Amendment to the Online Safety (Basic Online Safety Expectations) Determination 2023

Collective Shout

February 2024

Introduction

We welcome this opportunity to comment again on the BOSE Determination, made under section 45 of the *Online Safety Act 2021.* The proposed amendments are a significant improvement. We especially welcome stronger requirements relating to generative AI, transparency, terms of use, and accountability.

While platforms and providers need flexibility to adopt new technologies as they arise to meet the goals of the Online Safety Act, the BOSE could be further improved by being more specific in terms of elements included in the requirements and in terms of actions given as examples, so that no aspects of a service can fall through the cracks, either inadvertently or deliberately.

We have continued to express concern that large social media companies have put profits ahead of safety. Digital industry groups in Australia did not prioritise community safety in some of their draft Codes of Practice, and as such were not accepted by eSafety. Big Tech has not to date shown a strong propensity to develop new means of improving children's online safety. Typically, action comes only after harms are made public, via leaks or independent research. Accordingly, civil penalties must be more substantial so as to be compelling. And oversight must be independent, with audits by qualified assessors.

Australia, as a signatory to the UN Convention on the Rights of the Child, has a responsibility to advocate for children's rights in the digital space. Children's best interests should be independently evaluated, regularly and explicitly, in the context of *all* aspects of the platforms on which children can or do interact, or on platforms where children may be featured.

In some amendments, the term 'reasonable' should be changed to 'necessary,' and 'could' should be changed to 'should,' to protect children from grave harms and ensure that the digital industry - proven to be reluctant to implement certain safety measures- will do the right thing.

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 people are at special 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.

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

- Submission to Draft Online Safety (Relevant Electronic Services and Designated Internet Services – Class 1A and 1B Material) Industry Standard 2024.[3]
- Submission to the previous inquiry on this matter - Draft Consolidated Industry Codes of Practice for the Online Industry (Class 1A and 1B Material).[4]
- Submission to Select Committee on Social Media and Online Safety 2022;[5]
- Submission to eSafety Consultation on the implementation roadmap for a mandatory age verification (AV) regime relating to online pornography 2021;[6]

---

[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; 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

[3] Collective Shout (22 Jan 2024). *Submission to Draft Online Safety (Relevant Electronic Services and Designated Internet Services – Class 1A and 1B Material) Industry Standard 2024*. https://www.collectiveshout.org/tags/submissions

[4] 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

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

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

- Submission on Harm Being Done to Australian Children Through Access to Pornography on the Internet to the Senate Environment and Communication References Committee 2016;[7]
- Submission to the Inquiry into Age Verification for Online Wagering and Online Pornography 2019;[8]
- Submission to the United Nations' review Children's Rights in the Digital Environment 2020;[9]
- Submission to the inquiry into Law Enforcement Capabilities in Relation to Child Exploitation 2021;[10] and
- Numerous other publications and commentaries.[11]

We track the activities of online predators on popular social media sites, documenting and reporting thousands of accounts for preying on underage/prepubescent girls, attempting to engage with them privately, describing sex abuse acts they wish to carry out on these girls, and soliciting, selling and trading child exploitation material. We have also documented the tracking, tagging and sharing of the Instagram content of hundreds of underage girls to paedophile forums operating on the open web.

Our joint global #WakeUpInstagram campaign with the National Center on Sexual Exploitation (USA), Courtney's House (US) and Defend Dignity (Canada) exposed Instagram as a platform for predators to access children, pornography companies to promote and link to hardcore porn sites, for hosting offers of paid sexual content featuring children, and for facilitating other practices harmful to children and young people.[12]

Generative AI, recommender systems, and user controls

---

[7] 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

[8] 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

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

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

[11] 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.*

[12] See https://www.collectiveshout.org/_instagram.

Generative AI

Deepfakes and AI CSEM are of critical concern. They constitute some of the most harmful material online, and their use is escalating rapidly. Sexually exploitative material should be treated as a particularly harmful context in which AI is used. The BOSE Determination Amendment is a timely opportunity.

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, with males most often being the perpetrators and consumers of CSAM.[13] Males are also more likely to perpetrate image based sexual abuse, while females are more likely to be victimised by a partner or ex-partner.[14]

The term “AI generated” serves to dehumanise the act of creating abuse content and shield offenders - the men creating it - from critique and accountability. We urge eSafety to address the reality of how AI CSAM is created and clearly identify its perpetrators and victims.

Our recent investigations found:

- Child sexual abuse narrative and chat generators hosted on the popular AI frontend platform Chub. 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 reposts) and revealed paedophile networking and other child exploitation activity.
- Instagram hosting AI content fetishising young boys.[15]
- 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

[13] 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.
[14] Henry, N, Flynn, A and Powell A (2019). Image-based sexual abuse: Victims and perpetrators. Australian Institute of Criminology. https://www.aic.gov.au/publications/tandi/tandi572#:~:text=The%20nature%20of%20victimisation%20and,a%20partner%20or%20ex%2Dpartner.
[15] https://x.com/CollectiveShout/status/1746833429659087318?s=20

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 have also discovered AI child exploitation images created in the likeness of a child version of actor Emma Watson. As well, we recently commented on the AI porn created in the likeness of Taylor Swift and argued this technology poses risks for all women and girls. [16]

We recommend a change to the wording of Section 8, given the potential severe harms, so that additional requirements become much stronger and more effective. The word ‘reasonable’ should be replaced with the word ‘necessary,’ and the word ‘could’ to be replaced with ‘should’ to ensure that protecting users from severe harm is not an optional extra for digital platforms.

Risk assessment and harm mitigation must be prioritised at every step in the development and deployment of AI, with focus on the best interests of children (not only child users, but all children, as victims of CSAM may not necessarily be users of the service). Similar to the well-established and effective use of HACCP in the food industry, it should become mandatory and routine to minimise risk at every level.

Recommendations for Generative AI:

Section 8A(1): If the service uses or enables the use of generative artificial intelligence capabilities, the provider of the service will take *necessary* [replace the word ‘reasonable’] steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of artificial intelligence capabilities on the service.

Section 8A(2): If the service uses or enables the use of generative artificial intelligence capabilities, the provider of the service will take *necessary* [replace the word ‘reasonable’] steps to proactively minimise the extent to which generative artificial intelligence may be used to produce material or facilitate activity that is unlawful or harmful.

Section 8A(3): Without limiting subsection (1) and (2), necessary steps for this section should [replace the word ‘could’] include the following:

> **(a) ensuring that assessments of safety risks and impacts are undertaken, identified risks are appropriately mitigated, and safety review processes are implemented throughout the design, development, deployment and post-deployment stages of generative artificial intelligence capabilities;**
>
> **(b) providing educational or explanatory tools (including when new features are integrated) to end-users that promote understanding of generative artificial intelligence capabilities on the service and any risks associated with the capabilities;**

[16] See https://www.collectiveshout.org/ai_a_tool_for_abusing_women_and_children and https://www.collectiveshout.org/putting_women_in_their_place_ai_abuse_of_taylor_swift

(c) ensuring that training materials for generative artificial intelligence capabilities and models do not contain unlawful or harmful material;

(d) ensuring that generative artificial intelligence capabilities can detect and prevent prompts that generate unlawful or harmful material.

Recommendation: Add (e) "ensuring independent audits of the functions of AI systems" as another example.

Recommender Systems

Recommender systems cause harm to children as well as adult users by serving up sexualised content, self-harm and eating disorder material, among other kinds of harmful content. It can also connect children with predators, dramatically increasing children's risk of being groomed, exploited, and abused. Recommender systems operate on content, friend suggestions, follower suggestions, targeted advertising, suggested search terms, autocomplete, autoplay, trending lists, popular hashtags, nudges, and of course viral news, reels, and challenges etc.

Recommender systems have been central to businesses strategies to increase user engagement and advertising revenue. These systems prioritise content or make personalised suggestions to users, and are invisible to the user. It has become clear that recommender systems have been serving harmful content to users, including children, such as sexualised content and promotion of self-harm, eating disorders, suicide, and hate speech. For example:

- Content that normalises the sexualisation of children; or sexual adult content that may be harmful to children who access it;
- To children and teens, friend/follower suggestions that include unknown adults who may be potential predators;
- To potential predators, friend/follower suggestions of children or young people who share similar 'interests' such as modelling or gymnastics (which predators follow in order to find these vulnerable children);
- Content that promotes disordered eating behaviour like extensive fasting or binge eating;
- Content promoting ideal body and beauty standards, much of which is heavily edited or completely fake.

A pertinent example is the finding by the European Data Journalism Network and AlgorithmWatch that Instagram's algorithm of recommendations prioritised "scantily-clad" men and women.[17] Recommender systems are so pervasive that creators are afraid to

17 https://docs.google.com/document/d/1L7A5hmskm3Y3huSXHNtIIoiVijHD3dkDqubff4Yvkg8/edit#

speak out about them, for fear of being shadow-banned – another function of recommender systems.[18]

We have documented the harmful recommendation of child accounts (sometimes "parent-run") to adults who are following child models, dancers, gymnasts, influencers etc. Some adults clearly game this system by adopting interests similar to children in order to be connected with them. Children are more likely to accept a friend/follower request when they have a mutual friend, just as adults are.

In our submission to the Select Committee on Social Media and Online Safety (January 2022), we called for requirements for social media platforms to stop recommending unconnected adults to minors in Discover pages, and restrict all adults from seeing minors via search tools. Meta has responded to community pressure, especially to the #WakeUpInstagram campaign by ourselves, National Center On Sexual Exploitation (NCOSE in the USA), and Defend Dignity (Canada). Since May 2022, when children under 18 sign up for Instagram, default settings specify that unconnected accounts cannot tag, mention, or use their content. Platform users over 19 cannot send private messages to teens who do not follow them. Meta introduced many new policies to protect under 18 accounts from unconnected adults.[19]

Meta also rolled out parental controls, but it has become clear that few parents use them. By the end of 2022, less than 10% of teens on Instagram had enabled the parental supervision setting. And even fewer had parents who had actually adjusted their child's settings. Research shows that being time poor and having limited tech knowledge are among barriers to supervising their children online, yet social media platforms continue to expect parents to do this work.

The ACCC found in their Digital Platform Services Inquiry Interim Report (Sept 2022) that digital platforms of all sizes engage in unfair trading practices "including choice architecture that exploits consumers' behavioural biases and undermines consumer choice," software designs that are also referred to as Dark Patterns.[20]

Further, the ACCC observed public extensive distrust of digital platforms. Yet digital platforms have become essential to the economy and to the lives of vast number of ordinary people. For social and educational purposes, teens find that digital platforms are difficult to avoid. We agree with the ACCC that the economic benefits of the digital industry will be expanded only by strengthening public trust and safety. When everyone can safely engage with digital platforms, the benefits will truly outweigh the costs.

---

[18] https://algorithmwatch.org/en/instagram-algorithm-nudity/

[19] Meta (2024) About Instagram teen privacy and safety settings https://help.instagram.com/3237561506542117

[20] Commonwealth of Australia (September 2022). Digital Platform Services Inquiry Interim Report No. 5 – Regulatory reform. Australian Competition and Consumer Commission https://www.accc.gov.au/system/files/Digital%20platform%20services%20inquiry%20-%20September%202022%20interim%20report.pdf

We expect services to be much more responsible in how they develop, implement, and maintain recommender algorithms. We expect services to be transparent about how they are developed and used, and to prioritise the wellbeing of users, particularly children.

Recommendations:

> **Specify that all types of recommender systems are applicable under this section. Content moderation systems, advertising management and deployment systems, and all other systems and elements employed by service providers should be subject to these additional expectations.**
>
> **Paragraph 6(3)(e) ensuring that assessments of safety risks and impacts are undertaken, identified risks are appropriately mitigated, and safety review processes are implemented, throughout the design, development and post-deployment stages for the service. [Add: These should be independently audited.]**

The best interests of the child and access to age-inappropriate materials online

We welcome the best interests of the child being prioritised in these Amendments. Reflecting the unique vulnerability of children and the lifelong impact adversarial childhood experiences have on them as adults, the best interests of the child should be the primary consideration at every step. We include here any decisions about user controls, independent audits, generative AI, recommender systems, enforcements of terms of use, and more.

We continue to urge the Federal Government to implement an Age Verification system, in line with its obligations under the Convention of the Rights of the Child, Articles 19 and 34, and the Optional Protocol to the CRC on the Sale of Children, Child Prostitution and Child Pornography.[21] Government has failed to respond adequately to community concerns and the body of evidence testifying to significant harms to children.[22]

Of special concern to Collective Shout are sextortion and sex trafficking. The ACCC found in their Digital Platform Report that inadequate verification of digital platform users and content means digital platforms are increasingly used by scammers.

An example of this is children livestreaming on social media. Weak age verification is common among service providers - we know that many children, even as young as 9, have social media profiles, with parents known to allow children to falsify their ages, despite a minimum of age 13 on most social media sites.

---

[21] https://www.ohchr.org/en/instruments-mechanisms/instruments/optional-protocol-convention-rights-child-sale-children-child

[22] Open Letter: Women's safety and child protection experts call for age verification pilot. September 19, 2023. https://www.collectiveshout.org/open_letter_age_verification

Live streaming is a particularly risky feature of social media services. Social media has a format that encourages users to share videos from personal spaces such as bedrooms and bathrooms. Predators are provided the opportunity to connect in a private capacity and build trust in order to exploit. Services have often set children's livestream to public by default, making them visible to millions of people including strangers. They enable viewers to move from public interactions to private messaging. The 'like' visualisations particularly appeal to children who have a strong desire for love and affirmation. In live chat functions, 5Rights Foundation found that 6% of children have been asked by viewers to change or remove their clothing on film.[23]

We viewed a 13-year-old girl's live posts during which she was bombarded with sexual comments including a request for sex, questions about her underwear, a man telling her he wanted her to give him an erection, another saying he would pay to meet her and others pressuring her to remove her shoes and show her feet. We witnessed a 14-year-old girl's live video joined by a naked man masturbating. We documented these and other similar instances:[24]

> *An underage girl has just – with no moderation or intervention from the global multi-billion dollar Facebook-owned platform - broadcast a live video of a naked man masturbating. She and her friend - and fifty other people - just witnessed a serious criminal act, prohibited by Australia's Commonwealth, state and territory child exploitation material laws.*
>
> *Who else witnessed the live sex act? Other school friends? Perhaps younger children – cousins or neighbours who tuned in to catch up on some big-girl news? How widely did Instagram disseminate this piece of child exploitation material that it failed to moderate and helped produce? How many times is this scene being played out in Australia each day? How many kitchens and bathrooms and bedrooms of Australian homes are being infiltrated by predators who want to abuse underage girls in this way? How many men are using Instagram to broadcast live sex acts to children? Has this type of criminal behaviour become 'normal' for girls who have been desensitised to predatory advances because sexual objectification, harassment and predation are so entrenched in their everyday, lived experiences? Why - in flagrant disregard of human rights, law, child safety principles and common sense - is Instagram connecting predators to minors?*
>
> *Four days later our concerns that this event was not a one-off, that predators are targeting underage girls for the purpose of broadcasting live sex acts to them and that this is 'normal' for some girls were confirmed when we found the public Instagram account of a 9 year old girl based in Europe. She had saved a live post to her profile, allowing anyone to watch it for the 24-hour period that followed. We watched the video and saw that it was interrupted several times as the young girl accepted requests from different viewers to be in the broadcast. We counted three*

[23] https://www.riskyby.design/livestreaming-and-video-sharing
[24] Kennedy, Lyn (17 Mar 2020). School girl's Instagram 'live' post becomes sex predator webcam. Collective Shout. https://www.collectiveshout.org/schoolgirl_instagram_live_post_sex_predator_webcam

different viewers who filmed themselves masturbating. We then followed the girl. Within an hour we received a notification from Instagram that she had started a live post. We began viewing the video immediately and within seconds she accepted a viewer's request to be in the broadcast. It was another naked, masturbating man.

It is therefore very encouraging to see Proposed subsection 6(2)(a) creating new additional expectations to take reasonable steps to ensure that the best interests of the child are a primary consideration in the design and operation of any service that is used by, or accessible to, children.

Recommendations:

Add: The provider of the service will conduct regular assessments of the design and operation of the service in how it impacts the best interests of children accessing the service.

Add: The best interests of a child should be a primary consideration when making decisions about user controls.

Add: Assessments of the impact on children's wellbeing should be carried out before rolling out any new features or services.

Age Assurance Technologies

Paragraph 12(2)(a) provides an example of a reasonable step to ensure that technological and other measures are in effect to prevent access by children to class 2 material. We believe that it should be a *necessary* step to implement appropriate age verification mechanisms. This expectation remains tech-neutral but requires that children are much less able to access class 2 material.

It is clear that the digital industry does not want to implement age verification, while the community clearly wants age verification in place as soon as possible. Without mandated age verification we are concerned that digital platforms will find it 'reasonable' to have the weakest barriers. This is because young users represent, in a business sense, a steady stream of new customers.

Tech companies tend to only put controls in place after weaknesses or abuses are made public.[25] Whistleblower Frances Haugen proved with data that Facebook made decisions to maximise profit and growth at the cost of users' wellbeing.[26]

Based on this and other evidence – including the growing body of global literature demonstrating that pornography exposure harms children – we have continued to call for implementation of an age verification system across all social media platforms whose services are available for access and use by children, or where content featuring children is published, where class 2 material may be found.

For more information and a summary of research evidence, refer to our Submission to Inquiry into Age Verification for Online Wagering and Online Pornography[27] and our submission to the United Nations on Children's Rights in the Digital Environment.[28]

Recommendations:

Paragraph 12(1): The provider of the service will take necessary [remove '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.

Transparency

We welcome improved expectations in transparency. Mechanisms relating to complaints, terms of use, policies and procedures about end user safety, and standards of conduct are all very important for user safety.

Based on the ACCC's finding in September 2022 that digital platforms have ineffective complaints and dispute resolution processes,[29] we support the recommendation in that

---

[25] Nix, Naomi (30 Jan 2024). Meta says its parental controls protect kids. But hardly anyone uses them. *The Washington Post*. https://www.washingtonpost.com/technology/2024/01/30/parental-controls-tiktok-instagram-use/ Huang, K. (Jan 2024). Meta Rejected Efforts to Improve Youth Safety, Documents Show. *The Information*. https://www.theinformation.com/briefings/meta-rejected-efforts-to-improve-youth-safety-documents-show

[26] Zakrzewski C. and Albergotti R. (11 Oct 2021) The education of Frances Haugen: How the Facebook whistleblower learned to use data as a weapon from years in tech. *The Washington Post*. https://www.washingtonpost.com/technology/2021/10/11/facebook-whistleblower-frances-haugen/

[27] https://www.collectiveshout.org/submission_to_inquiry_into_age_verification_for_online_pornography

[28] https://www.collectiveshout.org/un_sub_children_digital_rights

[29] Commonwealth of Australia (September 2022). *Digital Platform Services Inquiry Interim Report No. 5 – Regulatory reform*. Australian Competition and Consumer Commission

report (which the present amendments reflect) that the Government establish mandatory minimum standards for internal dispute resolution processes, as well as an external ombudsman scheme for the digital services industry. This has also been our experience over almost 15 years of grassroots advocacy for a digital industry that is safer for children and free of sexploitation.

For transparency (subsection 18A) we recommend mandatory independent audits such as is required in the European Union under the Digital Services Act (DSA) for regulation of Very Large Online Platforms. The European Commission 2023 Commission adopts rules on independent audits under the DSA, plus access by appropriate academics and organisations to data about public safety and harms to users as is necessary to conduct research into compliance as well as detect new threats and harms which are sure to arise.

We recommend that transparency be extended to the provision of data to relevant academics and organisations that have a focus on digital safety, safety, and children's rights. In our submission to the Select Committee on Social Media and Online Safety (January 2022), we called for Meta to:

- Share its full research on children's mental health and well-being, and grant access to its data to independent researchers, civil society organisations and regulators;
- Set out what research has been conducted on how Facebook's services and design choices contribute to child sexual abuse, and publish the findings;
- Publish Facebook's risk assessments;
- Provide transparency on Facebook's product reputational reviews.

We also recommend that content moderation systems should be included in transparency requirements. Again, independent audits should monitor content moderation. Children and teens do not understand how content moderation works, and external nonprofit organisations have identified and researched Dark Patterns on social media platforms.[30]

The BOSE should list more detail in the expected information and metrics in transparency reports. This is because there is a risk that compliance reporting becomes simply a Public Relations exercise for companies. In the past, Facebook voluntarily and publicly reported on the most simple metrics that made their safety policies look effective and benevolent; for example, simple statistics on how much harmful material was removed. In reality, they were aware from their own research that grave harms were occurring on the platform, and they had clearly placed profits above children's wellbeing.

Recommendations:

- **Independent audits should be mandated.**

---

https://www.accc.gov.au/system/files/Digital%20platform%20services%20inquiry%20-%20September%202022%20interim%20report.pdf

[30] https://www.reset.tech/resources/risktominors/instagram-and-risks-to-minors/

- **The BOSE should provide greater detail on the expected information and metrics in transparency reports.**
- **Service providers should ensure that researchers are able to access data that is in the public interest with regard to safety and wellbeing.**

Enforcement of Terms of Use

We have documented many instances of social media platforms failing to enforce terms of use, to the detriment of users, especially children.

An example is the failure of social media platforms to undermine parents' monitoring of children's accounts. In multiple reports Facebook describes secret, secondary Instagram accounts hidden from parents or family known as Finstas as a "unique value proposition," indicating that secret accounts are a growth strategy to boost activer user metrics. At the same time, Facebook has rolled out tools to help parents navigate social media and keep their kids safe online. Parents are unable to apply these tools to accounts they know nothing about.[31] Given the history, parents cannot trust companies like Meta to prioritise children's safety over profit.

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

For example, one of our campaigners reported to Twitter/X that men were discussing raping and impregnating pre-teen girls and violently dismembering women. Twitter/X 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/X endorsed men's explicit desires to sexually abuse young girls, that a number of these accounts were suspended.

Other examples include:

- Instagram 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.
- 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

[31] US Congress (30 Sept 2021). *Protecting Kids Online: Facebook, Instagram, and Mental Health Harms*. Witness: Antigone Davis, Global Head of Safety. https://www.commerce.senate.gov/2021/9/protecting-kids-online-facebook-instagram-and-mental-health-harms

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 are still ‘in review.’ Of 100 reports of child exploitation we made during August 2022, Instagram 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 were 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/X 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."[32]
- Meta’s training manual for content moderators instructed 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.[33] 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.

We welcome this amendment to hold service providers more accountable, rather than merely relying on user reports and potentially ignoring or responding inappropriately to them.

However again we recommend that stronger language be used. Hash matching, for example, is widely used and proven technology that the community expects to be used. It is a technology that *should* be used to detect CSAM, rather than *could*. Services *must* respond to user reports of unlawful and harmful behaviour.

---

[32] 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

[33] 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-wlth-porn-websites