Submission to the Statutory Review of the Online Safety Act 2021

5 July 2024

Executive Summary

Google welcomes the opportunity to contribute to the independent review of the *Online Safety Act 2022* (the Act). The review, required under the Act, is a timely opportunity to reflect on how the Act has operated in practice since its commencement in January 2022. We appreciate the opportunity to share our insights on how the Act has worked to further the objectives of improving and promoting the safety of Australians online, as well as provide our thoughts on where improvements could be made to ensure these objectives are being met.

In the first part of our submission, we provide our thoughts on the broader operation of the Act since its commencement. In our view, the Act has evolved into a complex regulatory regime which presents challenges to users, government and industry in delivering on its objectives. We make the following recommendations **to support a simpler and more coherent approach to online safety regulation** in Australia:

1. The multiple schemes under the Act need to be harmonised.
2. The Act should adopt a simpler approach to regulated sections of industry, based on risk.
3. Any reforms to the Act should adopt a risk-based, technology-neutral approach to online safety that places equal emphasis on managing risk and protecting human rights.
4. Adopting global trends may increase regulatory harmonisation, but requires careful consideration.
5. This review should properly take into account broader regulatory initiatives of relevance to ensure a systematically coherent approach to digital regulation.

The second part of our submission is focused on the **operation of the separate schemes and processes** under the Act:

1. Codes and standards
    a. Google supports a co-regulatory approach to Industry Code development, with Industry Standards as a last resort, and a requirement for real and substantial consultation and more realistic timeframes built into the Act.
    b. The scope of the Industry Codes and Standards should be decoupled from the National Classification Scheme.
    c. Industry Codes and Standards should not extend to certain types of content, which should instead be subject to appropriate legislative oversight.
2. Basic Online Safety Expectations (BOSE)
    a. The Act should be amended to include better guardrails around the exercise of the eSafety Commissioner's powers to require service provider reporting.
    b. The intended scope of "unlawful or harmful" material or activity under the BOSE Determination should be defined, and how that interacts with, or relates to, other regulatory regimes and initiatives should be clarified.

    c. The reporting regime should be subject to a robust confidentiality protocol.
3. Complaints and content based removal schemes
    a. We support the operation of the complaints and notice-based removal scheme under the Act.
    b. Removal notices should include the basis for determining that content meets relevant thresholds.
    c. Content removals do not address underlying causes of harm.
    d. Most categories of content underpinning the removal schemes are well-defined, and sufficiently broad to cover a range of harms.

Introduction

Google is supportive of effective content regulation and shares the Australian Government's goal of ensuring regulation helps keep users safe from bad actors while protecting the core benefits of online environments, including the ability of users to express lawful speech openly, access useful information and connect with one another. We believe in the power of the open internet and how it acts as a catalyst for innovation, economic growth, education and social well-being.

At Google, we have been working to address this evolving challenge for years, ensuring the right policies are in place to protect our products and users, and using both technological tools and human reviewers to identify and stop a range of online abuse, ranging from disinformation to child sexual abuse material. A mix of people and technology helps us identify illegal and harmful content and enforce our policies, and we continue to improve our practices and remain committed to transparency through regular updates to our Community Guidelines Enforcement Report.

Google has not waited for regulation before acting to keep our users safe. We are constantly improving and introducing new policy changes to support online safety and continuing to invest in technology to help us tackle illegal and harmful content at scale.

We recognise that tackling this problem is a shared responsibility, and we want to offer our thoughts to contribute constructively to the conversation drawing on our expertise and experience.

Need for a simpler and more coherent approach to online safety regulation

We support regulation to better protect and empower people online. At Google, our goal is to ensure that regulation helps keep users safe from bad actors while protecting fundamental rights and the core benefits of online environments. We want our services to remain open and innovative, and for Australian users to be able to express lawful speech openly, access information, and connect with one another. We believe effective regulatory frameworks reflect the shared responsibility to tackle online safety, set out clear rules so services know how to fulfil their legal obligations, and remain flexible to accommodate new technology and innovative approaches.

But we are concerned by what is becoming an increasingly complicated, overlapping and confusing regulatory regime within Australia. Since its commencement in January 2022, the Act has evolved into an overly complex regulatory regime which includes not only a content and complaints based removal scheme and Basic Online Safety Expectations (BOSE), but also multiple industry codes and standards.

Examining the operation and effectiveness of the existing regulatory regime and considering ways to streamline and harmonise obligations to ensure the Act continues to meet its objectives should be the priority of this review and needs to be addressed before any expansion of the Act is contemplated to avoid further exacerbating existing challenges.

Where feasible, this review should also take into account other regulatory initiatives currently under consideration by the Government that impact digital service providers. Many of these initiatives overlap with or are directly relevant to the issues being considered as part of this review. Not doing so risks introducing systematic complexity and inconsistencies across digital services regulation in Australia.

1) **The multiple schemes under the Online Safety Act need to be harmonised**

We urge this review to consider how the different schemes under the Act interact with each other and how best to streamline the different - and sometimes conflicting - obligations under these schemes to reduce regulatory and compliance burden on service providers, while still achieving the intended purpose and objectives of the Act.

The relationship between the schemes under the Act - in particular the BOSE and Industry Codes and Standards - is unclear, and obligations under each of the schemes are not necessarily aligned across the different regulated sections of industry nor are they consistent for the same regulated section of industry under each scheme. This presents particular

challenges for companies, such as Google, who provide multiple services that are each subject to multiple separate regulatory regimes under the Act.

For example, Google services are subject to three separate regulatory regimes for addressing class 1 content under the Act. If a service meets the definition of a hosting service, equipment, social media service, app distribution services and/or a search engine service then they are subject to Industry Codes. Whereas, if the service meets the definition of a designated internet service and restricted electronic services they are subject to Industry Standards. The obligations between the Industry Codes and the Industry Standards are not aligned, and in many instances the differences do not appear associated with any particular or increased risk with the service. And for those services that meet the definition of social media services, relevant electronic service and designated internet services, they must also comply with the BOSE Determination for the same class 1 material. Again, the expectations overlap but are inconsistent with equivalent obligations under both the Industry Codes and Standards.

The practical impact of the above is that:

- **Providers cannot adopt uniform compliance solutions across 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 the online safety outcomes.
- **Providers can be compliant with the mandatory requirements under an Industry Code or a Standard, but still be non-compliant under the BOSE Determination for the same type of obligation or content.** 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.
- **Providers are subject to duplicative reporting and transparency requirements.** While the Industry Codes and Standards require certain compliance reporting requirements, the Commissioner can separately issue a notice to answer additional questions about the same content and/or obligation under the BOSE Determination.

The complexity will be further exacerbated by the introduction of an additional set of Industry Codes and/or Standards for class 1C and class 2 material ("**Phase 2 Industry Codes**"). To avoid this - and to ensure that these Codes are consistent with and informed by the outcomes of this review, as well as the other relevant government initiatives under consideration, notably the Government's age assurance trial - we suggest the finalisation of the Phase 2 Industry Codes be delayed. We note that the eSafety Commissioner is requesting final Phase 2 Industry Codes by 19 December 2024.

We made similar arguments in our submission on the proposed amendments to the BOSE Determination. We were, and remain, concerned about the significant overlap between many of the proposed amendments to the BOSE Determination with the issues being considered under this review. We were also concerned that proceeding with amendments to the BOSE Determination may in part pre-determine the outcome of this review.

2) **The Act should adopt a simpler approach to regulated sections of industry, based on risk**

The current approach of categorising different regulated sections of industry introduces unnecessary complexity and rigidity into the Act. As outlined above, Google's products and services fall within seven of the eight regulated sections of industry and are subject to different regulatory requirements for different products as a result.

As is rightly noted in the Issues Paper, the challenges of this approach have been most apparent in the context of the Industry Codes, where this categorisation has not allowed sufficient flexibility to take account of differences between services within each regulated section of industry and for certain categories, the breadth of services covered makes the application of consistent measures for that section challenging. This was most apparent in the development of a single code for Designated Internet Services given the different risks associated with the breadth of services covered (websites, apps and file-storage services).

At a more practical level, the complexity of the current approach has the potential to create confusion and uncertainty for Australian end-users. The Australian Government's own Principles for clearer laws states that legislation 'should enable those affected to understand how the law applies to them'. Yet in the context of the Industry Codes, for instance, Australians would be required to navigate a complex categorisation of services across six separate codes and two standards in order to understand obligations on industry with respect to class 1 content. This ultimately risks undermining the ability of the Act to achieve its objects and purpose.

Not only is the categorisation of regulated industry sections unnecessarily complex but it is also too broad. Google remains of the view that at the very least enterprise services, that is business to business (B2B) services, should be excluded from the scope of regulation. Providers of B2B services operate on a completely different model to consumer services. They are often subject to heavily negotiated service agreements and service providers are typically prohibited from exercising any control over their customers' content. For example, the cloud provider typically does not have visibility into its customers' content to meet the privacy, security, and regulatory demands of its customers (and of their end customers), and to comply with existing laws and regulations governing cloud based services. Users of such services - whether business entities, public sector organisations, or healthcare and education providers -

entrust service providers with their confidential data and need to be able to remain fully in control of it.

The inclusion of B2B services within the scope of the Act risks the loss of trust, confidentiality, and security for customers, which would ultimately undermine the foundations of cloud services in the Australian economy. It could also lead to the imposition of unworkable sanctions. Even if something was flagged by an external observer, it is often impossible for a cloud provider to remove individual pieces of content. For example, it could lead to a service provider needing to remove a customer's entire website where the provider does not have control over individual pieces of content - a clearly unworkable change to the way online services operate.

To effectively deal with online safety, services should be scoped in based on level of risk, and not based on size or business model. The regulatory regime should protect against illegal content migrating across platforms by ensuring a consistent set of rules for all market players. We acknowledge that not all services have the same level of resources. However, the migration of content from mainstream sites to less moderated platforms, often with niche user bases, is a worrisome trend that analysts have observed with terrorist content, violent extremism, and child sexual abuse imagery. This would ensure obligations were appropriately targeted to meet the objectives of the Act and be a better use of the resources of both Government and industry.

3) **Any reforms to the Act should adopt a risk-based, technology neutral approach to online safety that places equal emphasis on managing risk and protecting human rights**

We support a risk-based approach to regulation. Risk-based approaches can ensure a more targeted and proportionate approach to online safety, avoiding unnecessary burdens on lower-risk services. It can also help to appropriately tailor obligations to service, based on their functionality and degree of control over user content.

We support an equal emphasis on managing risk and protecting human rights. Online safety is broader than keeping Australians safe from harmful content. It also includes protecting Australians' privacy, security and personal information. And measures to mitigate harms necessarily involve consideration of various rights and freedoms, including the right to access information and rights of free speech and expression, which can be in tension.

Protecting and respecting the fundamental human rights of Australians should be a priority for both services and regulators alike. We strongly encourage safeguards to be added to the Act to protect fundamental rights and freedoms, including privacy rights, freedom of expression, and equality rights. We also suggest the Act should be amended to include concepts of

reasonableness and proportionality, which acknowledge the need for service providers to balance these interests in a way that is proportionate to harm and best protects user safety overall.

We believe that the Online Safety Act should remain technology neutral, focused on the outcome - that is harmful or unlawful material - rather than the technology that may facilitate it but is not of itself harmful. This ensures that the Act remains both focused and future proofed.

4) **Adopting global trends may increase regulatory harmonisation, but requires careful consideration**

Google is broadly supportive of regulatory harmonisation and global interoperability, reflecting the global nature of the internet. Adopting consistent approaches provides clarity and certainty to users, online services, and policymakers, and enables better and more consistent experiences.

The Issues Paper discusses a number of concepts from international regulatory approaches. We provide some comments on those approaches below, based on our experience. However, it is critical that these concepts are not considered in isolation from the regulatory context in which they operate. For example, the UK's duty of care principle operates in the context of regulation focusing on systemic protections. It would operate very differently under a complaint-based approach to individual items of online content. In the same way, before a fundamentally new concept is introduced, its interaction with the current Act must also be considered - layering a duty of care-type concept on top of the existing BOSE and Industry Codes and Standards would greatly exacerbate the existing complexity of the Act.

Duty of Care

A duty of care is appropriate in a systemic model, rather than a complaint-based approach to individual items of online content. It must be clear that services' responsibility should be limited to systemic failures to comply with the duty to act responsibly, and that, as in the UK, enforcement resides exclusively with the regulator rather than with individual users.

Should a duty of care be adopted, safeguards are required to ensure that risk mitigation measures do not raise undue risk for fundamental rights, reporting on risks does not expose sensitive information, and obligations are proportionate.

The duty of care must place significant emphasis on safeguarding fundamental rights. The regulator should have an obligation to consider risks to fundamental rights when evaluating the sufficiency of risk mitigation measures.

Obligations to undertake risk assessments should be proportionate to ensure they are manageable for services and regulators alike. They should also give services sufficient space and flexibility to manage risk. We routinely make changes to address harmful content on our platforms, including in response to real world crises, and need to be able to assess emerging risks and make changes swiftly.

To protect the integrity of risk mitigation systems, reporting on risk assessments and mitigation should only be made accessible to regulators. Public reporting on the vulnerability and mitigation measures services undertake would expose sensitive information, and open up our systems for exploitation by nefarious actors.

Best interests of the child and protections for children

At Google, we aim to balance delivering information with protecting users and society. We take this responsibility seriously. Our goal is to provide access to trustworthy information and content by protecting users from harm, delivering reliable information and partnering with experts and organisations to create a safer Internet.

We understand how critical this is, especially when it comes to children. We know that Australian children and teenagers are increasingly using digital devices. Government, parents, educators, child-safety and privacy experts are rightly concerned about how to keep our children safe, and we share those concerns. Our commitment to doing so is demonstrated through the systems and processes we have in place to respect first and foremost the laws and regulations of Australia, and then to apply Google's terms of service and content policies.

When designing our products and services, we consider the online harms children may face and work with experts to develop products, tools and policies to enhance the safety of children online.

We build age-appropriate products that align with kids' and teens' developmental stages and needs. Family Link is a downloadable app that helps parents set digital ground rules for their children, including through content controls. YouTube Kids is a standalone app with more parental controls for our youngest users and offers a safer and simple place where kids can learn and explore their interests. We also offer supervised experiences on the main YouTube platform, where a parent or caregiver creates and links a child's account to their own. Supervised experiences come with three tailored content settings as well as privacy protections, parental control and limited features.

We also offer a number of settings and tools that give families flexibility to manage their own unique relationships with technology. For example, Safe Search offers protections to help filter out explicit content - such as adult or graphic violent content - in Google's search results across images, videos, and websites when enabled. SafeSearch is on by default for users under 18. In addition, explicit imagery is blurred by default when it appears in Search results. On YouTube, Autoplay is turned off by default for all users younger than 18 across all of YouTube's products; "Take a Break" and bedtime reminders are on by default for users younger than 18 on YouTube and YouTube users with supervised experiences; and for users under 18, we set the default upload, livestream, and livechat settings to the most private setting available.

Finally, we have strict content and privacy policies in place to protect our young users across our products, including for the ads kids see. We regularly review and update these policies and roll out product improvements. On YouTube, our Community Guidelines outline the types of content that are not allowed, including cyberbullying, suicide and self harm, and content that endangers the emotional and physical well-being of minors. They also detail our approach to age-restricted content, which is only viewable by users who are signed-in and have an account age of 18 or older. When it comes to Ads, we do not allow personalised advertising to minors based on age, gender, interests, we restrict sensitive ads categories (e.g., tobacco and alcohol, dangerous activities, weight loss, sweepstakes, etc.); and for our youngest users on YouTube Kids, Made for Kids content and in supervised experiences, we prohibit ads in additional categories such as foods and beverages, religion, or politics, as well as ads with inappropriate content such as scary imagery, crude humour, or sexual innuendo.

Understanding the age of our users forms a key part of our efforts to ensure children and teens have appropriate experiences when using our products and services. We use various tools to understand the age of users or for age assurance purposes. We also use other tools and services - some of which are product-specific - to limit access to content that is inappropriate for children.

We agree that a smart and strong regulatory framework for children and teens starts by supporting their best interests. But it is important that any "best interests of the child" requirement should clearly define what those interests are, and do so in a holistic way that weighs considerations such as safety, physical and mental wellbeing, privacy, agency, access to information, and freedom of participation in society.

Well-crafted legislation should take all of those rights and freedoms into consideration. Online services used by children and teens should be required to assess the collective interests of children within comparable developmental stages, based on expert research and best practices, to ensure that they are developing, designing and offering age-appropriate products and services geared to the best interests of children and teens.

This is consistent with the Committee on the Rights of the Child's General Comment 25, which 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.

Transparency and Data Access

Data access by regulators must be proportionate and protect confidentiality. We understand that data requests are an important oversight component. However, regulators should be required to consider the burden on services and the risks of data disclosure. Access requests should be proportionate and data kept secure and confidential, used for a specific purpose, and then deleted. Emphasis should be given to working with companies to explain what the data means and how it should be used. Where further data is needed, services should be given appropriate timeframes to gather the information.

We note the eSafety Commissioner already has - and exercises - broad information gathering powers under the Basic Online Safety Expectations scheme and a number of the Phase 1 Codes include transparency reporting requirements.

Data access by researchers must also be proportionate and include adequate safeguards. We recognise that researchers need data to scrutinise or investigate issues of societal concern. Google has significant experience providing access to platform data through tools and datasets, including through its FactCheck Claim Search API which allows researchers and others to query fact checking information that is available to other users via Google's Fact Check Explorer tool and by making available datasets such as the Google Health COVID-19 Open Data Repository, YouTube-8M (a labeled video dataset of over 8 million YouTube video IDs), and Open Images (approximately nine million annotated URLs to images). In July 2022, Google launched the YouTube Researcher Program to provide scaled, expanded access to global video metadata across the entire public YouTube corpus via a Data API to eligible external academic researchers.

Through this experience, we are aware of the challenges in safely and securely providing that access to appropriate researchers. Any proposal to require researcher access to data should include robust safeguards around what data may be requested, how such data may be accessed, and what may be done with the data, such as:

- Define a "reasoned request" to set parameters around what information can be requested and shared with vetted researchers . For example, specific categories of data may need to be excluded from the scope of this provision to ensure that providing access to them does not interfere with law enforcement investigations.

- Allow online platforms to take additional measures to protect the privacy of data subjects (e.g. through pseudonymisation), where appropriate.
- Allow online platforms to object to methods of data transmission that they do not consider sufficiently secure, and to set limits on what can be done with the data and clarify that the data should not be further shared/disclosed.
- Require transparency on any funding researchers receive as part of their vetting process. "Commercial interests" might not cover researchers who, for example, have major academic projects funded by competitors or critics of the very large online platform at issue.
- Aligning proposal requirements with existing institutional research ethics processes would facilitate a shared understanding of ethical considerations between researchers and platforms.
- Allow services to appeal the vetting of a particular researcher and stop the flow of data access.
- Provide flexibility for services to respond to requests, depending on the scale of the data sought, and allowing room for queries and clarifications.

We also note that researcher access to data is being separately contemplated under other regulatory schemes in Australia. Given the potential complexity in administering and responding to these schemes, we urge any program to be streamlined under a single regulator.

Dispute resolution

The question of whether to introduce an ombuds scheme for digital platforms in Australia should be harmonised with the process already being conducted by the Government, and not as part of the review of the OSA. The digital platform industry has already been tasked with developing an internal dispute resolution standard by July 2024.

In our February 2023 submissions in response to the Government's consultation on the ACCC's Report on Platform Regulation, we detailed our view that, if the Government considers that an additional external ombuds scheme for digital platforms is required, the process and scope of that scheme need to be very carefully designed to ensure that the cost and complexity of adjudicating complaints can be kept proportionate to their seriousness. An ombuds scheme may be an appropriate, efficient and effective means of resolving transactional disputes. Any ombuds scheme should be limited to such disputes.

In contrast, the sort of disputes that might fall within the purview of eSafety under the OSA are likely to be highly challenging content-based disputes. Any user of the web, from anywhere in the world, may make a complaint to Google about products like Search, YouTube, or Maps, or indeed about a Google Ad they see online.

On YouTube, we provide internal appeals systems to allow users to contest decisions to

remove their content or terminate their accounts. For video sharing services, it would not be proportionate, efficient or effective to extend mandatory appeal systems beyond these categories. In particular we are mindful of the risk of bad actors overwhelming these mechanisms with spurious appeals. While we support functionality that enables users to flag potentially violative or illegal content, extending appeals to these flags is problematic due to the inaccuracy of user flags—e.g., during the three month period ending March 2024, less than 2% of the more than 22M videos flagged globally for review under YouTube's Community Guidelines were ultimately removed after human review of that content.

An internal appeals requirement does not make sense for all services. Notably, online search services do not host the pages they index and often do not have a relationship with the author of the content. Offering appeals to those whose content is delisted from a search index, therefore, is often impossible. The EU Digital Services Act (DSA) recognises the different functionalities and responsibilities of online services, and places internal complaint-handling requirements on online platforms only.

An ombudsperson would face considerable challenges in addressing the scale of individual complaints. Our platforms provide many tools that contribute to user control and platform accountability. As discussed above, YouTube provides users with the ability to flag content and view decisions on content they've flagged through a dashboard, and YouTube receives a high volume of flags with a very low actionability rate. Users often use the flagging tool to express dislike of a video, not because it violates any policy or is unlawful. If even a fraction of these users' flags resulted in a complaint through the individual complaints mechanism, then—compounded with the complaints from all other platforms—the system would be overwhelmed and paralysed, ultimately undermining the effectiveness and objectives of such a complaint process. For this reason, our view continues to be that any ombuds scheme should be limited to transactional complaints.

Regulation should not introduce independent recourse that can revisit or overturn platform decisions. The out-of-court redress provisions in the DSA are expansive—raising considerable concerns about manageability and risks for fundamental freedoms—but even they do not go so far as to allow for a binding ruling on services to remove or reinstate content. This is appropriate, as binding decisions that overturn content removals by online platforms, products, and services would amount to forcing a service provider to host or display content. We have already noted the considerable challenges of volume that a recourse body would face.

Google works hard to provide transparency to its removals processes and decisions. We publish transparency reports at https://transparencyreport.google.com/.

5) **This review should properly take into account broader regulatory initiatives of relevance to ensure a systematically coherent approach to digital regulation**

Beyond the Act itself, there are a number of other broader government initiatives impacting digital service providers, many of which overlap with the issues under consideration as part of this review. This includes:

- **The current review of the National Classification Scheme**, noting that the National Classification Scheme underpins the definition of class 1 and class 2 material in the Online Safety Act.
- **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 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 consideration of amendments to the Act 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 consideration of whether a "best interests of the child" principle should be included in the Act.
- **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 consideration of whether "hate speech" should be included as a harm protected under the Act.
- The development of **hate speech regulation** being explored by the Attorney-General**.**
- **Joint Select Committee on Social Media and Australian Society**, which will examine a number of the issues under consideration by this review and is due to report by 18 November.
- **Development of an industry Internal Dispute Resolution (IDR) Code**, which will be directly relevant to the consideration of IDR or EDR in the context of this review.
- The **Government's age assurance trial** which should inform the Phase 2 Codes.

This review should properly take into account these initiatives in considering reform of the Act to consider opportunities for harmonisation and avoid regulatory overlap and compliance burden on industry.

Online Safety Act: systems and processes

Codes and standards

1) **Google supports a co-regulatory approach to Industry Code development, with Industry Standards as a last resort, and a requirement for real and substantial consultation and more realistic timeframes built into the Act.**

Google is supportive of the existing co-regulatory approach to Industry Code development, with Industry Standards imposed as a last resort.

We are concerned that the complexity and breadth of regulated sections of industry (discussed above) has the potential to undermine this approach. The Act currently allows the eSafety Commissioner to move to the development of an Industry Standard where there is no body or association that represents a particular industry section. The Act is silent on what qualifies as 'representation". Given the breadth of services covered under certain regulated sections of industry, the Act should be amended to clarify that an industry association or body can draft a code where it represents a substantial part of a regulated section of industry.

The eSafety Commissioner should be required to undertake real and substantial consultation with industry, both before the Industry Codes are put into draft, and on any Industry Standards. This requirement should be reflected in the Act itself.

This would be particularly important should this review recommend that the eSafety Commissioner be empowered to draft Industry Codes. We note that under the UK Online Safety Act, the regulator Ofcom is responsible for drafting industry codes but undertakes both an informal consultation to seek views from industry on research, conclusions and assumptions that it has reached and uses this feedback to inform a formal statutory consultation required before the codes are formalised.

The Act should also be amended to include a more realistic minimum time frame for the development of Industry Codes. Section 141 currently requires a minimum notice period of 120 days. This time is not sufficient to support the development of quality codes that advance the objectives of the Act. We suggest a minimum period of 12 months would be more appropriate.

2) **The scope of the Industry Codes and Standards should be decoupled from the National Classification Scheme**

The National Classification Code (NCC) is not fit-for-purpose as the basis for categorising content subject to the Industry Codes/Standards. The NCC was designed to support a regime in which specific items of content are classified prior to commercial publication after being assessed individually against the NCC criteria. This requires nuanced judgement of the content item against broad standards including "the standards of morality, decency and propriety generally accepted by reasonable adults" or whether the content may be "unsuitable for a minor". This may be workable where decisions are made in respect of individual items of content, but it is not workable as the basis for a regulatory regime designed to apply to a defined content type.

Given the scale of online content, it is impractical to base a compliance regime on a classifier that requires each specific item of content to be assessed against broad criteria. The reliance on the NCC creates significant uncertainty for service providers in seeking to comply with the Industry Codes/Standards. In the context of the Phase 2 Industry Codes, for example, it will be exceptionally difficult for service providers to develop a set of defined compliance measures to address films which fall into the broad category of being "unsuitable for a minor to see".

When designing products or implementing other compliance measures, service providers need clear categories of content to which to apply compliance controls. Rather than referencing the NCC, service provider obligations should be defined with reference to clear content categories (as is the case, for example, for 'cyber-abuse material targeted at an Australian adult' in the Act). Industry has sought to address this issue to some extent by creating definitions of content under the Phase 1 Industry Codes. However, the root of the problem is the reliance on the NCC as the basis for determining service providers' obligations. The Act should be amended so that the Industry Codes/Standards and BOSE apply to clearly defined content types. For example, defined categories of content could be created to describe child sexual exploitation material, crime and violence material, pornography, drug-related material and other material.

3) **Industry Codes and Standards should not extend to certain types of content, which should instead be subject to appropriate legislative oversight**

Mandatory obligations for providers to prohibit and remove from the service class 1A and class 1B material should be limited only to circumstances where the use, storage or distribution of that material is illegal and/or where it is appropriate or proportionate to the potential harm caused to end-users.

Section 13 of the Standards impose broad and mandatory obligations for providers to prohibit (via terms of service) the use of the service to **solicit, access, distribute or store class 1A or class 1B material,** irrespective of whether the possession and/or use of the content is, in all circumstances, illegal. Section 15 (2) and section 24 (2) of the RES (but not equivalent provisions in the DIS) also impose mandatory requirements for providers to remove the material from the service (unless it is not technically feasible) and take steps to ensure that the service no longer permits access to or distribution of the material.

While it is illegal to possess and access some of the categories of content that falls within class 1A material (for example, child sexual abuse material), it may not be illegal for Australian adults to possess and privately view other class 1 material (for example, certain drug related content which falls within class 1B).

That the law makes a distinction between the private possession of content that has been Refused Classification (which is not illegal), and its sale, advertisement and distribution (which is illegal), is deliberate: it is to limit the unreasonable intrusion into the private lives of its citizens, particularly in circumstances where there is no identifiable harm to other members of society.

Similarly, there may be legitimate (or non-malicious) reasons why a user may possess class 1 material and may share that material to a limited audience using a relevant electronic service (such as an email, MMS or SMS). For example - bystander footage taken on a user's device of an extremely violent event (for example a terrorist attack or a war crime), uploaded to a user's personal end-user managed hosting service and emailed to a news organisation.

The Industry Codes and Standards should treat online content in the same way as offline content. Where the Government believes a category of content is sufficiently harmful such that even the private possession of that content should be prohibited, the Government may make that content illegal, through transparent and democratic processes and in a necessary and proportionate manner. It should not be done indirectly via Industry Codes or Standards and only applicable to online content.

Given the significant societal implications that flow from the regulation of private communications (and noting that is it inconsistent, for example, with the Telecommunications Act which restricts monitoring), we suggest that imposing obligations of this nature should be a legislative function overseen by parliamentary processes and subject to public consultation and debate.

Basic Online Safety Expectations (BOSE)

1) **The Act should be amended to include better guardrails around the exercise of the eSafety Commissioner's reporting powers**

The Act provides for the eSafety Commissioner to require providers of a social media service, relevant electronic service or designated internet service to provide reports about compliance with the Basic Online Safety Expectations.

To date, Google has received two non-periodic notices issued under s56(2) of the Act. The first notice, issued on 22 February 2023, required Google to provide detailed information in response to 43 questions relating to Child Sexual Abuse Material ("**CSAM**") across Google Drive, Google Meet, Google Chat, Google Photos, Google Messages, Gmail and YouTube. The second notice, issued on 18 March 2024, required Google to provide detailed information in response to 44 questions relating to terrorism and violent extremism content ("**TVEC**") across YouTube, Gemini and Google Drive.

The questions included in each of these notices are not generic questions seeking information on the steps Google is taking to meet a particular expectation. Instead, the questions seek detailed information on a range of specific actions, with each question linked to relevant expectations.

This approach suggests that these actions have been determined by the eSafety Commissioner to be the 'reasonable steps' service providers should be taking to meet the relevant expectations. While the BOSE Determination does include examples of 'reasonable steps' services providers could take to meet expectations, the intention of Parliament was to avoid overly prescriptive expectations to allow service providers to develop their own appropriate means of complying with them.

This is confirmed in the Explanatory Statement to the BOSE Determination, which states:

> *It is not intended that the Commissioner prescribe specific steps for service providers to take to meet the expectations. 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 of achieving the expectations.*
>
> *Notwithstanding that the Determination provides flexibility for service providers, it does outline a number of examples of reasonable steps that could be taken within the sections of the Determination. Not all reasonable steps have to be taken by all service providers. Rather, they are intended to provide guidance to service providers.*

We note also that the Explanatory Memorandum to the Act states in respect of reporting under the BOSE that 'most large companies are already producing such reports with the appropriately trained staff'. This suggests that the current approach goes well beyond what was originally envisaged.

To ensure that the eSafety Commissioner's powers are exercised in a fair and proportionate way, based on evidence and insights and recognising the importance of reducing regulatory requirements (as articulated in the eSafety Commissioner's own regulatory guidance on the BOSE), we recommend that the Act be amended to provide greater guardrails around the eSafety Commissioner's exercise of reporting powers under the BOSE scheme.

At a minimum, the Act should require the eSafety Commissioner to detail how the information sought under a BOSE Notice will demonstrate the meeting of the relevant expectation. For instance, we question whether knowing the internal names of tools used to detect TVEC images or livestreams is necessary to demonstrate Google is meeting the relevant expectation.

We also suggest that the Act should be amended to introduce articulated thresholds that must be met before the eSafety Commissioner can issue a reporting notice. The Act currently requires the eSafety Commissioner to have regard to certain factors when deciding to issue a reporting notice. This includes, for instance, the number of occasions during the previous 12 months on which complaints about material provided on the service. Google's most recently received BOSE notice referenced only four relevant complaints, none of which had been escalated to Google for review or action.

2) **The intended scope of "unlawful or harmful" material or activity under the BOSE Determination should be defined, and how that interacts with, or relates to, other regulatory regimes and initiatives should be clarified**

The existing statutory regime under the 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 specific categories of unlawful and harmful material are clearly defined by the Act, the BOSE Determination (and the amendment to the BOSE Determination) is much broader and adopts the language "*unlawful and harmful*" material and activity that is not tied or limited to those five categories.

The concept of "*unlawful and harmful*" material or activity (outside of the 6 categories above) is very broad. What may be unlawful or harmful is:

- dependent on context (a piece of content by itself may be harmful, but not if additional information or disclosures are provided);
- the nature of the service (for instance, content 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, content about wellness, diet and exercise may not alone be harmful but could be for a user who is suffering an eating disorder).

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

It is imperative that service providers know what material or activity is "unlawful and harmful" within the remit of the 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 and activity fails to provide service providers with the legal clarity they need to act.

3) Reporting regime should be subject to a robust confidentiality protocol.

We have serious concerns about the treatment of service providers' confidential information provided in response to the BOSE reporting regime and, in particular, the eSafety Commissioner's publishing of material that may materially impact service providers' ability to operate a safe and commercial online service.

In the current statutory landscape, service providers are unable to refuse to respond to a non-periodic reporting notice on the basis that its information is confidential to the service provider or to a third party. Further, despite claims of confidentiality, the publication of confidential information provided in response to a reporting notice has involved disclosure of

highly sensitive and commercial-in-confidence material, and potentially undermined service provider efforts to thwart bad actors on their services by exposing details of provider systems. A platform's ability to address online harms is often dependent on highly sensitive and confidential information remaining out of the hands of bad actors. The confidentiality of that information must be respected to avoid undermining the objects of the Act.

The eSafety Commissioner has expressed the view that the transparency and accountability objectives of the Act are most effectively met by making information received from industry in response to a reporting notice public, where appropriate. This approach seems inconsistent with the Explanatory Memorandum to the 2022 BOSE Determination, which states (emphasis added):

> *Reporting of in-confidence information*
>
> ***Where a particular service shares commercial-in-confidence features or information*** *with the Commissioner for the purposes of demonstrating compliance with the Determination,* ***this information would not normally be made public.*** *However, the Basic Online Safety Expectations are intended to enhance transparency and accountability of service providers.* ***Therefore, service providers are encouraged to make reports publicly available, or agree that the Commissioner may do so.***

Despite the sensitivities around disclosure of confidential information, there is no statutory mechanism under the Act that allows for service providers to claim confidentiality over material provided to the eSafety Commissioner. While the eSafety Commissioner has published guidance notes that state that the eSafety Commissioner will consider claims of confidentiality, in our experience, claiming confidentiality has been extremely difficult and the decision to publish confidential information remains at the sole discretion of the eSafety Commissioner.

In the absence of any firm basis or procedure through which service providers can make claims of confidentiality, service providers are left without sufficient avenues to protect their information and the safety of their online environments.

To address this, the BOSE reporting regimes should be updated to ensure information reported by service providers is subject to a robust confidentiality protocol. This should include:

- An express right for service providers to claim confidentiality, and a transparent process by which the Commissioner will assess that claim.
- If the Commissioner is considering publishing information which is the subject of a claim of confidentiality, a requirement to consult with the service provider in respect of the confidentiality of the material proposed to be published.

- If the Commissioner decides to publish information which is subject to a claim of confidentiality, a requirement to give reasons for doing so and to give service providers an opportunity to take necessary steps to protect the confidential information.
- Finally, an express right for service providers to challenge a decision by the Commissioner to publish information subject to a claim of confidentiality prior to the information being published. That challenge should be considered by a separate, independent body and information which is subject to a challenge should only be published if and until any challenge fails.

Complaints and content based removal schemes

1) **We support the operation of the complaints and notice-based removal scheme under the Act.**

The involvement of the eSafety Commissioner's case management team in processing victims' complaints allows for victims' experiences to be handled holistically. In particular, it allows for victims to receive support from other agencies, including law enforcement, which are better able to address the cause of the harm or abuse at the source. In our experience, the regime under the Act provides greater support to victims than regimes in some other jurisdictions, in which victims may be largely left to their own devices in dealing with online abuse.

Once the eSafety Commissioner has determined that content falls within scope of the Act, the notifications they submit to Google are typically comprehensive and helpful, providing the information we need to address the harm promptly. We are also able to use their notifications to help identify broader trends in abuse on our platforms. The eSafety Commissioner's case management team is easy to work with and is open to dialogue and feedback.

2) **Removal notices should include the basis for determining that content meets relevant thresholds**

While the Act sets out thresholds for each type of content subject to the notice-based removal scheme, the eSafety Commissioner often still needs to perform a detailed assessment of content to determine whether it meets the threshold. In many cases, the eSafetyCommissioner is required to make an assessment about what an 'ordinary reasonable person' would conclude. For material that is subject to the online content scheme, the eSafety Commissioner may be required to determine how the content would be classified by the Classification Board, potentially needing to balance the principle that adults should be able to read, hear, see and play what they want, with restrictions based on 'standards of morality, decency and propriety generally accepted by reasonable adults'.

Given the potential for removal notices to impact freedom of expression, when issuing a removal notice the eSafety Commissioner should be required to include the basis on which it has been determined that the content meets the relevant threshold. This would improve transparency and assist service providers in their own assessment of removal requests under the scheme.

3) **Content removals do not address underlying causes of harm**

It is important to note in this context that content removals, though important, do not address the underlying causes of harm. Focusing solely on access to harmful content does not stop it at the source. Creators of this content should be held accountable for harms caused. Harsher penalties for those responsible for posting harmful content would be a significant deterrent to behaviour that materialises on our services.

4) **Most categories of content underpinning the removal schemes are well-defined, and sufficiently broad to cover a range of harms**

We make the following broad observations on the application of the complaints and content-based removal notice schemes to particular content types:

- The threshold for '**child cyber-bullying material**' fails to clearly address seemingly innocuous content that may be harmful when included in a broader campaign of bullying. To help service providers assess the validity of a removal notice, the eSafety Commissioner should be required to provide contextual information which is relevant to identifying material as child cyber-bullying.
- It is important that the threshold for '**adult cyber-abuse'** remains high to protect freedom of expression. This is particularly the case for material relating to public figures. It is also important that damage to reputation remains excluded from the definition of adult cyber-abuse as this harm is addressed via Australia's defamation laws. We suggest that the existing adult cyber abuse scheme provides the eSafety Commissioner the ability to respond to volumetric attacks and tech-facilitated abuse and gender-based violence.
- We provide users with tools to request removal of explicit or intimate personal images from Google Search and or altered or synthetic content that mimics someone's face or voice on YouTube. Where users elect to request removal by engaging the eSafety Commissioner, in our experience the thresholds for the **image-based abuse scheme** are appropriate and useful, and sufficiently flexible to apply to the introduction of deep fake pornography.
- As noted above, the online content scheme's reliance on the NCC is unworkable. However, it is clear that **violent pornography** is a type of content that would be

covered by that scheme, on the basis that it would meet the NCC's criteria for RC or X18+ classifications for films.

- Similarly, **Social media posts boasting about crimes** would clearly be considered RC films or publications under the NCC on the basis that they would 'promote, incite or instruct in matters of crime or violence'.
- The eSafety Commissioner's power to require internet service providers to block **abhorrent violent material** is, in our view, sufficient. That power is supplemented by the provisions of the Criminal Code which can impose significant penalties on social media services, designated electronic services and hosting services which fail to 'expeditiously remove' abhorrent violent material.
- While we support the Government's objective to tackle **hate speech**, we urge the Government to address the harm caused by hate speech holistically and through broader regulation of hate speech, both on and offline. As outlined above, we are concerned at the increasingly fragmented approach to respond to harms across multiple regulatory mechanisms. We understand the Government is contemplating broader hate speech laws and encourage this review to recommend against the expansion of the Act to contemplate hate speech until the Government's intended approach to this issue is clarified.

Conclusion

We appreciate the opportunity to contribute to this review and are available to provide further information and answer any questions on these materials as required.