Meta

Response to the Basic Online Safety Expectations

NOVEMBER 2021

Executive summary

Meta welcomes the opportunity to contribute to the Australian Government's consultation on a set of Basic Online Safety Expectations (the BOSE), as part of the regulatory structures underpinning the new Online Safety Act.

We have supported the enhancement of Australia's online safety laws via the Online Safety Act. Meta has been calling for new rules for the internet - including content regulation - around the world for many years.[1] We have appreciated the continued close working relationship with the Office of the eSafety Commissioner in preparation for the legislation to take force in 2022.

We see the BOSE as a novel way to encourage greater progress by industry in protecting users' safety online. The BOSE could be an innovative regulatory instrument that complements the strict requirements in the legislation itself and underlying codes, by providing transparency of platforms' efforts to work towards specified objectives, rather than 'black letter' requirements that necessitate strict compliance. This flexibility allows for the BOSE to cover online safety problems that are still evolving, where solutions are being developed, or where platforms need to take different approaches - but the Australian Government would like to establish an impetus for greater progress.

We share the Australian Government's desire to see more collective action by industry, governments, NGOs and the broader community on many of the areas nominated in the BOSE, including more work to encourage age-appropriate experiences for teens and young people online, and greater protections for public figures who may experience mass harassment. Our submission contains suggestions to make the BOSE clearer and more workable for service providers, while retaining broad flexibility and working towards the Government's stated intentions for the regulation.

While we share the Government's ambition for greater progress on online safety issues, the current version of the BOSE sets very high expectations for significant progress in a short period of time. It would not be practicable to expect that all of industry can comply with all of these expectations within two months. Some of the requirements relate to complex problems where solutions are not yet clear and different members of industry will have different views on the best approach. We would suggest the Government's expectations in some areas are better described as 'advanced' rather than 'basic'.

[1] M Bickert, *Charting a way forward on content regulation*, Meta Newsroom, February 2020, https://about.fb.com/news/2020/02/online-content-regulation/ .

Given the complexity of some of the areas identified, we expect the regulatory impact of the BOSE would be exponentially greater than the Government's previous estimates.[2]

Our suggestion is that the Government can retain this level of ambition, but classify some expectations as 'advanced', requiring providers to report on the work they are doing against advanced expectations, but take a more practical compliance approach and recognise that platforms are not in a position to meet these expectations from January 2022.

Although the BOSE are voluntary, we have provided some more detailed feedback, on the assumption that platforms like ours will work hard to meet the letter of the BOSE to the extent possible. The definition of what "reasonable steps" platforms should take is central to understanding how they can meet the expectations in the BOSE. The current drafting vests significant discretion in the eSafety Commissioner in deciding what constitutes "reasonable steps" on a platform-by-platform basis, including by explicitly stating that platforms must consult with the Commissioner on a bilateral basis, on what constitutes 'reasonable steps' for their service (s7(1)).

While we appreciate our constructive working relationship with the current Commissioner, regulatory requirements that are contingent on company-specific guidance provided in private meetings risk setting expectations that are not transparent, fair, or consistent across similar services. We recommend instead that publicly-available regulatory guidance would be a more effective way to advise companies on how to interpret the BOSE for their services. The Frequently Asked Questions that the Department has issued in relation to the BOSE are helpful and instructive[3], because they also outline what the BOSE does *not* require (like scanning of private messages, or collection of authoritative identity). We suggest this document could form the basis of regulatory guidance related to the BOSE.

Our submission contains a series of constructive suggestions about the detailed components of the BOSE. Some of these suggestions include:

- defining and clarifying the scope of content that is caught by the BOSE, which currently applies to any online content that *may be* harmful (in addition to content that may be unlawful). The use of the phrase "may be harmful" in the definition

[2] Australian Government Department of Infrastructure, Transport, Regional Development and Communications, *Online Safety Reform Regulation Impact Statement*, https://obpr.pmc.gov.au/sites/default/files/posts/2021/03/online_safety_reforms_-_ris.pdf.
[3] Australian Government Department of Infrastructure, Transport, Regional Development and Communications, *Frequently Asked Questions - Basic Online Safety Expectations* https://www.infrastructure.gov.au/sites/default/files/documents/frequently-asked-questions--basic-online-safety-expectations.pdf.

creates a scope for the BOSE that is uncertain, much broader than that set in the legislation and could potentially capture a range of innocuous, lawful content that could be argued to have some risk of possible harm. An excessively broad or unclear definition poses risks of over-enforcement or the potential that the regulator could expect companies to be taking action against beneficial online speech (such as political speech), which may be out of step with international best practice standards on content regulation.

- adjusting the focus of expectations around mass harassment and “volumetric attacks” to focus more on the steps companies are taking to prevent mass harassment on their own platforms. We suggest establishing a working group between eSafety and industry to more precisely define “volumetric attacks” and explore possible areas of cross-industry collaboration.
- amending requirements around record-keeping to more proportionately applying only to “material complaints”.

We welcome the opportunity to discuss the BOSE with the Australian Government, and are ready to assist in providing any additional contributions that would assist.

Full list of recommendations

We make the following suggestions about amendments to the BOSE that could make them clearer and more practical while retaining the government's stated intention and level of ambition:

1. The existing elements of the BOSE could be classified as either 'basic' or 'advanced', with expectations that digital platforms will be able to comply with all 'basic' expectations by January 2022 and allowing a longer time frame to take steps to meet expectations that are 'advanced'.

2. The definition of 'reasonable steps' that a digital platform can take should be determined by transparent and publicly-available regulatory guidance, rather than via bilateral conversations between companies and the Commissioner (as per current s7(1)).

3. We suggest incorporating some of the language from the Government's Frequently Asked Questions into the BOSE, including:
   - Inserting in s8, for the avoidance of doubt, that this section does not constitute an expectation for digital platforms to monitor private communications or otherwise engage in practices that compromise the integrity of end-to-end encrypted communications.
   - Inserting in s9, for the avoidance of doubt, that this section does not constitute an expectation for digital platforms to be collecting and verifying the real identity of all users.

4. We suggest that sections 6, 8, 9 and 10 should clarify the scope of content to be caught under the BOSE (currently any content that *is or may be unlawful or harmful*) to either: content that *is* unlawful or harmful; or alternatively by drafting a new and clear definition to give certainty to businesses and the community about the regulation's scope.

5. We suggest amending section 10(2a) to require companies to report on the steps they are taking on their own platforms to combat the issue of mass harassment, in the first instance. At the same time, eSafety and interested digital platforms should establish a working group to more precisely define "volumetric attacks" and identify possible areas for collaboration that could be inserted into the BOSE in future.

6. We suggest that record-keeping obligations (s 19) should be amended to be limited to “material complaints” rather than all complaints and reports.

Table of contents

Overarching comments on the BOSE

Meta has supported the enhancement of Australia's online safety laws via the Online Safety Act. We have been calling for new rules for the internet - including content regulation - around the world for many years.[4] We have appreciated the continued close working relationship with the Office of the eSafety Commissioner in preparation for the legislation to take force in 2022.

We see the BOSE as a novel way to encourage greater progress by industry in protecting users' safety online. We support the voluntary and flexible nature of the BOSE as a regulatory instrument, and commend the Government on incorporating feedback from earlier consultations to ensure the BOSE is flexible and future-proof.

In this way, the BOSE plays an important role complementing the strict requirements in the legislation itself and underlying codes. While other regulations are 'black letter' requirements that necessitate strict compliance, the BOSE's voluntary and flexible nature means it can be more innovative and future-looking. The Government can nominate online safety problems where it would like an impetus for greater progress, but where the best solutions might be contested or where industry is still working through the most practical approach.

In order to ensure the BOSE are effective, we provide some overarching comments below about how to ensure they are clear and workable for digital platforms. The Australian community are best served by a set of BOSE that are clear and workable, because it ensures digital platforms and the Government are working towards common objectives that are well-understood.

Our overarching comments focus on three aspects that could make the BOSE clearer and more workable:

1. Ensuring the Australian Government is well-informed about the *required work* to comply with the BOSE
2. Discussing the *timing of commencement*, and
3. The *transparency and consistency* of the BOSE.

This is followed by more specific comments relating to the scope of the BOSE.

[4] M Bickert, *Charting a way forward on content regulation*, Meta Newsroom, February 2020, https://about.fb.com/news/2020/02/online-content-regulation/ .

Required work to comply with the BOSE

We anticipate it will be very complex and challenging for service providers (including digital platforms, messaging services, and all websites) to comply with many provisions of the BOSE. We would suggest the Government's expectations in some areas are better described as 'advanced' rather than 'basic'.

The BOSE include requirements that relate to complex problems where solutions are not yet clear and different members of industry will have different views on the best approach. To take a few examples:

- The BOSE sets an expectation that digital platforms will work together to detect and notify each other when users are experiencing a 'volumetric attack'. The definition of volumetric attack is not clear nor simple to define: for example, what percentage or volume is large enough to be considered volumetric; how to distinguish between 'attacks' versus situations where comments could be variously positive, negative and neutral; do volumetric attacks need to be coordinated or would platforms be expected to take steps when a large volume of uncoordinated users interact with a single user organically; how to map the difference in products, policies and mitigation measures between different platforms; and how to distinguish between 'attacks' and instances where users may receive negativity or criticism that is legitimate. To our knowledge, *no* digital platform currently automatically notifies other platforms of 'volumetric attacks' on their services.

- What types of safety review systems would be considered adequate, given the ongoing and continuous assessments digital platforms make in relation to the safety of their services.

- The BOSE includes a general expectation that all services will take reasonable steps to proactively minimise the extent to which material or activity on the service is or may be unlawful or harmful. However, it then includes an additional, specific obligation on encrypted services to detect and address such material (or accounts that could be spreading this material). It is not clear why encrypted services have been singled out in this way - with *additional* expectations beyond other services. While we are increasingly trialling and deploying safety mitigations on end-to-end encrypted services like WhatsApp, the online industry as a whole is at a relatively early stage in considering what mitigations are appropriate - and this is a question that is hotly contested between different digital platforms, civic society, and various governments around the world.

Given the BOSE set expectations in some areas that exceed all current industry practices, every single company has work to do. It is highly challenging to stand up new cross-industry processes in such a short period of time: a reasonable runway is necessary to develop policies and protocols, execute responsibly, and avoid the risk of inadvertent consequences. The Government's estimation of the regulatory impact of the BOSE does not account for this complex and significant amount of work.[5]

Even excluding the work required to change company-wide safety processes to respect the BOSE, the estimates of regulatory impact also underestimate the burden imposed on industry in relation to the two requirements that do incur strict compliance from platforms: development of reports explaining how a platform respects the BOSE; and responding to requests for information (RFIs) from the eSafety Commissioner. The Regulatory Impact Statement estimates regulatory costs of $178,000 annually (of which only $20,000 would be borne by large businesses like Meta). This is informed by an estimate that only 22.5 hours of staff time would be required to prepare a transparency report and respond to RFIs from the Commissioner's Office.

Given our experiences of preparing transparency reports for other Australian Government processes (such as the voluntary industry code on misinformation and disinformation, or the taskforce on terrorist content online) and our experiences of responding to questions from the eSafety Commissioner, we anticipate the regulatory impact would be exponentially greater than the Government's estimate. Meta was not consulted in the preparation of the Regulatory Impact Statement for the online safety legislation.

We are not suggesting that the regulatory impact under the BOSE is unjustified, but we raise these considerations to ensure the Australian Government understands the size of the task ahead. While we understand the BOSE represents the Government's expectations about digital platforms' practices, we should be clear that some of these provisions require significant work.

In order to help ensure the BOSE can remain a flexible regulatory instrument, and the Government can use it to signal areas where they would like to see further work even if industry is not yet in a position to comply with strict 'black letter' requirements, we recommend amending the design of the BOSE to separate out 'basic' versus 'advanced'

[5] Australian Government Department of Infrastructure, Transport, Regional Development and Communications, *Online Safety Reform Regulation Impact Statement*, https://obpr.pmc.gov.au/sites/default/files/posts/2021/03/online_safety_reforms_-_ris.pdf

expectations. In this way, the Government can clearly communicate those expectations where it understandably expects that digital platforms will comply immediately (such as having policies and procedures relating to safety) versus those expectations where further work is required. This would allow the Government to retain a high level of ambition and incentivise industry to work towards best practices but ensure the BOSE is practical and workable. We would still be happy to report against advanced expectations, but would appreciate an acknowledgement from the government that these are more challenging and aspirational areas that require more time to work on - and the eSafety Commissioner will take this into account before 'naming and shaming' certain providers.

The remainder of our submission should assist in illuminating the expectations that we consider to be 'basic' versus those that are 'advanced'.

Timing

While we share the Government's ambition for greater progress on online safety issues, the current version of the BOSE sets very high expectations for significant progress in a short period of time. It would not be practicable to expect that all of industry can comply with all of these expectations within two months.

While recognising the Government's desire for urgent action on online safety, one alternative could be - if the Government were minded to adopt our recommendation of classifying different expectations as either 'basic' or 'advanced' - to require compliance with all basic obligations from January 2022 but allowing a longer period of time for complying with advanced expectations.

Transparency and consistency of the BOSE

Although the BOSE are voluntary, we have provided some more detailed feedback, on the assumption that platforms like ours will work hard to meet the letter of the BOSE to the extent possible.The definition of what "reasonable steps" platforms should take is central to understanding how we can meet the expectations in the BOSE, while recognising that this is context-specific and should take into account the nature of the service, technical limitations, and other factors. The current drafting vests significant discretion in the eSafety Commissioner in deciding what constitutes "reasonable steps" on a platform-by-platform basis, including by explicitly stating that platforms must consult with the Commissioner, on a bilateral basis, on what constitutes 'reasonable steps' for their service (s7(1)).

While we appreciate the constructive working relationship we have with the current Commissioner, regulatory requirements that are contingent on company-specific guidance provided in private meetings risk setting expectations that are not transparent, fair, or consistent across similar services. We recommend instead that publicly-available regulatory guidance would be a more effective way to advise companies on how to interpret the BOSE for their services. The Frequently Asked Questions (FAQs) that the Department has issued in relation to the BOSE are very helpful and instructive[6], because they also outline what the BOSE does *not* require (like scanning of private messages, or collection of authoritative identity). We suggest this document could form the basis of regulatory guidance related to the BOSE. We also welcome the statement in the FAQs that the regulatory guidance will be based on evidence and industry consultation.

Specific comments on provisions of the BOSE

Scope of content to be captured under the BOSE

It is important to recognise that content caught under the BOSE is much broader than content covered under the legislation. It applies not just to online content that is unlawful, or even content that is harmful, but content that *may be* unlawful or *harmful* (sections 6, 8, 9 and 10).

While we understand the Government intends for the BOSE to have a scope broader than the legislation, tethering the regulation to such a broad and undefined definition creates uncertainty for businesses in knowing what type of content is within scope. Hate speech, misinformation, spam, IP infringement and even political advertising could all arguably fall within this definition - as well as swathes of innocuous, lawful content that could be argued to have some risk of possible harm (for example, truthful but negative business reviews).

It appears that this definition goes beyond even the Government's intention of a broad scope for the BOSE. The lack of definition also creates uncertainty for businesses in knowing what steps they should take to meet the expectations, and uncertainty for the community in knowing what they can expect from this regulation.

[6] Australian Government Department of Infrastructure, Transport, Regional Development and Communications, *Frequently Asked Questions - Basic Online Safety Expectations* https://www.infrastructure.gov.au/sites/default/files/documents/frequently-asked-questions--basic-online-safety-expectations.pdf.

Greater clarity and certainty could be brought to the BOSE by limiting its scope to content that *is* illegal or is seriously harmful to a person's physical, emotional or mental wellbeing (similar to the Online Safety Act itself), or alternatively by developing a new and clear definition of the broader suite of content intended to be caught by the BOSE.

Encryption

The BOSE have a broad remit. They set the same requirements and expectations for social media services as for "relevant electronic services" (including private messaging).

As we have outlined in previous submissions, we do not support applying the same safety regulatory schemes to private messaging as social media. Regulations that require the detection and removal of content are not suitable for private messaging, due to the technical limitations and different expectations of users. Human relationships can be very complex. Private messaging could involve interactions that are highly nuanced and context-dependent and could be misinterpreted as bullying, like a group of friends sharing an in-joke, or an argument between adults currently or formerly in a romantic relationship. It is not clear that private companies continuously monitoring private conversations (or government regulation requiring monitoring of these conversations) is warranted, given there are already measures to protect against when these conversations become abusive including those that users themselves can take (for example, reporting and blocking the offending user).

Moreover, the BOSE not only apply the same expectations to private messaging as social media; they set *an additional* expectation for private messaging services if they are end-to-end encrypted. If the service is encrypted, providers are required to "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" (s8). It is not clear why the existing s7 is not adequate for encrypted services, given it applies to all other social media services, messaging apps, and websites.

We appreciate the Government's clarification that the eSafety Commissioner will not interpret this as requiring companies to monitor private correspondence and the BOSE are intended to instead refer to detecting abuse via behavioural signals (the approach taken by WhatsApp to child sexual abuse material (CSAM), for example).[7]

[7] Australian Government Department of Infrastructure, Transport, Regional Development and Communications, *Frequently Asked Questions - Basic Online Safety Expectations* https://www.infrastructure.gov.au/sites/default/files/documents/frequently-asked-questions--basic-online-safety-expectations.pdf.

However, the wording of the BOSE is so broad and vague that a future Commissioner could interpret the BOSE this way. In order to assure Australians that the BOSE cannot be used in this way, we recommend that the Government insert a new component, consistent with the guidance in the FAQ, in section 8 to indicate the BOSE cannot be interpreted to mean monitoring of private communications, or requiring companies to proactively scan content on end-to-end encrypted services.

Online anonymity

It's important to separate out two related but distinct questions that are often conflated: (1) whether digital platforms should allow users to be anonymous or pseudonymous in their interactions with others; (2) whether digital platforms should collect information to allow for identification of users for integrity or safety reasons.

We recognise there can be value in requiring users to authentically represent who they are: for example, we have long had an authentic name policy on Facebook which asks users to use the name that they use in everyday life (noting this might not necessarily be their legal name). We believe that, on a service like Facebook where people make direct connections with each other, our community is safer and more accountable when people stand behind their opinions and actions.

We also invest significantly in detecting and removing fake accounts on Facebook. We removed 1.8 billion accounts in the last quarter alone.

However, it may be appropriate for different services to take a different approach: for example, Instagram allows users to be pseudonymous. That's because we believe a service like Facebook with a 'friends' model of connecting people should operate differently to a service with a 'follower' model of connecting people.

Anonymity and pseudonymity play a vital role online.

They enable people to be more open, including in providing support to others on sensitive topics like mental health, addiction issues, or gender transitions - matters where people do not always want to be publicly identifiable. It is also an important part of an open, democratic society that people are able to publicly criticise their elected officials, and anonymity allows them to do so without fear of reprisal by those in power. And there may be legitimate instances where someone does not want to be easily identifiable online: for example, domestic violence survivors have legitimate reasons to mask their real name.

We firmly believe in the principle of allowing people to operate anonymously or pseudonymously online. We also believe taking anonymity away from Australian internet users would not solve issues of bullying, harassment or hate speech: notwithstanding our authentic name policy on Facebook, users continue to bully or harass others. A 2016 German study found that non-anonymous users on a review site were more aggressive than anonymous users.[8]

However, that does not mean that we believe users should be beyond the reach of law enforcement, safety regulators like the eSafety Commissioner, or courts considering a defamation action. When we receive a valid legal request from these Australian authorities, we will generally provide certain information regarding the subscriber of the relevant account to the authority.

eSafety has long held powers to request this information, and we have cooperated with these requests when we began receiving them from eSafety in 2021. In the context of defamation reform, we have also suggested that Australian policymakers consider processes like preliminary discovery orders (similar to the Norwich Pharmacal orders process in the United Kingdom), to connect a complainant with the originator of potentially defamatory material online, whilst still ensuring due process and judicial oversight.

In some instances of safety and integrity, we take steps ourselves (independent of law enforcement or regulators) to verify that a user is not operating a fake account.

The most important consideration when designing regulatory obligations around anonymity is that any instances when platforms are required to authenticate a user are proportionate, risk-based, contain due process requirements with checks and balances, and respect the privacy principle of data minimisation.

In order to provide certainty to the community about the operation of the BOSE, we recommend that the Government's intention, as stated in the Frequently Asked Questions document, should be reflected in the drafting of the instrument. We recommend inserting in s9, for the avoidance of doubt, that this section does not constitute an expectation for digital platforms to be collecting and verifying the real identity of all users.

[8] K Rost, L Staehl and B Frey, 'Digital social norm enforcement: online firestorms on social media', *Plos* One, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4912099/

Cooperation between platforms

Digital platforms are increasingly taking steps in order to build out structures that enable cross-industry collaboration.

- The Global Internet Forum to Counter Terrorism (GIFCT) was transitioned to an independent organisation in 2020. The GIFCT has a number of initiatives to encourage coordination across its members, including the Content Incident Protocol (for instances where a platform experiences an incident, there are clear processes to trigger coordination across the industry to limit the potential for cross-platform abuse) and the Hash Sharing Database.
- In 2020, the Tech Coalition announced Project Protect, a renewed investment and ongoing commitment across industry to combat child sexual abuse material. This is in addition to existing cooperation with the National Center for Missing and Exploited Children, and the technology that companies have developed and shared on an open-source basis to assist others with detecting CSAM (for example, the PDQ + TMK +PDQF technology developed by Meta).
- The Digital Trust and Safety Partnership is a new and first-of-its-kind partnership between leading technology companies, which has set out a series of principles to promote a safer and more trustworthy internet.
- Locally, the industry association DIGI has built out structures for digital platforms to collaboratively develop industry codes, such as the voluntary industry code on misinformation and disinformation.

We agree that there should be increased cooperation and work across industry, especially in relation to areas (like child sexual abuse material) where companies have shared objectives and relatively similar policies and approaches.

Cross-industry cooperation becomes more challenging, however, as these models are broadened out to a larger set of companies and issues. Companies think about issues differently - especially when they touch on complex and contested topics like how to treat political commentary. One of these complex issues relates to how to treat mass harassment or, as the BOSE describes them, “volumetric attacks”.

Meta has been doing a significant amount of work in order to develop policies and tools to help combat mass harassment on Facebook and Instagram. For example, in October 2021, we strengthened our policies to remove content that targets individuals at heightened risk of offline harm (for example, victims of violent tragedies or government dissidents) if it is *coordinated* - even if the content itself would not otherwise violate our policies.[9] We have also developed products, such as Limits on Instagram, which allows users with a single click to prevent accounts that do not follow them, or have only recently followed them, from interacting with them via comments or DMs.[10]

However, the BOSE sets an expectation that companies “work with other providers to detect high volume, cross-platform attacks (also known as volumetric or ‘pile-on’ attacks)” (s10(2a)). While we are open to exploring opportunities for greater collaboration between companies, the issue of mass harassment requires much greater consideration and debate before the BOSE could ask companies to collaborate to detect “cross-platform volumetric attacks”.

There are a number of fundamental issues yet to be resolved:

- **The precise definition of “volumetric attacks”.** While we remove content that violates or policies, or coordinated attacks that target individuals at heightened risk of offline harm (even if the content would otherwise not violate), the only example of a “volumetric attacks” that the eSafety Commissioner has raised with us in Australia was an instance that (1) was not coordinated; and (2) included commentary that was negative (including political commentary) but did not rise to the level of bullying or harassment under our policies.

  We have concerns that defining a definition this broad in regulation could, in practice, essentially be used to shield any politician or public figure from any criticism, even if legitimate.

- **Distinguishing between “attacks” and other instances of high-volume comments.** Many individuals may find themselves the subject of a high-volume of comments on the internet in a short space of time, and this could be for a wide range of reasons. They could be associated with a high-profile event (such as a sporting match or concert), an inadvertent video or meme that suddenly attracts popularity online, or because they are the target of an advocacy campaign (recent examples include campaigns around climate change or raising awareness of human rights violations).

[9] We also announced changes to prohibit an increased number of degrading or sexualised attacks, if they are directed at a public figure. A Davis, ‘Advancing Our Policies On Online Bullying and Harassment’, *Meta Newsroom*, 13 October 2021, https://about.fb.com/news/2021/10/advancing-online-bullying-harassment-policies/.

[10] A Mosseri, ‘Introducing new ways to protect our community from abuse’, *Instagram Blog*, 10 August 2021, https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse.

These interactions will often involve a mix of positive and negative interactions. A sudden increase in the number of interactions alone is unlikely to be sufficient information for a digital platform to distinguish between whether an individual is facing a “volumetric attack” (which, under the Commissioner’s definition, need not be coordinated) or a positive event which they are comfortable with. A better approach is to put tools (such as Limits on Instagram) in the hands of the individual who can make decisions about whether they would like to receive this level of new attention or not.

- **Differences in bullying and harassment policies across companies.** Unlike CSAM or terrorist content, companies take different approaches to developing policies on bullying and harassment content. While Facebook and Instagram have detailed policies that we are regularly updating and strengthening, some other platforms have not taken the same approach to safety. In the instance of a “volumetric attack” that the eSafety Commissioner raised with us, an individual was receiving harassment on another platform and, when that was not adequately policed on that service, it spilled over onto Instagram.

  Establishing channels to coordinate across companies will not solve this problem. Companies should be scrutinised and held to account for their bullying and harassment policies in the first instance, which may negate the need for new cross-industry structures.

- **Compliance with privacy and other legal obligations.** The BOSE seem to set an expectation that companies would be detecting and notifying each other of an individual’s activity on their service - without their prior consent. There are also legal limitations on a company’s ability to share information about its users with other companies, which will impact the degree of possible collaboration.

While we share the Commissioner’s desire to see progress across the industry on this issue, the current wording of the BOSE sets an expectation for companies to establish new, cross-industry structures relating to mass harassment - while a number of fundamental issues need to be worked through. The eSafety Commissioner indicated to us in correspondence in August 2020 that she would be convening a cross-industry workshop to discuss possible collaboration in this area, but this never occurred.

We recommend that the eSafety Commissioner proceed with cross-industry workshops to work through some of the fundamental considerations of different platforms. These could be held regularly to facilitate ongoing dialogue on emerging threats. In the meantime, the relevant provision of the BOSE (s10(2a)) should be re-worded to require companies to report on the steps they are taking on their own platforms to combat the issue of mass harassment.

Defaults

The BOSE indicates an expectation that, for users that are children, privacy and safety settings should be set to the most restrictive options by default (s6(3b)).

While we agree that there should be robust default settings for young users on our services, it will be overly cumbersome to require all users to necessarily have their defaults set to the *most* restrictive standards, because we built even stronger controls for certain use cases that will not be appropriate to everyone. For example:

- The default for young users on Facebook is that their posts are 'private' (ie. shared only with their family and friends). There is however an even stronger privacy session ('only me', sharing posts only with the user themself) for circumstances where users might want to store photos on Facebook but not share them. Given users generally join Facebook to connect with their friends and family, it would be cumbersome to require all young users to begin with the strongest settings built for niche use cases.
- Similarly, Instagram has recently announced sensitive content controls. While young users are automatically defaulted into settings that limit their exposure to offensive or upsetting content, Instagram now allows users to strengthen the settings further and limit even more.[11]

Defaults should establish the settings that the majority of users expect, rather than setting the standard for niche use cases across all potential use of the service. While we agree the BOSE should require robust defaults for young users - and these defaults should be stronger than for older users - we recommend removing the reference to the "most restrictive level".

Record-keeping obligations

The BOSE also introduces a requirement for service providers to keep records of reports and complaints for 5 years. There is no definition of 'report' or 'complaint'. The scope of information potentially captured by this requirement is therefore wide and not necessarily limited to material reports or complaints. For example, one possible

[11] Instagram Blog, 'Introducing sensitive content control', *Instagram Blog*, 20 July 2021, https://about.instagram.com/blog/announcements/introducing-sensitive-content-control.

interpretation of this requirement would require a service provider to retain records of every single accidental, duplicative or erroneous in-app report filed for 5 years. Users frequently report content, not because it violates a policy, but because they simply don’t like it: examples like users reporting content from football teams playing their own team, or from pop stars who they don’t enjoy.

While we understand the rationale for such a requirement to provide transparency, we suggest there are ways this obligation could be re-worked to present a more manageable obligation. For example, it could set an expectation that companies regularly report publicly via transparency reports on metrics related to harmful content. Or it could be re-worked to be limited to material complaints (e.g. complaints that require investigation or some other action by the service provider). Otherwise, this provision could represent a potentially unmanageable record-keeping obligation on companies that are captured.