Microsoft

Microsoft submission to the Draft Online Safety (Basic Online Safety Expectations) Determination 2021 consultation

Introduction

Microsoft welcomes the opportunity to respond to the draft *Online Safety (Basic Online Safety Expectations) Determination 2021*. We recognize that government regulation has an important role to play in addressing digital safety risks, and we support the development of a principled and carefully calibrated regulatory framework. Our submission on the draft Determination will reference points made in our previous submissions on the Online Safety Bill.

We note that this consultation is taking place at the same time as industry associations and their members are engaging with the eSafety Commissioner on the approach for the industry codes required by the *Online Safety Act 2021 (the Act)*. To avoid any risk of duplication or inconsistency between these two evolving processes, we recommend that the Department and eSafety remain closely engaged. We also note the work concurrently underway on an age verification roadmap, which is still in early stages.

As outlined in our previous submissions on the Online Safety Bill, Microsoft considers that targeted interventions based on risk will be more impactful in preventing harm than a "one-size-fits-all" approach. Some platforms, by the nature of their service or functions, facilitate higher-risk interactions. Explicitly considering risk factors within the Determination, as well as in the industry code-making process, also accords with the Minister for Communications' stated preference for a 'risk-informed approach'. At the conclusion of parliamentary debate on the Online Safety Bill 2020, the Minister indicated that eSafety would release a position paper initially and publish its regulatory priorities annually — a process that would enable eSafety to "focus on the most harmful material as a priority". The Minister also noted that the articulation of eSafety's regulatory priorities would also "*support industry to focus its efforts towards products that have a higher chance of facilitating harm*".

Addressing implementation challenges through a risk-based approach

As a large, matrixed company with products and services across the consumer-facing top of the 'technology stack'— as well as with tools and infrastructure offerings at lower levels of the stack (or 'back-end') — Microsoft has a broad perspective on the challenges of balancing the safety, privacy, and cybersecurity needs of our global user base.

With the passage of the new *Online Safety Act 2021 (the Act)*, the scope of Australia's online safety legislation has been extended to a very broad range of digital services including email, messaging, search, and hosting, as well as specific-purpose products like games and apps that also allow communication between users, and infrastructure that enables organizations from schools to the Parliament of Australia to operate effectively and securely. Many technologies will have safety-related compliance obligations for the first time, without any roadmap or prior best practice to demonstrate compliance. As currently drafted, the proposed Determination takes a wide-ranging approach, applying a long list of requirements to a huge and diverse variety of providers and services across the technology ecosystem, including technologies which regulators may have had limited experience.

The range of individual technologies that could potentially fall within the scope of the Basic Online Safety Expectations Determination (BOSE) is also likely to create challenges for eSafety in managing the

required consultation processes and in monitoring alignment with the BOSE. The consultation process as conceived in the draft Determination presupposes the eSafety Commissioner has capacity to consult individually with each service provider on what measures will constitute "reasonable steps" across what may be a large number of products and services.

Recognising the regulatory burden across the digital economy, and potential strain on eSafety's resources, the Determination should instead focus the eSafety Commissioner's efforts on categories of services where the risks of harm are greatest, and where the best practices have been demonstrated as having effect.

Accordingly, as a starting point for implementing the BOSE, the Australian Government may wish to consider a staggered approach. We would recommend the Minister consider first prioritizing the release of a Determination only for social media services, as defined in the Act (for example, excluding online business interaction), and taking into account the provision of advertising material on the service. Separate Determinations could be developed later for relevant electronic services and designated internet services. We also recommend that the expectations in each of these three Determinations are tailored to recognise some of the differences between technologies in each of these categories. Some of the basic online safety expectations that are appropriate for a social media platform will differ, for instance, to what is appropriate in the context of an enterprise email service used by a government.

Through this process, the eSafety Commissioner would also be able to more efficiently manage the requirement that industry consult with the Commissioner with respect to the BOSE. Given the pace of technology innovation and that Australia can be used as a first-release market (providing Australian users with early access to new products and services, or to new features for existing products), this risks being an intensive process for both industry and the Commissioner. As a result, it may also be helpful to add into the Determination some further guidance and/or thresholds about when the eSafety Commissioner should be consulted by a service provider, as required by section 7(1). Introducing further clarity will help reduce the risk of either under- or over- consultation, potentially reduce compliance burdens, and help focus resources on the most impactful interventions. It may also be helpful to specify that such consultation should not require the disclosure of commercial or other sensitive information.

Enhancing definitional clarity

The absence of specific definitions for key terms used in the draft Determination creates a level of uncertainty for both providers and their users.

Providing further guidance in determining what constitutes "reasonable steps"

To further ensure the BOSE are developed and applied proportionately, we recommend adding into the Determination a specified list of factors to be considered when determining what constitutes "reasonable" steps. Industry is likely to be able to respond more directly and effectively, based on identified risk attributes that contribute to an objective determination of what are reasonable steps.

The Determination could be amended to outline a list of factors that the providers must consider when making a "reasonable steps" assessment, and that the Commissioner must consider when providing guidance. Such an objective list might include:

- Whether a service already takes voluntary actions to protect its users;
- The user-base of a service (including the average monthly number of active Australian users and the average age of those users [where known]);
- The purpose of a service;

- The functionality of a service;
- The security and privacy needs and expectations of a service;
- The relationship between the service and the end users for purposes of enforcing codes of conduct;
- The extent to which content is amplified or can go viral on a service;
- Whether a service enables interaction with or discovery of other, hitherto unknown users; and
- The number and type of removal requests issued to that service by the eSafety Commissioner, relative to other services of that type.

The discussion paper supporting this consultation notes that the BOSE are not intended to be prescriptive but to allow flexibility in the way that they are applied. It would be helpful to see explicitly in the Determination an acknowledgement that the BOSE should be applied flexibly, in ways that make sense for different services and technologies.

Creating greater certainty by using clearly defined terms

In addition to providing more detailed guidance to help providers calibrate what steps are reasonable for their service, we recommend against using undefined terms in the draft Determination. The absence of specific definitions for key terms used in the draft Determination risks creating a high level of uncertainty for both providers and their users. If retained, examples of important terms that need further clarity include "harmful" material or "activity", and "volumetric attacks". There is a risk that by using undefined terms and introducing new concepts through the draft Determination, the scope of the governing Act is significantly expanded.

In particular, the undefined and vague language of "is or may be … harmful", applied to both "material or activity", is likely to pose real challenges for service providers and their users. Undefined, material that "…[is] or may be harmful" could be an enormously wide range of content, especially when coupled with "activity", which could also be read very widely. Harm can be highly subjective and felt to varying degrees, depending on the individual. No reference is given to a specific complaint or person affected by the alleged harm, or the context in which it may occur. This stands in very stark contrast to the Act, which deals with a number of specific and defined harm types.

Rather than creating clear expectations for providers and users, the BOSE therefore risk being highly unclear and subjective, to the point of potentially being unworkable. It is likely to be challenging for providers to understand what is required to move towards these standards and, as outlined below, this uncertainty may also have human rights implications. We therefore recommend that the draft Determination focus only on the content and harm-types that have been defined in the Act.

Core and additional expectations

Requirements for proactive measures

We appreciate that the BOSE will focus on systems and processes for providers, rather than setting out "one-size-fits-all" solutions for providers to implement. That said, we are concerned about the scope of what is potentially required by the BOSE. This is exacerbated by the definitional issues outlined above.

Expectation 1(b) currently states that "[t]he provider of the service will take reasonable steps to proactively minimise the extent to which material or activity on the service is or may be unlawful or harmful". As it is currently drafted, this suggests that providers have an obligation to prevent a potentially huge range of material or activity from appearing on their service. Given the inherent subjectivity of what is "harmful", this could theoretically require providers to prevent almost all speech from appearing on their platforms. Providers are not well-placed to make judgments about what

material or activity may be subjectively harmful to individual users, let alone before that material or activity has actually manifested on the platform. Providers do not necessarily have knowledge of all the material they are hosting, given the volume at which user-generated content is created. Even where content is flagged to providers, whether that content is harmful will be subjective and context-dependent in many situations.

The expectation that providers will proactively detect or remove material or activity that *may* be harmful risks compromising the right to freedom of expression and access to information and envisages a potentially unwarranted level of intervention and responsibility on the service provider. Moreover, such interventions can have an impact on all users, most of whom will not be aiming to use a service for illegal or malicious purposes. Such "proactive policing" may not be the most effective or proportionate way to achieve the intended outcomes. Retaining this draft expectation risks providers feeling obliged to implement processes disproportionate to the risk of harm and/or that could compromise user privacy and other fundamental rights. Context, including the type of content or conduct at issue; the purpose, type, or functionality of the service on which it appears; and the degree of access to and availability of the content should all be considered when adopting a "proactive" approach. If Expectation 1(b) is retained, we strongly recommend redrafting it to focus solely on content that is unlawful.

Additional expectations

The draft Determination currently includes a list of 'additional expectations', creating a further layer of specific standards, including elements that may significantly overlap with industry codes being developed to deal with class 1 and class 2 content. In some areas, such as remedying systemic deficiencies — such as the absence of terms of use (section 14) or complaint mechanisms and contact points (sections 15 and 21) — explicitly setting additional expectations is appropriate and seems in keeping with the systems-based approach taken in the BOSE.

More problematically, some "additional expectations" and the associated examples of "reasonable steps" in the draft Determination seem prescriptive and concern certain product functions or features that may have a disproportionate impact on the rights of all users – not just those who might seek to abuse a service. These additional expectations may also present practical difficulties for industry. In particular, the elements proposed in sections 8, 9, and 10 may present challenges. If the Government sees these issues as policy priorities, it may be more appropriate to handle them through a guidance process on an ad hoc basis with individual platforms, rather than forming part of the (universally applied) Determination itself.

Section 8: Encrypted services

Specifying that processes to detect material on an encrypted service form part of a reasonable steps test is highly problematic. The issues related to encryption and detection of abusive material on encrypted services are extremely complex. Microsoft believes that governments should act with extreme caution in this area. While safety of digital services is an important societal interest, so too are security, privacy, and civil liberties related to private communications. Therefore, any regulatory efforts designed to facilitate the detection of material on encrypted services should be narrowly focused to minimize the risk of overbroad regulation and threats to privacy, security, and civil liberties.

An end-to-end encrypted (E2E) communications system can only be accessed by the intended users. If a third-party (including a cloud service provider) can access the data, it is not end-to-end encrypted. As a result, enabling the detection of materials on an E2E encrypted service is inherently inconsistent with the definition of E2E encryption. No matter how this issue is framed, any "technical solution" that is offered for access to E2E encrypted data involves some form of backdoor access. Microsoft believes that

policy makers must view such proposals considering the core problem that a backdoor for one means a backdoor for all - including rogue governments and criminal actors.

To acknowledge the importance of maintaining strong encryption while also seeking to preserve user safety, we suggest redrafting this expectation as follows: *"If the service uses encryption, the provider will take reasonable steps to develop and implement processes to detect and address material or activity on the service that is or may be unlawful or harmful, without compromising end-to-end encryption where this is enabled."*

Section 9: Anonymous accounts

Expectation 4 appears to be drafted based on an assumption that anonymous accounts present an unacceptable risk that necessitates identify verification for all services and in all circumstances. This presentation risks ignoring the many legitimate and protective benefits of anonymity for particular individuals, such as the victims of domestic violence, individuals exploring their sexuality or sensitive health topics, political dissidents, journalists, and whistle-blowers. Children and teenagers may also create user account names and email account addresses that do not reveal their true identity (or other personal information) to safeguard their privacy. Indeed, the eSafety Commissioner's own website recommends the use of anonymous accounts, under certain circumstances, such as when using dating apps or in reporting situations of domestic violence. Privacy may also be important for those affected by cyber-bullying, cyber abuse, and the sharing of non-consensual intimate images.

The need for accountability for illegal or violative actions must be balanced with the broader human right of privacy. This depends heavily on the context in which the technology is being used. Feasibility must also be taken into consideration. The deliberate provision of false user details is also not something a service provider can directly control or safeguard against; nor can it investigate or verify in all use cases. Methods do already exist to facilitate accountability for infringing account holders, such as the use of associated mobile phone numbers or email addresses, or technical solutions such as device identification. While not foolproof, these may provide a more expedited means of taking action against infringing account holders without collecting additional personal data. The *Act* also already empowers the eSafety Commissioner to obtain user account details under particular circumstances. Care must also be taken to ensure any personal details are safeguarded and that their retention does not create conflicts of law or risk misuse in countries with fewer democratic and other safeguards.

Section 10: Cooperation with other service providers

Cross-industry collaboration is an important part of how we work to ensure the safety of our users and the integrity of our services. Microsoft is a founding member of the Global Internet Forum to Counter Terrorism (GIFCT) and the Technology Coalition – two non-government organisations that facilitate cooperating on two of the most challenging online harms. Other organizations such as the Digital Trust and Safety Partnership (DTSP) are dedicated to the development of best practices in safety governance, risk assessment, and enforcement procedures. Such efforts, initiated by industry, that work at a systemic level to reduce very serious online harms, are achieving positive outcomes without the need for explicit regulatory prescriptions. We encourage other providers, where possible, to join these industry efforts to help reduce harm across the internet ecosystem.

However, we have some concerns about the examples of "reasonable steps" set forth in this draft expectation. The GIFCT, Technology Coalition, and the DTSP allow member companies to work together – but importantly, also acknowledge the differences between members. Member companies share information and cooperate but will take action on specific content and conduct according to their own policies, processes, and procedures. Among other things, this includes their respective commitments to

user privacy. The draft Expectation could be usefully qualified to note that providers are not expected to share information in a way that might compromise people's rights, including privacy or cybersecurity.

Another concern involves the provided example describing how providers may work together to prevent "high-volume, cross-platform attacks (also known as volumetric or 'pile-on' attacks)". As noted earlier, this is not a harm type that is defined in the *Act* or the BOSE, nor is it a term of art that is clearly understood by practitioners of Trust and Safety. It is also not (as we understand it), necessarily unlawful content or conduct. It is therefore difficult to suggest should be expected to work together against an ill-defined form of material/activity, that may differ in both its definition and impact across services.

Facilitating a collective response of this kind implies that providers should share/consult others about a specific person's behaviour, necessitating identification of both that individual's personal data and information about their content or conduct. This is inherently highly problematic from a user privacy standpoint. It is arguably inappropriate for private companies to follow individual users across services offered by different players in the technology industry.

Moreover, as above, members of the GIFCT, Technology Coalition and DTSP enforce their own terms of service, community standards or other policies in response to violating content. Providers should be expected to do so - but with respect only to content or conduct appearing on their own platform or service and in line with their own policies. To do otherwise would risk disproportionately impacting user rights. It would also make it almost impossible for a platform or service to provide clarity for people using that service about the types of activity that are prohibited and that could jeopardize that person's ability to continue using these services. This cuts against basic concepts of fairness and transparency – which are all the more critical in the context of online services that affect basic aspects of life. We, therefore, strongly recommend that this example is removed from the draft expectation.

Expectations regarding record-keeping

The reporting requirements at the Commissioner's disposal are among the most important and reflect a welcome step towards transparency, although the expectation that records should be kept for five years may be an issue depending on the type of data retained and other data privacy requirements. We, therefore, recommend giving further consideration to defining precisely what records should be kept. In keeping with good privacy practices, providers should retain only necessary data, for the shortest time possible.

Conclusion

Microsoft thanks the Department for the opportunity to provide comment on the draft Determination. We are available to meet with officials to discuss these points, as required.