Submission to Consultation on the Draft *Online Safety (Basic Online Safety Expectations) Determination 2021*

International Justice Mission (IJM) welcomes this opportunity to comment on the draft *Online Safety (Basic Online Safety Expectations) Determination 2021*. Our comments should be read in conjunction with our previous Submission to the Consultation on the Exposure Draft *Online Safety Bill 2020*.

We applaud the Australian Government's continued efforts to protect children and adults from online abuse and harms, through the passage of the *Online Safety Act 2021*. We fully agree with the key principle of the Act which states that the rules and protections we enjoy offline should also apply online and commend the Act's recognition of the instrumental role that tech companies must play to prevent their platforms and services from being used to cause online harm to people. We encourage the Government to implement clear guidelines to ensure that tech companies have appropriate systems and processes in place to improve the safety of children and adults who use or are abused on their platforms.

Background:

IJM's recommendations for the BOSE relate to measures that incentivise tech companies to prevent, detect and disrupt online sexual exploitation of children taking place on their platforms and services.

Since 2011, IJM has focused on a particular form of online sexual exploitation of children - the trafficking of children by adults to create new child sexual exploitation material, including via livestream video. IJM considers this trafficking form of online sexual exploitation of children as one of the most serious and devastating forms of abuse. The crime involves the actual livestreamed sexual exploitation of a child, often pre-pubescent, by a trusted individual in real time as directed and paid for by a sex offender, often in another country. Many Australians are involved in perpetrating livestreamed child sexual abuse.[1]

This form of abuse is complex because it allows offenders to engage in child sexual abuse production in real-time while leaving little evidence. Detection of this type of online abuse is critical because the victims – often young children – are being repeatedly abused "live"; however, because of the ephemeral nature of a livestream, detection methods in common use do not typically recognise livestreamed sexual abuse of children. The number of reported incidents involving livestreamed child sexual abuse has steadily increased in recent years and has further intensified during the COVID-19 pandemic.[2]

Strengthening the BOSE provisions:

We are generally in agreement with the provisions in the Determination setting out reasonable steps that digital service providers could take in meeting each of the core expectations and additional expectations. We note, however, that the provisions tend to focus on *safety of end-users*. Unfortunately, many children are subjected to devastating abuse and harm through the use of online platforms at the hands of adult end-users – without being users of the platform themselves.

[1] For example, a study by the Australian Institute of Criminology found that 256 Australians spent more than $1.3 million over 13 years to commission and watch livestreamed sexual abuse of Filipino children.
[2] EUROPOL, Serious and Organized Crime Threat Assessment 2021, p. 41.

In IJM's study, *Online Sexual Exploitation of Children in the Philippines: Analysis and Recommendations for Governments, Industry and Civil Society* (May 2020), of the 92 case files of livestreamed sexual abuse of children reviewed, none of the victims were themselves end-users. They had, however, suffered horrific sexual abuse at the hands of adults, with their abuse livestreamed via online platforms to other adult end-users. The median age of the child victims was 11 years old, with the youngest being 3 months old – clearly too young to be an end-user of an online platform.

The tech industry's responsibility in terms of preventing and mitigating the harmful effects of the use of their platforms should extend to protecting *all* vulnerable people, especially children, from online harms via the use of their platforms - end-users and non-end users alike. The online service providers' responsibility should also extend to restraining adult end-users from using those platforms to cause online harm.

In this light, IJM encourages the Government to provide clear guidance to the tech industry on expectations, and reasonable steps to achieve those expectations, to ensure:

1) Safety of end-users (both adult and children);
2) Protection of children who may not themselves be users of the platform or service but who may be exploited by end-users; and
3) Identification and restraining of Australians who pose an online threat to vulnerable children around the world.

Our specific recommendations, below, are measures that encourage tech companies to proactively detect and disrupt first-generation child sexual exploitation material (CSEM) and livestreamed CSEM, including:

a) deployment on their platforms specific tools informed by indicators of the production of first generation and live-streamed CSEM;
b) cross-sector collaboration to share data of suspicious activity in order to quickly surface the worst offending and violations of terms of service; and
c) development and implementation of Artificial Intelligence (AI) tools to detect harm while it is happening.

We also provide industry examples of technology and tools that are being deployed in each of these areas.

Recommendations:

We recommend amending the wording of the draft BOSE provisions, as indicated with yellow highlighting. An explanation for each amendment is provided in an endnote.

Division 2—Expectations regarding safe use

6 Expectations—provider will take reasonable steps to ensure safe use

Core expectation

(1) The provider of the service will take reasonable steps to ensure that end-users are able to use the service in a safe manner.

Additional expectation

(2) The 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.

Reasonable steps that could be taken

(3) Without limiting subsection (1) or (2), reasonable steps for the purposes of this section could include the following:

(a) developing and implementing processes to detect, block,[1] moderate, report and remove (as applicable) material or activity on the service that is or may be unlawful or harmful;

(b) if a service or a component of a service (such as an online app or game) is targeted at, or being used by, children (the children's service)—ensuring that the default privacy and safety settings of the children's service are robust and set to the most restrictive level;

(c) ensuring that persons who are engaged in providing the service, such as the provider's employees or contractors, are trained in, and are expected to implement and promote, online safety;

(d) continually improving technology and practices relating to the safety of end-users;

(e) continually improving technology and practices relating to the safety of populations who are vulnerable to abuse by end-users (e.g. children who are too young to use social media platforms but may be subjected to abuse by end-users on these platforms including via livestreaming);[2]

(f) ensuring that assessments of safety risks and impacts are undertaken, and safety review processes are implemented, including consideration of on-device options[3], throughout the design, development, deployment and post-deployment stages for the service.

7 Expectations—provider will consult with Commissioner and refer to Commissioner's guidance in determining reasonable steps to ensure safe use

Core expectation

(1) In determining what are reasonable steps for the purposes of subsection 6(1), the provider of the service will consult the Commissioner.

Additional expectation

(2) In addition, in determining what are reasonable steps for the purposes of subsection 6(1), the provider of the service will have regard to any relevant guidance material[4] made available by the Commissioner

8 Additional expectation—provider will take reasonable steps regarding encrypted services

(1) If the service uses encryption, the provider of the service 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.

Reasonable steps that could be taken

(2) Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following: [5]

(a) Implementing on-device artificial intelligence and machine learning tools targeted to detect and address material or activity on the service that may be unlawful or harmful

(b) Implementing on-device detection capabilities to prevent and disrupt unlawful material before it enters encrypted channels

(c) Engaging in secure cross-sector data collaboration to detect and address material or activity on the service that is or may be unlawful or harmful.[6]

9 Additional expectation—provider will take reasonable steps regarding anonymous accounts

Additional expectation

(1) If the service permits the use of anonymous accounts, the provider of the service will take reasonable steps to prevent those accounts being used to deal with material, or for activity, that is or may be unlawful or harmful.

Reasonable steps that could be taken

(2) Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following:

(a) having processes that prevent the same person from repeatedly using anonymous accounts to post material, or to engage in activity, that is unlawful or harmful;

(b) having processes that require verification of identity or ownership of accounts.

10 Additional expectation—provider will consult and cooperate with other service providers to promote safe use

Additional expectation

(1) The provider of the service will take reasonable steps to consult and cooperate with providers of other services to promote the ability of end-users to use all of those services in a safe manner

(1.1) The provider of the service will take reasonable steps to consult and cooperate with providers of other services to prevent end-users from creating or sharing material or engaging in activity on the service that is or may be unlawful or harmful.[7]

Reasonable steps that could be taken

(2) Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following:

(a) working with other service providers to detect high volume, cross-platform attacks (also known as volumetric or 'pile-on' attacks);

(b) working with other service providers and relevant non-service providers to detect cross-platform users sharing material or engaging in activity on the service that is or may be unlawful or harmful, including livestreamed class 1 material;[8]

(c) sharing information with other service providers on material or activity on the service that is or may be unlawful or harmful, for the purpose of preventing such material or activity.

Division 3—Expectations regarding certain material and activity

11 Core expectation—provider will take reasonable steps to minimise provision of certain material

(1) The provider of the service will take reasonable steps to minimise the extent to which the following material is provided on the service:

(a) cyber-bullying material targeted at an Australian child;

(b) cyber-abuse material targeted at an Australian adult;

(c) a non-consensual intimate image of a person;

(d) class 1 material, including in the form of livestreaming of any child;[9]

(e) material that promotes abhorrent violent conduct;

(f) material that incites abhorrent violent conduct;

(g) material that instructs in abhorrent violent conduct;

(h) material that depicts abhorrent violent conduct.

Additional expectation

(2) The provider of the service will take reasonable steps to minimise the extent to which the following material is provided on the service:[10]

(a) material that instructs in preparatory child sexual exploitation and abuse activity ("grooming")

(b) material that instructs in how to avoid detection and prosecution for producing and distributing class 1 material;

(3) The provider of the service will take reasonable steps to ensure that technological or other measures are in effect to prevent access by children to any class 1 material on the service.[11]

Reasonable steps that could be taken

(4) Without limiting subsections (1) and (3) of this section, reasonable steps for the purposes of those subsections could include the implementing of on-device solutions.[12]

12 Core expectation—provider will take reasonable steps to prevent access by children to class 2 material

Core expectation

(1) The provider of the service will take reasonable steps to ensure that technological or other measures are in effect to prevent access by children to class 2 material provided on the service.

Reasonable steps that could be taken

(2) Without limiting subsection (1) of this section, reasonable steps for the purposes of that subsection could include the following:

(a) implementing age assurance mechanisms;

(b) conducting child safety risk assessments;

(c) implementing on-device solutions.[13]

Division 4—Expectations regarding reports and complaints

13 Core expectation—provider will ensure mechanisms to report and make complaints about certain material

(1) The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable end-users to report, and make complaints about, any of the following material provided on the service:

(a) cyber-bullying material targeted at an Australian child;

(b) cyber-abuse material targeted at an Australian adult;

(c) a non-consensual intimate image of a person;

(d) class 1 material;

(e) class 2 material;

(f) material that promotes abhorrent violent conduct;

(g) material that incites abhorrent violent conduct;

(h) material that instructs in abhorrent violent conduct;

(i) material that depicts abhorrent violent conduct.

Additional expectation

(2) The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable end-users to report, and make complaints about, any of the following material provided on the service:[14]

(a) material that instructs in preparatory child sexual exploitation and abuse activity ("grooming")

(b) material that instructs in how to avoid detection and prosecution for producing and distributing class 1 material

14 Additional expectation—provider will ensure service has terms of use, certain policies etc.

The provider of the service will ensure that the service has:

(a) terms of use; and

(b) policies and procedures in relation to the safety of end-users; and

(c) policies and procedures for dealing with reports and complaints mentioned in section 13 or 15; and

(d) standards of conduct for end-users (including in relation to material that may be posted using the service by end-users, if applicable), and policies and procedures in relation to the moderation of conduct and enforcement of those standards.

Note 1: See section 17 in relation to making this information accessible to end-users.

Note 2: For paragraph (b), the policies and procedures might deal with the protection, use and selling (if applicable) of end users' personal information.

Explanatory Notes

[1] Insert "block" in 6(3)(a) - Platform and service providers need to develop and implement processes that would prevent harmful or unlawful material from being uploaded, and harmful and unlawful activity taking place on the service in the first place, as opposed to merely detecting, moderating, reporting and removing the material/activity once it is already on the service. This is particularly important when dealing with online harm that is ephemeral, such as livestreamed abuse – there is no material uploaded that remains to be removed, but online abuse has occurred.

Automated technologies that are highly accurate in identifying illegal child sexual exploitation and abuse content on their platforms can be deployed for blocking harmful content or activities. Such

technology already exists, and tech companies should be encouraged to continually innovate for more accurate tools.

One example is a tool developed by Safe-to-Net, which has created on-device software that detects Child Sexual Exploitation Material (CSEM) in real time, identifying high-risk images using a machine-learning algorithm. It can also be implemented by social medial companies to prevent graphic content from being uploaded and distributed.

[2] 6(3)(e) *[new]* – Online safety should encompass all those at risk of abuse through the use of digital services, not just end-users of the services. Service providers need to incorporate measures to prevent end-users from harming other people through the use of their platforms. One vulnerable population at risk of abuse by end-users are children who are too young to use social media platforms themselves but may be subjected to abuse by end-users on these platforms, including via livestream. These victims urgently need to be identified and rescued from a slavery situation where they are being repeatedly abused and exploited.

[3] 6(3)(f) *[renumbered]* – in setting out the principle of Safety-by-Design in this provision, we recommend adding in the expectation that there be consideration of whether an on-device solution can or should be implemented to enhance online safety of the platform or service. See note 5 for a fuller discussion on on-device tools.

[4] Although this would not form part of the BOSE Determination, we recommend that the following be amongst the guidance material the Commissioner provides: Voluntary Principles to Counter Online Child Exploitation and Abuse; Guide for tech companies considering supporting the "Voluntary Principles to Counter Online Child Exploitation and Abuse".

[5] 8(2) *[new]* - A reasonable step that can be taken to detect and address material or activity that may be unlawful or harmful on an encrypted service, is to implement *on-device* artificial intelligence (AI) and machine learning tools that are highly accurate to identify such material/activity, and to flag, report and suspend such material/activity *before* it enters the encrypted stream.

Subsection 8(2)(a) references detecting and addressing harmful material/activity *on the service*, while subsection 8(2)(b) refers to addressing such material/activity *at the device level*.

We recommend explicitly encouraging tech companies to use the best, most advance AI tools and measures currently available and that will become available in the future, along with encouraging continued innovation of more accurate tools by industry. Client-side, on-device AI solutions could become the most effective method of preventing the production of CSEM in the first instance by blocking streaming or image/video capture, while maintaining high levels of privacy for users as it operates on-device and pre-entry into private channels. Such automated technologies protect the privacy of users as they are trained to only scan for CSEM, thus reducing the need for human eyes to review private content.

For example, Safe-to-Net's on-device software is an effective method of preventing the production of CSEM in the first instance by blocking streaming or image/video capture, while maintaining high levels of privacy for users as it operates on-device and pre-entry into private channels. This software can also be implemented by social medial companies to prevent unlawful or harmful content from being uploaded and distributed.

[6] Cross-industry collaboration via data-matching detection systems for the purpose of detecting and reporting crimes against children would aid in quickly identifying victims and offenders of livestreamed abuse. This is already done in the auto industry to detect insurance fraud, in the financial industry to detect key financial crimes, and should be done across industries to detect the worst forms of crimes against children imaginable.

IJM is working on a private sector solution with other companies to address this problem.

In another example, U.S.-based NGO Child Rescue Coalition (CRC) and Western Union have joined forces, with CRC providing unique data from its world-renowned technology to bolster Western Union's global Anti-Human Trafficking initiative. CRC's technology works by monitoring file sharing networks in real time, indexing 30 to 50 million records of online suspects involved in trading CSEM every day. This information allows CRC to help expose hidden networks of abusers and report their activity.

[7] 10(1.1) [*new]* – This provision makes explicit the responsibility of tech companies to ensure that their platforms or services are not being used to cause harm to people, and that their duty is not met by protecting only end-users. [see note 1, above]

[8] 10(2)b) *[renumbered]* – Cross-sector collaboration is critical to detecting unlawful or harmful material or activity on a service that is otherwise difficult to detect. Non-service providers holding relevant data in the case of online child sexual exploitation include financial companies. This provision explicitly encourages data sharing and cross-matching of suspicious data (in tokenised format, to protect privacy) between service providers and non-service providers, which when taken together can help to quickly identify both offenders and victims of first generation CSEM and livestreaming abuse.

IJM has developed a tool, *Tech and Financial Sector Indicators of Livestreaming Online Sexual Exploitation of Children,* which lays out specific language and behaviours of offenders and traffickers indicative of the production of new CSEM, including via livestreamed video. These indicators reveal actions that are often not illegal on the surface, but when combined with each other, reveal a high likelihood of this form of abuse against children occurring. Please contact endosec@ijm.org to gain access to this document.

See also note 6.

[9] 11(1)(d) – We recommend making explicit reference to livestreaming child exploitation/abuse as a form that class 1 material can take, to ensure that tech companies take steps to minimise this form of material on their platforms, despite the ephemeral nature of the material.

Another drafting option would be to specify the inclusion of livestreamed child abuse within "class 1 material" in a note:

> '*Note: "class 1 material" in subsection (1)(d) includes material in the form of livestreaming of a child where the content of the livestream would be classified as class 1'*

[10] 11(2)*[new]* – This provision places an expectation on tech companies to take steps to limit online communities of offenders and web forums that exist for the purpose of encouraging discussion about child sexual abuse among its members, facilitating the online distribution of child abuse material, and providing tips and techniques for grooming children and evading detection.

Such communities normalise offenders' behaviour, provide encouragement and validation, and enable offenders to share and learn tradecraft, thus decreasing the likelihood that individuals will seek help and increasing the chances of their offending escalating, and helping them to evade detection and prosecution.

[11] 11(3) *[new]* – this provision makes explicit the responsibility of service providers to take reasonable steps to prevent children from accessing class 1 material. Although the previous subsection puts the onus on tech companies to minimise the extent to which class 1 material is available on the service, this provision seeks to prevent children from accessing the class 1 material that may exist on the service despite the measures in subsection (2).

[12] 11(4)(c) *[new]* – A reasonable step that could be taken to ensure that provision of certain material are minimised on the service, and that children are prevented from accessing class 1 material on a

service, would be to implement on-device solutions. [See notes 1 & 5 for discussion of on-device solutions currently available]

[13] 12(2)(c) *[new]* – A reasonable step that could be taken to ensure that children are prevented from accessing class 2 material on a service would be to implement on-device solutions on devices that are used by children. [See notes 1 & 5 for discussion of on-device solutions currently available]

[14] See explanatory note 10, above.

About IJM

International Justice Mission (IJM) is a global organisation that protects people in poverty from violence. IJM and our partners are helping local authorities protect more than 400 million people from violence. As the largest anti-slavery organization in the world, IJM partners with local authorities in 24 program offices in 14 countries to combat slavery, violence against women and children, and other forms of abuse against people who are poor. Our model works side-by-side with local authorities and governments to rescue and restore survivors, hold perpetrators accountable in local courts, and strengthen the public justice system so it can better protect people from violence. This model is replicable and has worked to reduce modern day slavery and violence in programs against commercial sexual exploitation of children, among others.

About IJM's Center to End Online Sexual Exploitation of Children

IJM's Center to End Online Sexual Exploitation of Children partners with governments, industries, NGOs, and other stakeholders to expose, neutralise, and deter the online sexual exploitation of children around the world. Leveraging practices proven effective in IJM's ongoing program against OSEC in the Philippines, the Center helps (1) improve technology and financial sector platforms detection and reporting of livestreamed sexual abuse, (2) strengthen international collaboration in law enforcement and prosecution, and (3) support effective justice system (law enforcement, prosecution, and aftercare) responses in source and demand-side countries, resulting in sustainable protection for children and accountability for perpetrators.