as part of existing section 6) or, alternatively, should be deferred until after the OSA Review, and developed in consultation with broader government policy - in particular, in the context of generative AI, the Government's broader approach to the regulation of AI.

- **Hate Speech.** While we support the Government's objective to tackle hate speech, we strongly urge the Government to address the harm caused by hate speech holistically in a democratic and transparent manner, and through broader regulation of hate speech, both on and offline.

- **Best Interest of the Child**. While we agree that service providers should take into account the "best interests of the child" when designing and implementing services likely to be accessed by children, we recommend that this is best addressed as part of the OSA Review. We also recommend that it is undertaken in consultation with other overlapping policy initiatives, including development of the Children's Online Privacy Code to ensure a coherent and non-fragmented approach is adopted.

- **Transparency**. We recommend against the inclusion of additional transparency or reporting requirements. The proposed transparency requirements introduce significant regulatory compliance burden on industry and are unnecessary to achieve the objective of improving transparency and accountability of service providers in light of existing reporting obligations under the Industry Codes, draft Industry Standards and powers of the Commissioner to request the information (on a per service and as needed basis) under the Online Safety Act.

- **Enforcement of terms of service and conditions**. We recommend that proposed subsection 14(1A) is better focused on requiring service providers to take appropriate and proportionate action to enforce its terms of service, policies and procedures and not on requiring service providers to take steps to "proactively detect" breaches of its terms of service and conditions. A service provider's terms of use, policies and procedures or standards of conduct cover a large number of different matters that pose varying levels of harm and risk to other end-users. Proactive detection of all breaches may neither be feasible, appropriate or proportionate for every type of potential breach and for every service.

- **Timely resolution of complaints and reports.** We recommend that proposed subsection 14(3) is limited only to an obligation to review and take appropriate action in response to a report of harmful material within a reasonable period of time. At the scale that many service providers operate, a requirement that service providers provide feedback in response in every user report or complaint would be extremely resource intensive and not practical.

General comments

I. Timing of the amendments

The Government should defer any consideration of changes to the BOSE Determination until after the statutory review of the *Online Safety Act 2021* (OSA Review), to ensure that a coherent and consistent approach to the regulation of online safety is achieved.

There are a number of reviews and consultations either in progress or that are due to commence this year that overlap with, and are likely to impact on how the BOSE Determination and the Online Safety Act operates.

This includes:

- **The ongoing consultation and subsequent implementation of draft Relevant Electronic Services (RES) and Designated Internet Services (DIS) Industry Standards for "class 1 material".** The Industry Standards include new and additional obligations relating to ongoing investment in the development of new technologies and resourcing of trust and safety, new categories of generative AI services, and reporting obligations.
- **Commencement of the development of Industry Codes for "class 2 material".**
- **The statutory independent review of the *Online Safety Act 2021* .** This will include a review of the following matters that overlap with one or more of the proposed amendments:

    a. The operation and effectiveness of the BOSE Determination regime under the Online Safety Act.
    b. Whether additional arrangements are warranted to address online harms not explicitly captured under the existing statutory schemes including (amongst others) online hate and potential harm caused by a range of emerging technologies such as generative AI and recommender systems.
    c. Whether the regulatory arrangements, tools and powers available to the Commissioner should be amended and/or simplified including thorough consideration of ensuring industry acts in the "best interests of the child".

The above is in addition to other, broader government initiatives, including:

- **The proposed review of the *National Classification Scheme* in the first quarter 2024.** The National Classification Scheme underpins the definition of class 1 and class 2 material.
- **The release of the Government Interim Report into *Safe and Responsible AI in Australia* ("AI Review")**. The interim response by the Government highlights the intent of the Government to regulate AI in "high risk" settings, in a risk based and proportionate approach, which will involve further consultations to determine whether mandatory regulation will be via amendments to existing laws or via an alternative approach. This approach will likely overlap with the proposed new amendments in the BOSE Determination to address "generative AI".
- **Reforms to the *Privacy Act 1988* (the "Privacy Act")**. As part of those reforms, the Government has committed to the development of a Children's Online Privacy Code, which will apply to online services likely to be accessed by children, and to the extent possible, align with the *UK Age Appropriate Design Code* (**UK AADC**). This is likely to overlap with the introduction of a new "best interests of the child" obligation in the BOSE Determination.
- **The *Communications Legislation Amendment (Combating Misinformation and Disinformation) Bill 2023* (the "Misinformation Bill").** The current draft of the bill defines "harm" as including "hatred against a group in Australian society on the basis of ethnicity, nationality, race, gender, sexual orientation, age,

religion, or physical or mental disability". This will likely overlap with the proposed concept of "hate speech" under the BOSE Determination.

Each of these processes has the potential to supersede one or more of the proposed amendments made to the BOSE Determination, and more concerningly, result in a confusing, duplicative and inconsistent approach to the regulation of online safety across a number of different regulatory areas. It is also difficult to provide meaningful feedback to the draft amendments while many of the above processes are still ongoing. We consider that amendments to the BOSE Determination ahead of the completion of these above processes is premature.

We recommend that the Government defer any consideration of changes to the BOSE Determination until after the outcome of the OSA Review. This will ensure that any amendments are consistent with the Online Safety Act and support a more coherent approach to the regulation of online safety in Australia. Separately, the OSA Review should properly take account of relevant broader government policy initiatives and objectives, including the regulation of AI, privacy, misinformation/disinformation and scams, to ensure consistency and minimise regulatory duplication and overlap.

ii. Clarification as to intended purpose and operation of the BOSE Determination in relation to the Industry Codes/Standards

The Government should consider providing greater clarity as to the intended purpose and operation of the BOSE Determination in relation to the Industry Codes and/or Industry Standards for class 1 and class 2 material. This is necessary to ensure that regulation of online safety is clear, proportionate, consistent and does not impose an unnecessary compliance burden on service providers.

We support regulation to better protect and empower people online. However, we are concerned with what is becoming an increasingly complicated, overlapping and confusing regulatory regime within Australia.

Companies, such as Google, who provide multiple services, are currently subject to three separate regulatory regimes for addressing class 1 material on their services.

If a service meets the definition of a Hosting Service, Equipment, Social Media Service, App Distribution Services and/or Search Engine Service then they are subject to Industry Codes. Whereas, if the service meets the definition of a Designated Internet Service and Relevant Electronic Services they will be subject to Industry Standards. The obligations as between the Industry Codes and the draft Industry Standards are not aligned, and in many instances, the differences do not appear associated with any particular or increased risk with the service. We have made separate submissions to the eSafety Commission highlighting our concerns with the present approach. And for those services that meet the definition of Social Media Services, Relevant Electronic Service and Designated Internet Services, they are also expected to take reasonable steps to meet expectations articulated in the BOSE Determination for the same class 1 material. Again, the expectations overlap but are still inconsistent with equivalent obligations under both the Industry Codes and Standards. The complexity will be further exacerbated by the introduction of an additional set of Industry Codes and/or Standards for class 2 material, which we understand Government intends to progress in advance of the finalisation of the OSA Review and the review of the National Classification Scheme.

The practical impact of the above is that:

- **Service providers cannot adopt uniform compliance solutions across all products or services where it makes sense to do so (because the Industry Standards, Codes and BOSE Determination do not align).** This increases regulatory burden without improving online safety.

- **Service providers can be compliant with the mandatory requirements under an Industry Code or a Standard, but still non-compliant under the BOSE Determination for the same type of obligation or material.** Conceptually it is difficult to understand how meeting a mandatory requirement that the Commissioner is satisfied is an "appropriate community safeguard", could at the same time be determined by the Commissioner as insufficient to meet a similar requirement to take "reasonable steps" under the BOSE Determination.
- **Service providers are subject to duplicative reporting and transparency requirements.** While the Industry Codes and draft Industry Standards impose certain compliance reporting requirements, the Commissioner can separately issue a notice to answer additional questions about the same material and/or obligation under the BOSE Determination.

At the time the BOSE Determination was introduced, the process for developing the Industry Codes and Industry Standards had not yet commenced. Now that the the Industry Codes for class 1 material have been completed, and the eSafety Commission is in the process of finalising the Industry Standards for class 1 material, we believe it is a good opportunity for Government to review the intended purpose of the BOSE Determination, how it interacts with the Industry Codes and Industry Standards, and how best to streamline the different conflicting obligations to reduce regulatory and compliance burden on service providers, while still achieving the intended purpose and objectives of the Online Safety Act.

The Government has committed to reviewing the effectiveness and the intended purpose of the BOSE Determination as part of the OSA Review. We support this review to ensure the BOSE Determination operates effectively and is implemented in a way that is consistent with Parliament's intention. **We recommend that** the Government use the OSA Review to provide clarity on the relationship and interaction between the BOSE Determination and the Industry Codes and Industry Standards, and that any changes to the BOSE Determination be deferred until after this process and review has been completed.

iii. Clarification as to the scope of "unlawful or harmful" material or activity

The Government should define the intended scope of "unlawful or harmful" material or activity under the BOSE Determination, and how that interacts with, or relates, to other regulatory regimes and initiatives.

The existing statutory regime under the Online Safety Act explicitly identifies and defines six categories of unlawful or harmful material. These are:

- Cyber-bullying material targeted at an Australian child;
- Cyber-abuse material targeted at an Australian adult;
- Non-consensual sharing of intimate images of a person (image-based abuse);
- Class 1 material under the Online Content Scheme;
- Class 2 material under the Online Content Scheme (preventing access to children); and
- Material promoting, inciting, instructing in or depicting abhorrent violent conduct.

While these categories of unlawful or harmful material are clearly defined by the Online Safety Act, the BOSE Determination (and the amendment to the BOSE Determination) adopts the language "unlawful or harmful" material or activity, which is not defined, and it is unclear from the Determination whether this is limited to those six categories.

The eSafety Commissioner's Basic Online Safety Expectations Regulatory Guidance suggests that 'harmful' material is that which is within the scope of the Online Safety Act but also 'material or activity that *should* (emphasis added) fall under a provider's terms of use, policies and procedures and standards of conduct for end-users (as outlined in section 14 of the Determination)'. Rather than providing clarification or indeed certainty to industry, this instead

introduces further subjectivity and ultimately leaves interpretation of the potentially broad scope of the BOSE Determination at the discretion of the Commissioner.

The concept of "unlawful or harmful" material or activity (outside of the six categories above) is very broad. What may be unlawful or harmful is:

- dependent on context (a piece of material by itself may be harmful, but not if additional information or disclosures are provided);
- the nature of the service (for instance, material may be harmful when disseminated publicly but not privately, or if stored in a user's private file-storage service);
- the intended or targeted audience for the service (for example, whether the service is targeted at adults or children or is likely to be accessible by children); and
- the personal preferences or circumstances of the individual user (for example, material about wellness, diet and exercise may not alone be harmful but could be for a user who is suffering an eating disorder).

In the context of "unlawful material", what may be an unlawful statement or material differs from country to country, and in many instances may not reasonably be able to be determined by a service provider without other information or context.

In other instances, material or activity that would ordinarily fall within a definition of "unlawful or harmful" (and may be contained within a provider's terms of service or policies) is subject to other regulatory regimes or laws. For example, scams (which falls within the remit of other regulators and is subject to a separate consultation) and misinformation/disinformation (which the government proposes to be addressed via separate legislation to be regulated by ACMA).

It is imperative that service providers know what material or activity is "unlawful or harmful" within the remit of the Online Safety Act to understand what the obligation or expectation is that they have to meet, and which regulatory regime applies. Requiring action against ill-defined categories of "unlawful or harmful" material or activity fails to provide service providers with the legal clarity they need to act. It also makes it extremely difficult to provide meaningful feedback as to the feasibility of many of the proposed amendments to the BOSE Determination.

The Government has committed to considering whether additional arrangements are warranted to address other online harms not captured under the existing statutory scheme as part of the OSA Review. **We recommend that** the Government, as part of that review define the intended scope of "unlawful or harmful" material or activity that is to be addressed under the Online Safety Act and/or the BOSE Determination, and how that interacts with, or relates, to other regulatory regimes and initiatives in this area (as outlined in i) above.

Substantive comments on proposed amendments to the BOSE Determination

1. Generative AI, recommender systems and user controls

In our view, new categories or requirements that are directed at "generative AI" and "recommender systems" are not required. The harms that may be attributed to the use of this technology are already captured and addressed under the existing, broad and flexible terms of the BOSE Determination. The role of the Online Safety Act and the BOSE Determination in regulating specific technologies should otherwise be deferred until after the OSA Review (which will consider this issue separately) and developed in consultation with broader government policy.

We acknowledge concerns about risks associated with certain uses of recommender systems and generative AI. For our part, we have already acted to implement systems, processes and technologies to minimise harms that may arise from the use of those technologies, and in fact, many of our existing systems, processes or tools appear to align with what we understand to be the intended objective of the draft amendments.

However, the introduction of new expectations directed at the use of "recommender systems" and "generative AI" are a departure from the intended flexible approach initially adopted in the BOSE Determination and Online Safety Act.

We note for example that the *Explanatory Statement to the Online Safety (Basic Online Safety Expectations) Determination 2022* makes clear:

> "The Determination itself also does not prescribe how expectations will be met. This is intended to provide the highest degree of flexibility for service providers to determine the most appropriate method for achieving the expectations".

We support this approach to regulating online safety. Importantly, the Consultation Paper acknowledges (and we agree) that section 6 of the BOSE Determination, which requires that a service provider "take reasonable steps to ensure that end-users are able to use the service in a safe manner" and "will take reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful", is broad enough to capture uses of both "recommender systems" and "generative AI" by a service provider. This section is rightly focused on the outcome, that is harmful or unlawful material itself, rather than the technology that may facilitate it but is not of itself harmful or unlawful. It is unclear that new subsections 8A and 8B add anything to this broader expectation. This conclusion is supported by the eSafety Commission's own regulatory guidance on the BOSE Determination which includes significant detail on the reasonable steps that a service provider can take if they use a "recommender system" in meeting the expectation in section 6(2).

In any case, and notwithstanding the above, we also add:

- The introduction of a new category for "generative AI" appears to overlap with the Government's stated policy response on AI safety, as articulated in the *Australian Government's interim response to the Safe and responsible AI in Australia consultation*, released in mid-January 2024. This includes the development with industry of voluntary labelling and watermarking of AI-generated material in high-risk settings and voluntary AI safety standards.
- The OSA Review will separately consider whether additional arrangements are warranted to address online harms not explicitly captured under the existing statutory schemes including (amongst others) the potential harm caused by a range of emerging technologies such as generative AI and recommender systems.

The introduction of new expectations directed at "recommender systems" and "generative AI" ahead of these processes is premature, and in the case of generative AI, may not align with the Government's stated approach to regulation of AI. While we do not believe new categories are necessary, **we recommend that** if included, consideration as to the form and content of these requirements should be deferred until after the completion of the above processes (or considered in conjunction or consultation with these processes).

I. Specific comments on Proposal 1 - Generative AI

We also make the following specific comments and recommendations regarding the proposed drafting and scope of section 8A:

- **The expectation should make clear that the "providers of the service" are the "organisations that deploy the AI" and not the developers.** We recommend there is a clear delineation of roles and responsibilities between AI developers and organisations deploying particular use cases. The organisation deploying an AI application should be solely responsible for any disclosure and documentation requirements about the AI application because it is best positioned to identify potential uses of a particular application, monitor its performance and mitigate against misuse. Even in cases where an application is provided by a developer directly to the deployer, and no modifications are made, deployers will often be best positioned to understand downstream use cases and their attendant risks, implement effective risk management strategies, and conduct on-going monitoring, which developers of general use systems are not equipped to do.
- **The expectation should define what is "unlawful or harmful" material or activity and the obligation imposed on the use of generative AI should be proportionate, and consistent with the same obligation imposed on a non-generative AI service or use.** We have set out our concerns regarding the lack of clarity of what is meant by "unlawful or harmful" above. Our concerns are particularly relevant in the context of section 8A(2) given the large number of potential uses for generative AI within a service. Additionally, we also add that section 8A(2) is directed at service providers taking reasonable steps to proactively minimise the extent generative AI is used to "create" and "facilitate" unlawful or harmful material or activity, which appears to be much broader than the standard imposed on non-generative AI applications, which is to take reasonable steps to "minimise the extent to which material or activity on the service is harmful or unlawful". As a practical matter, any technology or service could be used to create or facilitate "unlawful or harmful" material or activity. For example, a digital camera or a word processor could be used to create and facilitate "unlawful or harmful material or activity". The standard that a generative AI application is expected to meet should be consistent with the same standard of a non-generative AI application.

ii. Specific Comments on Proposal 2 - "recommender systems"

We also make the following specific comments and recommendation regarding the proposed drafting of section 8B:

- **For the purposes of section 8B, the obligation should be directed at identified categories of harmful and unlawful material.** Our concerns regarding the lack of clarity or definition of what constitutes "unlawful or harmful" material are set out in above.
- **The example of a reasonable step provided at subsection 3(b) to provide educational or exploratory tools to end-users on the risks associated with recommender systems and the example of a reasonable step provided at subsection 3(c), to enable a user to complain or make enquiries about the role a recommender plays in presenting material or activity, are neither practical or feasible and we recommend against including.** We agree that service providers can (and should) provide general and transparent information about how their recommender systems work. YouTube already does this. However, because of the number of different factors that may influence what and how material is recommended, it is unlikely that a service provider could meaningfully respond to a complaint or an inquiry about why a specific piece of material has been shown to a particular user at a particular point in time. Also, given that the number of ways in which a user may interact with and make use of recommender systems is diverse and the risks would also differ depending on the profile of the user, it is unlikely that a service provider could meaningfully capture the top risks relevant to the specific user. We suggest the examples of reasonable steps are better directed at service providers having in place readily accessible mechanisms to "report harmful content" (which enables a service provider to take action to remove the harmful material from the service) or to provide appropriate "user controls" to optimise their recommendation experience (e.g., removing recommended material, adjusting the personal information used to influence recommendations).

2. The best interests of the child and preventing access to age-appropriate materials online.

i. Proposal 1 - Best interests of the child

While we share the Government's objective to protect children online, an express obligation to give effect to the objectives of Article 3 of the Convention on the Rights of the Child is best addressed holistically, with the involvement of a range of stakeholders, as part of the OSA Review. If new obligations are to be introduced, we recommend that it align, and guidance be provided that service providers can meet this obligation by showing compliance with one or more of the global initiatives or codes in this area (for example, the UK AADC).

We share the Government's objective to protect children from online harms. When designing our products and services, we consider the online harms children may face and have developed a number of special features to enhance the safety of children online. Through Family Link, we allow parents to set up supervised accounts for their children, set screen time limits, and more. Our Be Internet Awesome digital literacy program helps kids learn how to be safe and engaged digital citizens. The dedicated YouTube Kids app is a separate app for our youngest users and offers a safer and simple place where kids can learn and explore their interests; then as children get older, parents have the choice to allow their family to explore more of YouTube with a supervised experience. Kids Space and teacher-approved apps in Play also offer experiences that are customised for younger audiences.

The "best interests of a child" is a concept drawn from the Convention on the Rights of the Child. The Committee on the Rights of the Child, in its General Comment No.25 on children's rights in relation to the digital environment, recognises the "best interests of a child" as one of four general principles that should inform measures needed to guarantee the realisation of children's rights in this context. The General Comment recognises that this principle is a dynamic concept that requires an assessment appropriate to the specific context and in considering the best interests of the child, regard should be had to all children's rights, including their right to seek, receive and impart information, be protected from harm and to have their views given due weight - many of which may not directly relate to online safety. Importantly, General Comment No. 25 also states that States parties should ensure transparency in the assessment of the best interest of the child and the criteria that have been applied. As drafted, it is unclear how service providers are expected to interpret and give effect to this concept without further articulation of those factors which might be relevant to its assessment in this context.

The Government has committed to considering "best interests of the child" obligations as part of the OSA Review. Separately, in the Government's response to the review of the Privacy Act released 28 September 2023, the Government also committed to the introduction of a Children's Online Privacy Code, which will apply to online services likely to be accessed by Children, and to the extent possible, align with the UK AADC.

We suggest that any express obligation to give effect to the objectives of Article 3 of the Convention on the Rights of the Child is best achieved holistically via the above two processes. Introducing a new obligation in the BOSE Determination ahead of these processes risks duplication and an inconsistent regulatory regime.

If however section 6(2A) is to be included, **we recommend** the following amendment:

> *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* ~~*operation of any service that is used by, or accessible to, chil*~~ ***and development of online services likely to be accessed by a child***

The purpose of this amendment is to ensure:

- The obligation is appropriately directed at the design and implementation of a service. We note that the Consultation Paper states the intent of section 6(2A) is to direct consideration of the interests of the child at the "design and implementation" stage. This change ensures the drafting is consistent with that intent.
- It appropriately targets those services that are likely to be used by children, and not any service that may happen to be used by an individual "child".

We also recommend that given the global initiatives already in place, the Government provides guidance that service providers can meet this obligation by complying with one or more of these global initiatives or codes (for example the UK AADC).

3. Hate Speech

While we support the Government's efforts to tackle hate speech, we strongly urge the Government to address the harm caused by hate speech holistically, in a democratic and transparent manner, as part of the OSA Review or indeed through broader regulation of hate speech, both on and offline. We recommend that any consideration of amendments to the BOSE Determination that specifically address hate speech be considered as part of this review, or deferred until after this process has been completed.

We support the Government's efforts to tackle hate speech in Australia. Tackling hateful material online is a top priority for Google. Hate speech is not allowed on our products that host user-generated content. Under our Community Guidelines and content policies, we will remove material promoting violence or hatred against individuals or groups based on attributes like race, religion, gender, and ethnicity. We also receive and evaluate legal removal requests to remove material under local law (where applicable).

We are, however, concerned with an approach that targets only those service providers who have voluntarily sought to address legal but harmful material with additional compliance and reporting/transparency obligations under a subordinate legislative instrument. This does not promote online safety nor address the harm caused by hate-speech across society.

In Australia, hate speech regulation has proven complex. There is no uniform legislative definition in Australia of unlawful "hate speech".

The Government has committed to reviewing "online hate" as part of the OSA Review. The Government has separately proposed that "hate speech" is included as a 'harm" under the proposed Misinformation Bill. We urge the Government to address the harm caused by hate speech holistically, at a society level, through a transparent, democratic process rather than indirectly via subordinate legislation that is reliant on a private entities' own interpretation of what they consider to be an appropriate framework.

We recommend that any amendments to specifically address hate speech in the BOSE Determination are deferred until at least the completion of the OSA Review.

4. Safety impacts of business and resourcing decisions

<u>Proposal 1 - Reasonable steps regarding business decisions affecting user safety</u>

While we agree that the obligation to ensure that a service is "safe to use" should extend to not only ensuring the service is safe to use during the design and implementation phase, but also its ongoing

operation, we believe that the matters intending to be addressed by subsection 6(3)(f) are already adequately addressed by the examples provided in section 6(3)(c) - (e) of the BOSE Determination, and as such, section 6(3)(f) is unnecessary.

We understand that the intent of new section 6(3)(f) is to clarify that service providers need to consider safety impacts not only in the design and implementation of their products, but also in the making of business decisions that will likely have an adverse impact on the ability of the service provider to operate their service is a safe manner. The Consultation Paper suggests that, while this is only an example of the type of steps a service provider may take to comply with section 6, the type of business decisions will be contained in regulatory guidance, and would include, for instance "major staffing cuts" or "removing staff from certain countries". It also notes that the framing of "significant adverse impact" is designed to provide flexibility to the Commissioner in issuing reporting notices relating to a range of organisational decisions.

We agree with the proposition that service providers should take steps to ensure that not only must they take steps during the design and implementation phase of their service, but also in their ongoing operation, to ensure that their service is safe to use for end-users.

However, in our view, this is already adequately captured by the examples set out in section 6(3)(c), (d) and (e), which provide that:

- "persons who are engaged in providing the service, such as the provider's employees or contractors, are training in and are expected to implement and promote, online safety";
- that service providers must "continually improving technology and practices relating to the safety of end-users"; and
- that service providers must take steps to ensure "that assessments of safety risks and impacts are undertaken, and safety review processes are implemented, throughout the design, development, deployment and post-deployment stages for the service".

Further, the Industry Codes and draft Industry Standards, include additional, specific obligations that are directed at ensuring adequate resourcing of trust and safety teams and ongoing investments in improving systems to detect and address class 1 material. We assume the intent of proposed section 6(3)(g) and 6(3)(h) in the BOSE Determination are to broadly replicate those requirements for a broader range of harmful material.

We therefore question what purpose section 6(3)(f) is intending to serve in addition to the examples of reasonable steps already provided. Further, we are also concerned that the premise or reason for the inclusion is to interrogate a broader range of business decisions with regard to matters such as budgeting or corporate restructures (which are complex and highly commercially sensitive) on the assumption that such changes must be inherently bad for user safety, which is not necessarily the case.

We recommend that the example suggested by section 6(3)(f) is removed.

5. Transparency

We are broadly supportive of efforts to improve transparency and accountability of service providers. However, the additional transparency expectations proposed by section 18A are extremely prescriptive, onerous and will be difficult for service providers to comply with. They are also duplicative of existing compliance reporting obligations under the Industry Codes, Standards and powers of the Commissioner to request the information (on a per service and as needed basis) under the Online Safety Act.

We agree on the importance of openness and have launched a number of tools to ensure there is transparency on how we are tackling online harms. Google has been an industry leader in transparency. We publish a range of transparency reports, including reports that share data on the number of government content removal requests, along with a separate report on YouTube Community Guideline Enforcement and Google's efforts to combat online child safety abuse. We also share useful information about how our key products work, and how users can stay safe online, and for YouTube information about how our creators can stay safe, when using our products.

We are committed to continuing our transparency efforts; however we are concerned that the transparency reporting requirements proposed by section 18A are onerous, prescriptive and will be very difficult for service providers to meet.

As drafted, the new expectation requires service providers to provide information, broken down by country and individual service (irrespective of risk or potential harm) about:

- The service's enforcement of its terms of use, policies and procedures and standards of conduct at an individual country level.
- The safety tools and processes deployed by the service and their effectiveness.
- Metrics on the prevalence of harms, reports and complaints and the service's responsiveness.
- The number of active end-users in Australia (including children) for each month during the reporting period.

Given the complexity of our enforcement systems, it takes considerable effort to produce accurate and reliable reporting suitable for public release. Producing multiple iterations for each individual service (irrespective of the risk of the service or the harm) across all markets would create an undue burden for businesses, and especially for smaller companies, and may not meaningfully add additional insight into how a service provider is progressing in addressing online harms. And reporting requirements that may be appropriate for one service may be ill-suited for another. The benefits of transparency (particularly metrics as to the effectiveness of internal processes to detect and identify certain harms) must also be balanced with the need to ensure that bad actors do not game a platform's systems through manipulation, spam, fraud and other forms of abuse, and the need to protect a provider's commercially sensitive information.

We also emphasise that these reporting obligations are in addition to existing reporting requirements that already exist under the Industry Codes, draft Standards and BOSE Determination:

- **Social Media Services Online Safety Code for class 1 material.** Tier 1 Social Media Services are required to produce an annual compliance report which must at a minimum contain information as to the steps taken to comply with the minimum compliance measures, the volume of CSEM and pro-terror material removed from the service and an explanation as to why the measures are appropriate. This is in addition to a separate obligation to provide at a dedicated location, information to end-users about online safety, including information for parents and carers to manage a child's access to class 1 material and available safety settings.
- **Draft DIS and RES Standards for class 1 material.** In the current draft, certain service providers are to provide reports to the Commissioner on any risk assessments undertaken, technical feasibility, outcome of development programs and compliance reports with each individual obligation under the Standard in respect of "class 1 material".
- **BOSE Determination**. The Commissioner has the power to request ***all*** *of* the information identified in proposed section 18A on a service specific or "as needs" basis, via: (a) non-periodic reports as to the extent to which a service provider is complying with one or more basic online safety expectation; and (b) periodic reports as to the extent to which a provider complies with one or more basic online safety expectations.

We agree transparency reporting can provide valuable information for the public and regulators in understanding the safety measures in place, and how effective those measures are - however reporting requirements must be proportionate and appropriately targeted. In the context of the Online Safety Act, this is already achieved via the existing reporting required under the Industry Codes and Standards (which are targeted at the most serious types of harms) and the existing powers of the Commissioner to request information via its existing powers under the Online Safety Act. In our view, new and additional transparency requirements are not required.

If the Government believes that there are gaps in existing transparency requirements under the Online Safety Act, we suggest that this should be considered as part of the OSA Review with a view to harmonising transparency requirements and reducing regulatory burden on industry.

6. Enforcement of terms of use

I. Proposal 1 and 2 - Detecting breaches of terms of use, policies and procedures and standards of conduct

A service provider's terms of use, policies and procedures or standards of conduct cover a large number of different matters, which pose varying levels of harm and risk to other end-users. Service providers should be provided with the flexibility to enforce their terms of service in a manner that is appropriate and proportionate taking into account the nature of the service, the type of risk and the potential harm to other end-users. We recommend that, rather than focusing on an expectation that service providers take reasonable steps to "detect (including proactive steps)" such breaches, proposed subsection 14(1A) is instead focused on requiring service providers to take appropriate and proportionate action to enforce its terms of service, policies and procedures.

Proposed subsection 14(1A) is a new additional expectation that requires a provider of the service to take "reasonable steps (including proactive steps) to detect breaches of its terms of use, and where applicable, breaches of policies and procedures in relation to the safety of end-users or standards of conduct."

While we appreciate the intent, it is important to emphasise that a provider's terms of use or policies or standards of conduct, cover a large number of different matters which pose varying levels of harm or risk to other end-users. Proactive detection of every type of potential breach may not be feasible or proportionate for all types of breaches and for all types of services. Service providers should be provided with the flexibility to assess and decide the appropriate action that is required to enforce their terms of service depending on the type of service, the nature of the risk and the potential harm to other end-users.

To the extent that this obligation is directed at ensuring that service providers have processes, systems and technologies in place to minimise the extent material or activity on the service is unlawful or harmful, this is already captured by section 6. For the most serious and harmful material (child sexual abuse material and pro-terror material), the Industry Codes and draft Industry Standards impose mandatory requirements for service providers to implement "systems, processes and/or technology" to "detect and remove" and to "disrupt and deter" such material from its service. A further expectation that a service provider is expected to "proactively detect" any breach of its "terms of service", irrespective of harm or risk or type of service, is not warranted.

We recommend that the expectation instead focus on requiring service providers to take appropriate and proportionate action to enforce its terms of service, policies and procedures.

II. Proposal 3 - Timely resolution of complaints and reports

At the scale that many service providers operate, a requirement that service providers provide feedback in response in every user report or complaint would be extremely resource intensive and not practical. We

recommend that proposed subsection 14(3) is limited only to an obligation to review and take appropriate action in response to a report of harmful material within a reasonable period of time.

Proposed subsection 14(3) creates a new additional expectation that a provider must within a reasonable period of time, review and respond to reports and complaints and provide feedback on the action taken.

We agree that service providers should have in place mechanisms for users to report harm on a service. However, given the scale that many service providers operate at, to provide feedback (or at least meaningful feedback) in response to individual reports would be extremely resource intensive, and in many instances, not practical or feasible. This is especially so when user reports can be extremely unreliable signals of material that is unlawful or harmful, as users may flag material for different reasons, including to express displeasure. To put this into context, in the period July 2023 to September 2023, YouTube received over 22 million reports of videos that were flagged by human reviewers. Even providing feedback to a small share of vexatious or frivolous reports would mean an unnecessary take-up of resources that could be better allocated to other uses addressing the highest risks to users, or that could more effectively keep users safe.

We recommend that proposed subsection 14(3) is limited only to an obligation for service providers to review and take appropriate action in response to a report or complaint within a reasonable period of time.

Conclusion

We thank the Government for the opportunity to review and comment on the draft to the amendments to the BOSE Determination and remain available to provide further information and answer any questions on these materials as required.

[END]