IJM

IJM Submission On:

Revised Draft Industry Codes of Practice for the Online Industry

International Justice Mission (IJM) welcomes the opportunity to provide input into this second public consultation on the *Industry Codes of Practice for the Online Safety Industry* ("Industry Codes"). We have previously submitted detailed comments on the initial draft Industry Codes. This second submission on the Revised draft Industry Codes should be read in conjunction with our previous comments: International Justice Mission Submission 1.[i]

Summary of Recommendations

IJM recommends that the draft Industry Codes v2.0 be revised as follow:

Social Media Services Code
1. Require social media services providers to take action to disrupt and deter both CSAM and CSEM, in SMS MCM10.
Relevant Electronic Services Code
1. Require relevant electronic services providers to take action to disrupt and deter both CSAM and CSEM, in RES MCM10. 2. Relevant electronic services with end-to-end encryption technology are capable of reviewing and assessing Class 1A materials and should be required to do so. Encrypted RES and closed communication RES providers should be required to commit to MCM8, to use systems, processes, and technologies to detect and remove CSAM. 3. Relevant electronic services should require approval from eSafety to be found "not capable of reviewing and assessing materials". In applying for an exemption, service providers must provide a detailed report of all detection technologies examined as potential tools for their platform and the reason(s) they are not capable.
Designated Internet Services Code
1. Require designated internet services providers to take action to disrupt and deter both CSAM and CSEM, in DIS MCM9. 2. Require designated internet services providers to take action to deter and disrupt first-generation CSAM, including livestreamed CSAM, similar to the requirement in SMS (MCM10) and RES (MCM10). 3. Amend DIS MCM9 to require ongoing investment in systems, processes and technologies that detect first-generation CSAM, including CSAM present in livestreamed video. 4. DIS MCM9 must apply to end-user-managed hosting services, requiring ongoing investment in systems and processes to detect and remove known and new CSAM. 5. DIS MCM 8 must apply to end-user-managed hosting services, requiring them to use systems, processes, and technologies to detect and remove known CSAM.

Background

International Justice Mission (IJM) is a global organisation that protects people in poverty from violence. We partner with local authorities in 29 program offices in 17 countries to combat slavery, violence against women and children, and other forms of abuse against people living in poverty. IJM works with local authorities and governments to safeguard and restore survivors, hold perpetrators accountable, and help strengthen public justice systems so they can better protect people from violence.

Since 2011, IJM has worked with the Philippine Government, international law enforcement, and relevant stakeholders to combat online sexual exploitation of children (OSEC), in particular the trafficking of children to produce new child sexual exploitation material (CSEM) especially via livestreaming video. To date, IJM has supported 324 law enforcement-led rescue operations, leading to the rescue and safeguarding of 1081 victims or at-risk individuals, the arrest of 333 suspected traffickers, and the conviction of 184 perpetrators. Leveraging IJM Philippines' promising practices in combatting livestreaming of child sexual abuse and exploitation, IJM's Center to End Online Sexual Exploitation of Children works to strengthen the global response against this crime, including via improved industry detection and reporting. The Center is available for consultation to industry, government, and NGOs.

*Note: in this submission, references to "livestreamed video" or "livestreaming" describe any technology whereby live video is transmitted, including, but not limited to, through video-chat or video conferencing platforms.

Strengths of the Industry Codes

IJM applauds the measures included in the revised draft Industry Codes that strengthen the actions to be taken by Tier 1 social media services and Tier 1 relevant electronic services to detect, report, remove, disrupt, and deter new CSAM. The requirement in Social Media Services (SMS) Code Minimum Compliance Measure #10 (MCM10) and Relevant Electronic Services Code (RES) MCM10 that service providers invest in detection and disruption technologies is a significant step forward in protecting children online. These measures demonstrate an understanding of the serious safety risks posed by first-generation (or "new") CSAM and the urgency to prevent imminent harm. As rightly set out in the Guidance under both of these compliance measures, "Newly generated material is more likely to indicate current and ongoing safety risks such as against a child being groomed and coerced into producing new abusive images."

We are encouraged to see that SMS MCM10 requires implementation of techniques that enable providers to identify and monitor areas of highest safety risks on their services, and where feasible and appropriate, deploy safety technologies such as AI or deep machine learning techniques to detect and remove new CSAM. Anticipating how products can be abused will ultimately protect children across the globe from being harmed in the first place.

RES MCM10's requirement to invest in systems, processes and/or technologies aimed at disrupting and deterring end-users from using the service to create, post, or disseminate CSAM captures and can prevent many of the ways offenders sexually abuse and exploit children. In the Philippines, children are being abused in person by an adult facilitator, with the abuse directed in live video by a paying offender online, and eventually that abuse spreads globally as it is shared in images and recorded videos. These three categories of offending (facilitator (or "trafficker"), buyer (or "demand-side offender"), distributor), all abuse a child

and can be deterred through relevant electronic services implementing disruption and prevention technology.

Opportunities to Further Strengthen the Industry Codes

1. **Disrupt and Deter Both CSAM and CSEM.**

We note that the Head Terms limit the definition of "CSAM" to only "visual depictions of child sexual abuse." Designated Internet Services (DIS), Relevant Electronic Services (RES), and Social Media Services (SMS) Codes should require companies to take action against both CSAM and CSEM, in order to disrupt and deter such materials, to combat the many types of illegal, exploitative, and/or harmful materials that may be distributed on their platforms.

> **Recommendation:** Require platforms to take action against both CSAM and CSEM, in DIS MCM9, RES MCM10, and SMS MCM10.

2. **Whether a service provider is capable of reviewing and assessing materials.**

The applicability of certain minimum compliance measures under both the RES Code and DIS Code depend upon whether a provider is capable of reviewing and assessing materials. Some of the caveats and assumptions set out in the RES include the following:

> "Due to the nature of relevant electronic services, and the manner in which they are otherwise regulated, providers of these services may not be capable of reviewing and assessing private communications shared by end-users on their services." (under 3. Definitions)
>
> *"Certain relevant electronic services such as closed communication relevant electronic services and encrypted relevant electronic services will often not be capable of reviewing and assessing materials or capable of removing materials because they are legally not permitted to detect the relevant material and/or do not have access to relevant messages to enable providers to review material being shared or any surrounding communications to assess context in accordance with the National Classification Scheme. For example, providers of SMS, MMS services and encrypted relevant electronic services often do not have access to the content of any communications…"* (under MCM3 Guidance)

The above statements fail to acknowledge the existence of technological tools that can be used by end-to-end encrypted platforms to detect CSAM, in compliance with platform Terms of Service and without violating legally protected user privacy rights. Encrypted service providers should be required to employ available tools to detect CSAM and key words or behavioural signals indicative of CSAM production and distribution to prevent the use of encrypted platforms for the production and distribution of both known and new CSAM. See section 2(b) for a non-exhaustive sample of existing technologies that are applicable for CSAM detection and prevention.

a) Known CSAM

The detection of known CSAM is critical to preventing the revictimisation of survivors through continued distribution of images and videos of their sexual abuse. RES Minimum Compliance Measures 3, 8 and 11 indicate that some RES providers may be incapable of reviewing and assessing Class 1A and Class 1B materials. RES MCM 3 and 11 require only those RES providers that are capable of reviewing and assessing materials to implement systems and processes that

enable the provider to take action for breaches of its terms of service with respect to Class 1A and Class 1B material. RES MCM8 exempts encrypted RES providers from complying with the commitment to implement systems, processes and/or technologies to detect and remove known CSAM.

The following are two examples of technologies that would enable RES providers to review, assess, and remove material, even in an encrypted environment:

1. Apple created a technology using **client-side hashing** called NeuralHash[ii] that ensures privacy for both end-users and survivors of child sexual abuse and exploitation. As explained by Dr. Hany Farid, client-side hashing technology extracts and encrypts the hash from the sender, decrypts the hash at the server level, and only sends on non-CSAM content.[iii]

2. Similarly, technology exists whereby **secure enclaves** within company servers decrypt a message, compute a hash, and block CSAM from being sent beyond the server. This type of detection technology is hosted at the service provider but is not visible by anyone at the company.[iv]

Client-side hashing technology is currently already being used by end-to-end encrypted service providers. For example, WhatsApp detects harmful content such as 'suspicious links' through scanning text messages and flagging them. As described on the company's website,

> *"WhatsApp automatically performs checks to determine if a link is suspicious. To protect your privacy, these checks take place entirely on your device, and because of end-to-end encryption, WhatsApp can't see the content of your messages."* [v]

This detection and scanning technology maintains legally protected user privacy rights (it does not "break" encryption) while still protecting end-users from malware. **Similar technology can be used to detect CSAM and protect vulnerable children, while also keeping illegal and harmful content off of platforms.**

> **Recommendation:** Relevant electronic services with end-to-end encryption technology are capable of reviewing and assessing Class 1A materials and should be required to do so. Encrypted RES and closed communication RES providers should be required to commit to MCM8, to use systems, processes, and technologies to detect and remove CSAM.

Under RES MCM3 and MCM11, the assessment of whether a service provider is capable of reviewing and assessing material must include consideration of existing technological tools, including client-side hashing, secure enclaves, amongst others.

To ensure full compliance with existing technological capabilities to detect and remove Class 1A and Class 1B materials, relevant electronic services should not be exempted from the requirement on the basis of "not [being] capable of reviewing and assessing materials" without approval from eSafety. Providers who seek an exemption must provide a detailed report of all detection technologies examined as potential tools for their platform and the reason(s) they are not capable.

> **Recommendation:** Relevant Electronic Services should require approval from eSafety to be found "not capable of reviewing and assessing materials". In applying for an exemption, service providers must provide a detailed report of all detection technologies examined as potential tools for their platform and the reason(s) they are not capable.

b) First-generation (or "new") CSAM

There is no provision in the DIS Code that requires a commitment to disrupt and deter first-generation CSAM. Further, although designated internet services are required to make ongoing investment in systems and processes and/or technologies and personnel to detect and take action concerning child sexual abuse material under DIS MCM9, the requirement is limited to known CSAM. This is not a reasonable or appropriate limitation, given the immense harm suffered by children due to the production, distribution, and dissemination of first-generation CSAM, including via livestreamed video.

To protect children from online sexual abuse and exploitation, it is critical that the digital industry address first-generation CSAM. It is equally vital to address livestreamed CSAM, including CSAM that is produced in a context where younger children (usually pre-pubescent) are abused by financially motivated offenders. INTERPOL reports that "Live-streaming of child sexual exploitation for payment has seen an increase in recent years," as demand surged during the pandemic as an alternative to 'in-person' abuse.[vi] Similarly, Europol warns that "livestreaming of child sexual abuse increased and became even more popular during the COVID-19 pandemic."[vii] WeProtect Global Alliance reported in the 2021 Global Threat Assessment, "Livestreaming is on the rise, enabled by connectivity and the availability of inexpensive streaming devices. It often manifests as a cross-border crime that demand a coordinated international response."[viii]

Victims of CSAM production urgently need to be identified and safeguarded from a situation where they are actively being abused and exploited. Tech companies have the ability not only to support this identification through existing detection technology, *but tech companies can also disrupt or prevent livestreamed abuse in real time.*

Technologies and processes aimed at detecting first-generation CSAM currently exist and are being deployed on a range of services. Below are some examples of technological tools or actions taken by platforms in real-time to address livestreamed CSAM.

- Safety technology company, SafeToNet, has created a real-time video & image threat detection technology, SafeToWatch,[ix] capable of determining whether visual data represents undesirable and illegal content such as pornography, sexually suggestive imagery, cartoon pornography, and/or CSAM. The machine-learning algorithm will hash images with harmful content and render the content harmless. SafeToNet can provide more information to the eSafety Commissioner or industry associations upon request.
- Thorn, a child protection technology developer, created Safer to scale CSAM detection, increase content moderation efficiency, and optimise detection using advanced AI technology.[x] It identifies known and first-generation CSAM, leveraging cryptographic, perpetual hashing and machine learning algorithms to detect CSAM at scale and disrupt its viral spread.
- Google's Content Safety API and CSAI Match uses programmatic access and artificial intelligence to help platforms classify and prioritise billions of images for review.[xi] The higher the priority given by the classifier, the more likely the image contains abusive material, helping platforms prioritise human review and make their own content determinations.
- The social livestreaming platform, Yubo,[xii] proactively screens live video to keep children safe online, implementing automated prompts to users to change behaviour and disabling violative livestreams.

- DragonflAI[xiii] is a prevention and disruption tool that moderates livestreams completely on-device before they are streamed to platform. It detects illegal content such as CSAM and prevents content from being uploaded.
- Cyacomb Safety,[xiv] a detection technology designed for end-to-end encryption protects personal privacy while anonymously matching and detecting known CSAM with shared user content.

Commitments by both Tier 1 and end-user managed hosting services to invest in systems, processes, and technologies, for both known and first-generation CSAM (including livestreamed), are critical in preventing violations of platform Terms of Service, and in detecting and reporting CSAM to the CyberTipline and other clearinghouses. This will support efforts to safeguard children from situations of ongoing abuse and exploitation.

> **Recommendation:** Require designated internet services providers to take action to deter and disrupt first-generation CSAM, including livestreamed CSAM, similar to the requirement in SMS (MCM10) and RES (MCM10).
>
> **Recommendation:** Amend DIS MCM9 to require ongoing investment in systems, processes and technologies that detect first-generation CSAM, including present in livestreamed video.

3. Cloud Storage Platforms and other end-user managed hosting services.

In the DIS Code, end-user managed hosting services are exempt from the minimum compliance measures relating to detection and removal of known CSAM (MCM8) and ongoing investment in systems and processes to detect and take appropriate action (MCM9). These exclusions fail to recognise the key role cloud storage and cloud sharing (which are end-user-managed hosting services) have played in the proliferation of CSAM. As stated in WeProtect Global Alliance's 2021 Global Threat Assessment:

> Cloud sharing apps fuel explosion of user interactions with harmful content. Offender populations have come to rely on the ease of use, security and privacy of cloud file sharing apps to store and distribute illegal images and videos. Cloud storage makes it possible to share child sexual abuse material by simply posting a link in a forum, on a platform or through direct messaging, to thereby reach more offenders, more quickly.[xv]

Platforms like Google Drive and Dropbox (and others) are currently designed in such a way that needlessly allows this proliferation of both new and known CSAM distribution.[xvi] There is no valid legal, policy, or other reason for this failure to proactively detect and prevent these violations of Cloud storage platforms' terms of service.

Detection technologies that already exist and can be deployed by these cloud storage platforms include:

- PhotoDNA by Microsoft
- Safer by Thorn
- Content Safety API and CSAI Match by Google. This tool is also used by Meta.
- Hash Sharing by NCMEC (with Hash Matching API support from Google)
- Hash Matching API support from Google)

Recommendation: DIS (MCM 8) must apply to end-user-managed hosting services, requiring them to use systems, processes, and technologies to detect and remove known CSAM.

Further, the DIS Code contains no requirement for the detection of first-generation (new) CSAM. New child sexual abuse material indicates immediate danger of present and future harm, therefore requiring urgent intervention. This is acknowledged in the SMS (MCM10) and RES (MCM10); a similar requirement should be included in the DIS, applicable to all designated internet services providers.

Recommendation: DIS (MCM9) must apply to end-user-managed hosting services, requiring ongoing investment in systems and processes to detect and remove known and new CSAM.

Contact:

John Tanagho
Executive Director
IJM's Center to End Online Sexual Exploitation of Children
LinkedIn| osec.ijm.org

Hiroko Sawai
Analyst, Advocacy Research
IJM Australia
hsawai@ijm.org.au| IJM.org.au

Endnotes

[i] https://onlinesafety.org.au/wp-content/uploads/wpforms/31-9e10405917e4c106ebe4ec5e69a7bc86/IJM-Online-Industry-Codes-Submission-Sept2022-fbb2fc2b31ab7278c4d829640acb7bf2.pdf
[ii] https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf
[iii] https://protectchildren.ca/en/hany-farid-photodna/
[iv] See https://protectchildren.ca/en/hany-farid-photodna/ at 15:15
[v] https://faq.whatsapp.com/393169153028916/?cms_platform=web
[vi] https://www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-highlights-impact-of-COVID-19-on-child-sexual-abuse
[vii] https://www.europol.europa.eu/publications-events/main-reports/internet-organised-crime-threat-assessment-iocta-2020
[viii] https://www.weprotect.org/wp-content/plugins/pdfjs-viewer-shortcode/pdfjs/web/viewer.php?file=/wp-content/uploads/Global-Threat-Assessment-2021.pdf&dButton=true&pButton=true&oButton=false&sButton=true#zoom=0&pagemode=none
[ix] https://safetonet.com/safetowatch/
[x] https://safer.io/
[xi] https://protectingchildren.google/tools-for-partners/#learn-about-our-tools
[xii] https://www.yubo.live/newsroom/yubo-joins-forces-with-ncmec-to-combat-the-spread-of-sexually-inappropriate-photos-and-videos-of-minors-online
[xiii] https://www.dragonflai.co/
[xiv] https://www.cyacomb.com/company/news/2022/september/first-line-of-defence-cyacomb-launches-online-safety-software-to-combat-child-sexual-abuse-whilst-protecting-privacy/?utm_source=ActiveCampaign&utm_medium=email&utm_content=News+and+opportunities+from+across+the+Alliance&utm_campaign=September+2022+Newsletter
[xv] https://www.weprotect.org/global-threat-assessment-21/#report
[xvi] "The Internet is Overrun with Images of Child Sexual Abuse. What Went Wrong?" New York Times, Sept. 29, 2019) ("While the material, commonly known as child pornography, predates the digital era, smartphone cameras, social media and cloud storage have allowed the images to multiply at an alarming rate. Both recirculated and new images occupy all corners of the internet, including a range of platforms as diverse as Facebook Messenger, Microsoft's Bing search engine and the storage service Dropbox"), https://www.nytimes.com/interactive/2019/09/28/us/child-sex-abuse.html?mtrref=undefined&gwh=87D8702D4E20AACBE46104A68FFCF9DC&gwt=pay&assetType=PAYWALL