Collective Shout for a world free of sexploitation

Industry Standards - Public Consultation

Draft Online Safety (Relevant Electronic Services - Class 1A and 1B Material) Industry Standard 2024

Draft Online Safety (Designated Internet Services - Class 1A and 1B Material) Industry Standard 2024.

January 22 2024

Introduction

For many years Collective Shout has been observing and critiquing developments in the production and online distribution of Child Sexual Abuse Material (CSAM). We have a long history of documenting harms facilitated and enabled by online platforms and calling for their regulation. We thank eSafety for taking this next step to compel the online industry to address these harms. Our submission pertains only to CSAM rather than pro-terror material, as sexual exploitation is our area of expertise.

In our Submission on Draft Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material), we called for major changes that would close loopholes that would have allowed CSAM to continue to proliferate on digital services.

We support the requirements outlined in the draft standards for both Designated Internet Services (DIS) and Relevant Electronic Services (RES). We strongly support the inclusion of closed communication and encrypted services in minimum compliance measures. We are pleased to see a strong start made on regulating generative AI.

The proposed additional requirements placed on the industry are proportionate to the extreme harms caused by the production, proliferation, and consumption of CSAM. These requirements constitute new opportunities for the digital sector, allowing the development of a safer and more inclusive internet.

We are hopeful that the framework of investing in research and development of new technologies, combined with the requirement to adopt whatever appropriate technologies are available, will result in a growing capacity for the online industry to address CSAM on and through its platforms.

We appreciate the opportunity to comment again on this important legislative development and propose further recommendations.

Summary of Recommendations

1. More specific guidance should be given in the technical feasibility exception.
2. A service should be required to establish and implement investment and development programs if they meet the monthly user threshold and/or a specified annual turnover threshold.
3. Standards for RES and DIS must require a mandatory time limit in which class 1A and 1B material is removed, and in which further risks are effectively dealt with.
4. The Standard should include the provision of support to personnel who are required to deal with CSAM.
5. Providers must be required to share detection models and information about known bad actors.

About Collective Shout

Collective Shout (www.collectiveshout.org) is a grassroots campaigning movement challenging the objectification of women and sexualisation of girls in media, advertising and popular culture. We target corporations, advertisers, marketers and media which exploit the bodies of women and girls to sell products and services, and campaign to change their behaviour. More broadly we engage in issues relating to other forms of sexploitation, including the interconnected industries of pornography, prostitution and trafficking as well as the growing market in the sale of children for Live Distant Child Abuse[1] and in child sex abuse dolls and replica child body parts.[2]

Our work puts us in touch with the unique and specific ways children are at risk, especially in their vulnerability to online grooming by predators and exposure to pornography. Young

[1] Tankard Reist, Melinda (2017). Why are Australian Telcos and ISPs enabling a child abuse pandemic? ABC Religion and Ethics. https://www.abc.net.au/religion/why-are-australian-telcos-and-isps-enabling-a-child-sexual-abuse/10095644; For a summary of recent global campaigns relating to on-line child protection see Collective Shout (6 Sep 2021). National Child Protection Week 2021: Join our campaigns to protect children and young people. https://www.collectiveshout.org/child_protection_week_2021

[2] Roper, Caitlin (2022). Sex Dolls, Robots, and Woman Hating: The Case for Resistance. Spinifex Press. https://www.spinifexpress.com.au/shop/p/9781925950601; See also Roper, Caitlin (9 Jan 2020). “Better a doll than a real child:” The spurious logic used to justify child sex dolls. ABC Religion and Ethics. https://www.abc.net.au/religion/spurious-logic-used-to-justify-child-sex-dolls/11856284

people are at unique risk of sexualisation, objectification and exploitation online. They are vulnerable to cyberbullying, sexual harassment, image-based abuse, predatory behaviour, grooming and exposure to pornography. This causes physical and psychological harm, hampering healthy development.

We have documented these harms for the past 14 years, including in the following:

- Submission to the previous inquiry on this matter - Draft Consolidated Industry Codes of Practice for the Online Industry (Class 1A and 1B Material).[3]
- Submission to Select Committee on Social Media and Online Safety 2022;[4]
- Submission to eSafety Consultation on the implementation roadmap for a mandatory age verification (AV) regime relating to online pornography 2021;[5]
- Submission on Harm Being Done to Australian Children Through Access to Pornography on the Internet to the Senate Environment and Communication References Committee 2016;[6]
- Submission to the Inquiry into Age Verification for Online Wagering and Online Pornography 2019;[7]
- Submission to the United Nations' review Children's Rights in the Digital Environment 2020;[8]
- Submission to the inquiry into Law Enforcement Capabilities in Relation to Child Exploitation 2021;[9] and
- Numerous other publications and commentaries.[10]

[3] Collective Shout (Oct 2022). Submission on Draft Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material). https://www.collectiveshout.org/submission_draft_codes_class1a_1b

[4] Collective Shout (Jan 2022). Submission to Select Committee on Social Media and Online Safety. https://www.collectiveshout.org/submission_social_media_online_safety

[5] Collective Shout (2021). Submission: eSafety Consultation on implementation roadmap for a mandatory age verification (AV) regime relating to online pornography.

[6] Collective Shout (2016). Harm being done to Australian children through access to pornography on the internet: Submission to the Senate Environment and Communications References Committee. https://d3n8a8pro7vhmx.cloudfront.net/collectiveshout/pages/1019/attachments/original/1457408234/CS_Submission_Harms_of_Pornography_Inquiry_March_2016.pdf?1457408234

[7] Collective Shout (2019). Submission to Inquiry into Age Verification for Online Wagering and Online Pornography. https://www.collectiveshout.org/submission_to_inquiry_into_age_verification_for_online_pornography

[8] Collective Shout (30 Nov 2020). UN Submission: Children's Rights in the Digital Environment. https://www.collectiveshout.org/un_sub_children_digital_rights

[9] Collective Shout (20 Aug 2021). Submission: Law Enforcement Capabilities in Relation to Child Exploitation. https://www.collectiveshout.org/submission_law_enforcement_child_exploitation

[10] For example, see Tankard Reist, Melinda (2016). Early sexualisation and pornography exposure: the

detrimental impacts on children, Australian Childhood Foundation blog. https://professionals.childhood.org.au/prosody/2016/07/melinda-tankard-reist/; Tankard Reist, M. (2016). Growing Up in Pornland: Girls Have Had It with Porn Conditioned Boys, ABC Religion & Ethics. https://www.abc.net.au/religion/growing-up-in-pornland-girls-have-had-it-with-porn-conditioned-b/10097244; Tankard Reist, Melinda (2018). Never Again? Addressing Sexual Violence Must Include Pornography, ABC Religion & Ethics.

Discussion questions 4 and 23: Is the technical feasibility exception in the obligation to detect and remove known child sexual abuse material and pro-terror material appropriate? How effective will this obligation be with this exception?

The exception in section 7 is overly subjective and could become a major loophole. More specific requirements must be given about the expectation of costs incurred in relation to the risks averted.

CSAM causes grave harms in production, distribution, and consumption. We are not confident that providers are always willing to sacrifice profit to prevent harm. From observing self-regulation in the advertising industry, we already know that there are too many companies who do not care about the harm they cause, even when claiming a commitment to Corporate Social Responsibility.

There is a growing body of evidence about the outcomes of specific technologies as they are applied in the real world. eSafety could be a repository of research findings that would, over time, be able to better inform businesses on the potential impact of introducing new technologies to address CSAM. Nonprofits that monitor CSAM like NCMEC, Tech Coalition, WeProtect, Internet Watch Foundation, Thorn, and others can also advise on the expected benefits of strategies. Hash matching, for example, is widely used and its outcomes can be quantified. All major platforms keep statistics on the outcomes of their hash matching technologies. Providers will be able to access data for all existing technologies, and new ones as they emerge. Thus, more specific guidance should be included in the technical feasibility exception.

Recommendation: More specific guidance should be included in the technical feasibility exception.

Discussion questions 8 and 24: Do you agree with the monthly active user threshold for the investment obligation? Are there other appropriate thresholds that should be considered to ensure the obligation is proportionate to the size and reach of the relevant electronic service?

https://www.abc.net.au/religion/never-again-addressing-sexual-violence-must-include-pornography/10094568; Tankard Reist, Melinda (2021). Consent education does not stand a chance against pornography, ABC Religion & Ethics, https://www.abc.net.au/religion/consent-education-does-not-stand-a-chance-against-pornography/13231364.

Section 23, Development Programs, applies only to services with an average monthly number of active end-users of one million or more. Under the Draft Standard, these services must establish and implement, for the calendar year, a program of investment and development activities in respect of systems, processes and technologies.

A similar approach should be taken as with modern slavery reporting requirements under the NSW Modern Slavery Act;[11] a minimum annual individual turnover threshold should additionally be applied to the proposed monthly user threshold. A service would need to invest in development if they meet the monthly user threshold *and/or* the annual turnover threshold.

Some services may have fewer than one million monthly active end-users in Australia, but may have a very high revenue either because their global reach is significant, or because they run a highly profitable small company. For the calendar year *following* the EOFY in which the particular turnover threshold is exceeded, the service should be required to establish and implement investment and development programs.

This recommendation would also impact Section 37, under which services should also report their annual turnover.

Recommendation: A service should be required to establish and implement investment and development programs if they meet the monthly user threshold *and/or* a specified annual turnover threshold.

Discussion question 10: Should the requirement on certain relevant electronic services to respond to reports of class 1A and class 1B material on their service be limited to a requirement to take 'appropriate action'?

Standards for RES and DIS must require a mandatory time limit in which class 1A and 1B material is removed, and in which further risks are dealt with. We propose this because in our years of experience advocating for children's online safety, we have found some large digital services to be unwilling to take action in reasonable time, if at all.

In the Discussion Paper in the previous consultation report, eSafety states it has "encountered services that lack effective content reporting mechanisms, fail to find and address overt CSEM and are non-responsive to removal requests. These services may deny

[11] Burn, Jennifer (2019). NSW Modern Slavery Reporting Requirement Guidance Material. NSW Government. https://www.parliament.nsw.gov.au/lcdocs/other/12778/Tabled%20document%20-%20Professor%20Jennifer%20Burn%20-%20NSW%20Department%20of%20Premier%20and%20Cabinet.pdf

ownership or connection to harmful content, refuse to acknowledge its presence on their systems or show a lack of will to remove it."

This corresponds with our observations over past years of monitoring CSAM and grooming on various platforms. The community expects wealthy tech companies, with vast resources at their disposal, to prioritise the best interests of every child,[12] not their bottom line.

In our many years dealing with self-regulation in the advertising industry, and in making complaints to Instagram and Twitter about exploitation on their platforms, we have experienced and documented how legitimate complaints are frequently ignored or dismissed.

- Ad Standards, for example, has a years-long documented history of failure to do its job. It frequently dismisses complaints from the public about advertisements that are clearly degrading to women, pornographic and in view of children, including BDSM themes.[13] Once an ad is denied panel review, there is no other avenue for having a complaint heard.[14] While waiting for the panel to review community complaints, months can pass and the ad remains in place; when the decision is made, often the advertising campaign has ended anyway.
- One of our campaigners reported to Twitter that men were discussing raping and impregnating pre-teen girls and violently dismembering women. Twitter had responded to say its Safety Policies had not been broken. It was only after we tweeted CEO Parag Agrawal, Chair Bret Taylor, major shareholder Vanguard, and Elon Musk, asking why Twitter endorsed men's explicit desires to sexually abuse young girls, that a number of these accounts were suspended.[15]
- Instagram had responded to our campaigners' report to say that the relevant account did not go against Community Guidelines. This account had published BDSM-themed pictures and videos of a prepubescent girl in sexualised poses, in fetish wear and chains. It was only after media coverage that Instagram pulled this account dedicated to promoting pre-teen "models" – it had over 33k followers. Even after the account was pulled, a hashtag containing the page's name returned 11k posts featuring adultified pre-teen and toddler girls.[16]
- When we reported child sex abuse comments made on reels featuring pre-teen girls, Instagram responded that it was too busy to review them and suggested we hide the

[12] ECPAT (Sept 2022). Proposal for a regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse - Publications Office of the EU (europa.eu). Submission of ECPAT International in consultation with ECPAT members in the EU. https://op.europa.eu/en/publication-detail/-/publication/13e33abf-d209-11ec-a95f-01aa75ed71a1/language-en/format-RDF

[13] Roper, Caitlin (8 Jun 2020). The sexist ads endorsed by Ad Standards. Collective Shout. https://www.collectiveshout.org/how_ad_standards_justifies_sexist_advertising?fbclid=IwAR1SMBgcenz6jCfXTyKJicXuHsrW3xfvMTTaGMh7GN2T6nqFWt0qV6aS0c0

[14] Kennedy, Lyn (27 Aug 2019). Self-regulated advertising: How many more examples of failure do we need? Collective Shout. https://www.collectiveshout.org/self_regulated_advertising_more_examples_of_failure

[15] https://www.collectiveshout.org/takedown_twitter_instagram

[16] https://www.collectiveshout.org/takedown_twitter_instagram

content if we find it ‘upsetting’. One particular video of young girls dancing sexually attracted 77k views at the time of reporting. Instagram should have taken the opportunity to investigate the account, with nearly 15k followers, dedicated to videos of young girls dancing sexually for mostly male followers, rampant predatory activity, and paedophile networking including invitations to off-site chat groups.

- We have dozens of examples of Instagram failing to review our reports of child exploitation. Some of our reports from January 2022 (10 months ago) are still ‘in review.’ Of 100 reports of child exploitation we made during August 2022, Instagram has reviewed only half. In every case, the account user promoted sales and/or trade of child sexual abuse material - often via links to Mega (NZ cloud storage company) folders and files. Of the half which have been reviewed, Instagram took action to remove just three accounts/pieces of content. Of the remaining reports, Instagram said it did not remove the content as it did not go against its Community Guidelines.
- An investigation conducted by cybersecurity group, Ghost Data, identified the more than 500 accounts that openly shared or requested child sexual abuse material over a 20-day period during September 2022. Twitter failed to remove more than 70% of the accounts. Of the accounts which remained online, many were soliciting materials for "13+" and "young looking nudes."[17]
- Meta’s training manual for content moderators instructs them, in cases where the age of the subject of suspected child exploitation material was unknown, to “err on the side of adults.”
- A young woman, Rose, emailed Pornhub multiple times to take down videos of men raping her at age 14.[18] The videos were left live and monetised by Pornhub until she resorted to impersonating a lawyer. Rose says that dozens of women have reached out to her with similar experiences.

In the RES Standard, this recommendation of mandatory minimum timeframes for removal of CSAM once identified would apply to:

- Section 25(2)(c)
- Section 15(2) and (3)
- Section 17(2)(c)

In the DIS Standard, it would apply to:

- Section 15(2) and (3)
- Section 17(2)(c) and (d)
- Section 21(4)

[17] Collective Shout (4 Oct 2022). Big brands pull ads from Twitter after child exploitation investigation. https://www.collectiveshout.org/big_brands_pull_ads_from_twitter

[18] McNamara, Haley (24 Feb 2020). Credit card companies should stop partnering with porn websites. Washington Examiner. https://www.washingtonexaminer.com/opinion/op-eds/credit-card-companies-should-stop-partnering-wIth-porn-websites

Recommendation: Standards for RES and DIS must require a mandatory time limit in which class 1A and 1B material is removed, and in which further risks are effectively dealt with.

Discussion questions 11 and 25: What are your views on the likely compliance costs and, in particular, the impact of compliance costs on potential new entrants?

An additional cost, not currently required by the Standard but essential to sustainable ongoing compliance with the Standard, is psychosocial support for personnel involved in these compliance requirements who may experience trauma by exposure to CSAM and pro-terror materials. This is a serious occupational hazard.

Research suggests that professionals may not access internal mental health support because of stigma or lack of trust in the process.[19] It may even be the case that providers would not offer this support to their employees. It was not mentioned in the Standard as a cost but we believe it should be a requirement to ensure the wellbeing of employees as well as sustainability of compliance.

Recommendation: The Standard should include the provision of support to personnel who are required to deal with CSAM.

Discussion questions 18 and 19: In relation to high impact generative AI designated internet services, do the proposed obligations provide appropriate safeguards? Are there specific challenges to deploying these measures in a generative AI context?

We are deeply concerned about the use of Artificial Intelligence (AI) technologies in the production of image based abuse and child exploitation material, and the growing proliferation of this content on mainstream social media platforms.

We believe it is important to acknowledge that AI itself is not creating CSAM or image based abuse material. AI content is generated by real people who prompt machine learning software. This software is trained on a vast body of digitised images and videos including real CSAM, images of real children and other real pornography also created by real people. For this reason we question the validity of the term “AI generated” in connection with CSAM and image based abuse material. We know that these types of abuse are highly gendered,

[19] Redmond, T., Conway, P., Bailey, S., Lee, P. and Lundrigan, S. (10 May 2023). How we can protect the protectors: learning from police officers and staff involved in child sexual abuse and exploitation investigations. *Frontiers in Psychology*. 14:1152446. doi: 10.3389/fpsyg.2023.1152446

with males most often being the perpetrators and consumers of CSAM.[20] We believe the term “AI generated” serves to dehumanise the act of creating abuse content and shield sexual offenders - the men creating it - from critique and accountability. We hope that eSafety’s approach to AI always directly addresses the reality of how it is created and who its perpetrators and victims are.

Our recent findings include:

- On the popular AI frontend platform Chub, child sexual abuse narrative and chat generators have been found. For example, one character was a 14 year old girl confined to a hospital bed in a coma. The character description implied a male doctor's desire to abuse the defenceless child. Another character was designed to generate chats for men to fantasise about raping teen girls with disabilities. Many ‘NSFW’ characters were tagged ‘little sister’ and were designed to generate incest themed child exploitation material.
- Highly realistic sexualised imagery of prepubescent girls distributed on X (formerly Twitter). Content was often tagged #stablediffusion (denoting generation by text-to-image model created by StabilityAI). The content often had extensive reach and engagement (millions of views; thousands of likes and shares) and revealed paedophile networking and other child exploitation activity.
- Instagram has been hosting AI content fetishising young boys.[21]
- Pornified, objectifying and sexualised AI content produced using the likeness of real women and girls. For example, Neural.Love AI ‘art generator’ hosted images of 16 year old ‘model’ and ‘influencer’ Presley Elise. The creator titled the images ‘Presley Elise with little clothing’. Presley Elise is a known victim of child exploitation. We also documented AI child exploitation images created in the likeness of a child version of actor Emma Watson.

We commend eSafety on making a start to address this rapidly emerging risk. We will continue to monitor, report, and engage the public in advocacy for online safety. We will also continue to call on the heads of technology companies to demonstrate corporate social responsibility and ensure their efforts to eliminate child exploitation and image based abuse extend to AI content.

Discussion questions 12 and 26: Is there any additional information eSafety should consider in determining the Relevant Electronic Services/Designated Internet Services Standard?

[20] Sexual Assault - Perpetrators: Sexual assault statistics for offenders proceeded against by police, criminal court outcomes for defendants, and prisoners in adult custody. (2 Feb 2022). *Australian Bureau of Statistics.* https://www.abs.gov.au/articles/sexual-assault-perpetrators; Child Sexual Abuse Material: The Facts (Feb 2023). *National Children’s Advocacy Center.*

[21] https://x.com/CollectiveShout/status/1746833429659087318?s=20.

We would like to see, in Section 23 of the RES Standard, a requirement that providers share detection technologies and information on bad actors that they have identified on their platforms. As one platform cleans up its user base, malicious users tend to migrate to other platforms. This would additionally make entry-level players in the industry more able to meet minimum compliance standards.

In Section 24(7) of the DIS Standard, these should be requirements, rather than examples:

- Sharing information on best practice approaches that are relevant to the services;
- Working with the Commission to share information, intelligence, best practices and other relevant information;
- Collaborating with non-government or other organisations that facilitate sharing of information, intelligence, best practices and other information.

Thiel, Stroebel and Portnoff, in their 2023 analysis of generative machine learning and CSAM, recommend that providers should collaborate with regard to sharing of detection models and sharing of information about known bad actors.[22] We believe this should be a requirement in the Standards.

Recommendation: Providers must be required to share detection models and information about known bad actors.

[22] Thiel, D., Stroebel, M. and Portnoff, R. (24 June 2023). Generative ML and CSAM: Implications and Mitigations. *Thorn and Stanford Internet Observatory Cyber Policy Center.*