Industry Standards – eSafety Public Consultation

January 2024 [Extended Deadline]

Organisation: WeProtect Global Alliance

Submitting officers: Shailey Hingorani and Eleanor Linsell

Contact: [redacted] and [redacted]

About WeProtect Global Alliance

WeProtect Global Alliance ('the Alliance') is a non-profit that brings together people and organisations with the knowledge, experience and influence to transform the global response to child sexual exploitation and abuse online. As of January 2024, its membership is comprised of 102 government members – including the Australian Government – 70 companies, 94 civil society groups and 10 international organisations.

The Alliance strongly supports the Online Safety Act (2022) which provides for industry bodies to develop new codes to regulate 'class 1' and 'class 2' illegal and restricted online material and for eSafety to register codes if they meet the statutory requirements. The Alliance supports the two draft industry standards drafted by the eSafety Commissioner ('eSafety') for which consultation was opened on 20 November 2023.

As a multi-sector membership organisation spanning governments, civil society, the private sector and international non-governmental organisations, WeProtect Global Alliance occupies a unique position in the child protection sector and thus has a comprehensive viewpoint of both the global threat landscape for children in a digital environment and the current response to child sexual abuse online. As a consequence, the Alliance views rigorous codes and standards that improve online platform regulation as an integral part of the wider response to tackling child sexual abuse and exploitation online; one which involves coordinated, consistent and strategic action by a range of stakeholders – as set out in our well-established Model National Response (MNR) framework.

Responses to questions set out in the Consultation Discussion Paper *Draft Online Safety (Relevant Electronic Services – Class 1A and 1B Material) Industry Standard 2024* and *Draft Online Safety (Designated Internet Services – Class 1A and 1B Material) Industry Standard 2024*

Question 1: Are the requirements for risk assessment in the draft Standards targeted at the right services and at the right points in a service's development journey? Are the risk factors appropriate?

The Alliance is supportive of the outcomes- and risk-based approach outlined in the discussion paper by eSafety. It is true that different services and technologies have different risk profiles,

particularly when it comes to child sexual exploitation and abuse online. Illegal and harmful content, where the severity and immediate impact of the harm is most dangerous, or that poses an imminent threat to the lives and safety of vulnerable users, should attract the most stringent risk mitigation measures. Child sexual abuse content is a serious crime that can have devastating emotional, social and physical consequences for victims and survivors.

Platforms which are most at risk of causing serious and immediate harm through facilitating the most abhorrent crimes, such as child exploitation and abuse online, should be prioritised. Size, impact and market share should also be considered when determining how to prioritise the regulation of such a wide variety of platforms. In addition to these criteria, there is also a strong case to be made for a higher level of scrutiny of platforms that are particularly popular with and regularly used by children. It is essential that such platforms adhere to the rules and provide products which provide a safe, positive experience for child users, based on a sophisticated understanding of likely risks. They have a unique and critical responsibility in ensuring services are safe by design, that their teams have the resources and infrastructure to quickly detect and tackle harm and provide support to their more vulnerable users.

Industry stakeholders should have proactive, robust and comprehensive risk mitigation measures in place to detect, report, block and remove this content (both new child sexual abuse material and material which has already been identified), as well as to report it to the authorities in order for offenders to be brought to justice. In order to be effective, the type and way in which the mitigation measures are implemented should be flexible and allow the private sector to adopt solutions based on the specific risks on their services. There will need to be explicit and strict safety obligations on high priority issues such as child sexual abuse and exploitation online, and for platforms with a larger market share or high-risk profile.

In explaining the risk mitigation measures adopted, tech companies should be able to show that user safety was prioritised in product design and engineering decisions and that they have adopted a Safety by Design approach.

Question 2: Do the obligations on each relevant electronic service and designated internet service category appropriately reflect the above considerations? Are other considerations relevant?

Yes. WeProtect Global Alliance agrees that the obligations placed on industry take into account the need to balance flexibility and enforceability, the risk-based and outcomes-based approach, technological feasibility and the paramount importance of balancing fundamental rights, such as the right to privacy and children's rights.

Given that the online world is constantly evolving, and new challenges, harms and threats to safety are also developing, changing and responding to existing safety measures, we are particularly supportive of the technology neutral approach and the need to be flexible and adaptable, so that we can collectively respond to new and emerging threats and harms in the space of child sexual exploitation and abuse online, such as AI-generated child sexual abuse material or sexual violence that occurs in the extended reality environments (to name but a few emerging threats).

Question 3: Is the test in section 5 workable? Is further guidance required to assist providers to determine whether this standard, or another code or standard, applies to a particular online service?

Many companies offer diverse ranges of digital products and governance structures and levels of independence of services vary. The test should provide as much information as possible to ensure companies are classified under the most appropriate code and that subsidiary services receive appropriate individual scrutiny.

Question 4: 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?

Automated content detection and moderation are essential elements of the response to tackle child sexual exploitation and abuse online. There are different ways in which automated technologies can be used to detect, report, remove and block child sexual abuse online. Effective detection of 'known' child sexual abuse material is made possible by two linked techniques called 'hashing' and 'hash-matching'. These techniques have significantly accelerated the identification and removal of known child sexual abuse material from the internet.

In addition to these techniques, the development of AI classifiers has been valuable in the detection, reporting, removal and blocking of 'unknown' or 'new' child sexual abuse material online. Such automated or semi-automated moderation systems identify illegal content by following rules and interpreting many different examples of content which is and is not harmful. In a 2021 survey of tech company practices, conducted by WeProtect Global Alliance and the Tech Coalition, 84% of the companies surveyed said they had at least partly automated processes for forwarding reports of child sexual abuse online, suggesting that report management is relatively efficient. Industry should continue to work – in partnership with safety tech experts and industry – on enhancing the accuracy of classifiers to detect 'unknown' child sexual abuse content (including livestreamed content) and grooming in both non-encrypted and encrypted video sharing environments. Open sourcing (with appropriate controls in place) should be used to encourage collaboration between relevant actors and help set consistent standards for safety technologies.

When it comes to the exception in the obligation to detect and remove known child sexual abuse material for end-to-end encrypted (E2EE) environments, a practical and solution-orientated approach should be adopted. E2EE digital services should be able to demonstrate their best efforts and intentions of detecting and removing illegal activity on their services, with clear evidence of the effectiveness of measures deployed (particularly where the technologies mentioned above cannot be deployed). The safety and privacy of child users (including survivors of abuse) must remain paramount. There are many emerging possibilities to help overcome the challenges of detecting, removing and reporting in E2EE environments which are further explored in Question 5.

Question 5: Are there other examples of systems, processes and technologies that can detect, flag and/or remove known child sexual abuse material and known pro-terror material at scale, which should be highlighted in the Standards or accompanying guidance?

As stated in WeProtect Global Alliance's 2023 Global Threat Assessment, there is growing understanding about to the potential for detecting child sexual abuse material in end-to-end encrypted (E2EE) environments. These include but are not exclusive to:

- Client-side scanning, which involves scanning messages on devices for matches or similarities to a database of illegal child sexual abuse material before the message is encrypted and sent;
- Homomorphic encryption, which is the use of a different type of encryption that allows operations to be performed without data decryption at any point; and
- Intermediate secure enclaves, which decrypt the message at the server level by a third party and use tools to detect child sexual abuse materials.

Such technologies are still relatively nascent and require further testing and exploration. However, given that the technological landscape is fast developing – both in terms of offender behaviour and in the response – it is important to consider how such technologies may evolve and be deployed in the future, with the best interests of child users of digital services as the determining factor in decision-making.

Question 6: Are there any limitations which would prevent certain service providers from deploying systems, processes and technologies to disrupt and deter child sexual abuse material and pro-terror material on relevant electronic services? If there are limitations, how might these be overcome? Is it appropriate for this requirement to apply to gaming services with communication functionality?

Disruption and deterrence are key elements of a preventative response to the wave of child sexual exploitation and abuse online and play a key role in secondary prevention (which focuses on individuals at higher risk of violence perpetration). One of the key findings from the Alliance's 2023 Global Threat Assessment is that all stakeholders – from governments to the private sector – need to invest more in prevention and that preventative responses must be prioritised. The Alliance is therefore very supportive of the inclusion of guidance to improve disruption and deterrence.

A key element of the deterrence response is providing avenues for offenders, potential offenders and those concerned about their thoughts to get help. From our global perspective, there is a real need for industry to work together with partners to signpost those who are potentially at-risk of committing harm to the right resources and help channels. According to the Lucy Faithfull Foundation, a UK charity dedicated to preventing child sexual abuse, the number of people seeking advice or support via online self-help, or their confidential helpline has trebled since 2020. A distinct barrier to investment in deterrence programmes is the public perception that, in doing so, society supports those that are harming children or are intent on doing so. This perception extends to therapists too. Research indicates that when people in the US who are concerned about their sexual interest in children seek help, they are often met with mental health providers who are ill-informed or prejudiced against them, which deters them from engaging in preventive therapy. From a public health perspective, it is crucial that people at risk of offending have Access to stigma-free, competent, and empathic treatment and that this is also signposted in deterrence messaging.

When it comes to disruption, discussions with the Alliance's members from across different sectors – including the gaming industry – have highlighted a general agreement that industry must make their services as disruptive and inhospitable as possible to bad actors. We believe that all relevant services and sectors that have child users have a responsibility to ensure these are as safe as possible, deploying all appropriate measures.

Question 7: Are there other examples of systems, processes and technologies that can disrupt and deter the use of a relevant electronic service to solicit, generate, distribute or access child sexual abuse material and pro-terror material, which should be highlighted in the guidance?

Deterrence messaging can take the form of warnings, blurring or blocking content, or providing safety information to end-users. The use of default settings, particularly for child users, is a promising area for development. On the simpler end, deterrence messaging can take the form of a pop-up box on search engine results pages when specific words are used to trigger a response and on the more comprehensive end, more complex interventions such as the recent trial of a chatbot created by Stop It Now, the Internet Watch Foundation and Pornhub to deter potential offenders before they commit a criminal offence. Different types of deterrence messaging require varying amounts of resources and financial support – services must determine the appropriate response based on their assessment of risk posed to users.

Question 8: 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?

Yes. The Alliance agrees that services with more than 1 million monthly active users in Australia should be required to have a program of investment and development to disrupt and deter child sexual abuse material, including first generation material.

Question 9: Are the end-user reporting requirements workable for the relevant service providers? Are there practical barriers to implementation?

End-user reporting is one of the more long-established mechanisms for content moderation. It is of paramount importance that end-users should be able to report a safety complaint on any type of online service and this should be a minimum requirement for all online service providers covered by the Standard. However, recognising that there are significant barriers to end user reporting (e.g. the stigma associated with child sexual abuse), this must only be one element of a service's response to the risk of child sexual abuse on their platforms and products. Child users must not feel that the onus is solely on them to address abuse and exploitation.

Digital services need to provide users with different reasons to flag content such as child sexual abuse content, hate speech, violence, or privacy violations and they should also be able to provide additional comments, feedback or evidence to ensure that claims of illegal and harmful content can be as robust as possible. Users should also be able to contest decisions by digital services if content flagged by them as inappropriate is not actioned.

Processing user reports requires robust and well-resourced Trust and Safety teams and processes, which can prove costly for industry players. However, high-performing services with

more than 1 million monthly active users in Australia have a responsibility in hiring enough content moderators, training and supporting their wellbeing and mental health and ensuring that they have the tools and resources to act swifty to identify and takedown illegal content.

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'?

The most dangerous and severe harms, that have the potential to cause real and significant emotional, physical and social harm, both immediately and in the longer term, should be prioritised. WeProtect Global Alliance therefore believes that class 1A material, such child sexual exploitation material, pro-terror material, and extreme crime and violence material, should be treated more expeditiously than class 1B material. In the case of child sexual exploitation and abuse, there should be clear expectations set out for industry regarding the reporting, removal and blocking of child sexual abuse material online.

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

The likely compliance costs of the standard will vary depending on the size and complexity of the service provider. Those with greatest resources and active users should take a proportionately greater level of responsibility. Each company is different in how it deals with hiring and training Trust and Safety staff, the type of technological systems they use to implement safety policies and how they conduct risk assessments, audits and transparency reports. For some technology companies, they are already committing to much of the above, whereas others are not as significantly invested. It is important for all industry players that provide digital services to take responsibility in ensuring that their services are safe by design and respond to the greatest extent possible, in light of the severity and impact of child sexual abuse and exploitation.

Question 12: Is there any additional information eSafety should consider in determining the Relevant Electronic Services Standard?

N/A

Question 13: Are the categories in Table 2 sufficiently clear for designated internet service providers to identify which category they fall within and therefore what obligations apply? What are the benefits and/or challenges of the categories as they are currently proposed?

The categories in Table 2 are sufficiently clear.

Question 14: Are the section 9(5) matters in the draft Standard appropriate and sufficiently clear to help designated internet service providers accurately self-assess which tier their service falls within?

The section 9(5) matters in the draft Standard are appropriate and clear, as is the tiering which clarifies which services fall into which category.

Question 15: eSafety is seeking to place requirements on service providers that are best-placed to prevent the use of generative AI features to create and disseminate class 1A and class 1B material. Does the proposal achieve this?

WeProtect Global Alliance's recent Global Threat Assessment 2023 report identified AI-generated child sexual abuse material as one of the key emerging trends that is further fuelling the sexual exploitation and abuse of children online. Generative AI tools are making it easier and quicker to produce illegal content rapidly and at scale. Open-source AI is already being used to produce results of infinite images by generating new images of abuse as well as modifying and distorting existing images of children. This includes increasingly violent and harmful versions of victims' abuse. Offenders are also using AI to share tips and generate demand, creating an ever-faster cycle of production. When real images are used to generate AI images such as sexual "deepfakes", the impact on those affected can be catastrophic. On top of this, the potential volume of new material threatens to overwhelm law enforcement agencies in identifying and rescuing at-risk children and bringing offenders to justice. As AI-generated content becomes increasingly realistic and convincing, the challenge for law enforcement and hotline services will be to differentiate between "synthetic" and real abuse material in order to save children from immediate danger, increasing the burden on already over-stretched services.

The Alliance supports the risk-based approach to identifying digital services where it is reasonably foreseeable that the model may generate high impact material, including child sexual exploitation and abuse material.

Question 16: Do the draft definitions for a high impact generative AI designated internet service and machine learning model platform service capture the right services? Are there types of providers that should not be included or should be excluded?

The right platforms and services are captured. By focusing on the potential for misuse, the threshold reflects the reality that even machine learning and artificial intelligence models with limited direct exposure to sensitive data or datasets containing illicit data may still be misused to create illegal content, such as "synthetic" child sexual abuse material and sexual deepfakes.

Open-source AI technology will present specific challenges to deployment. While closed model tools are maintained by digital service providers and cannot be modified by their users, open-source AI models can be downloaded and altered by users to evade safeguards and otherwise facilitate abuse and exploitation.

Question 17: The high impact generative AI designated internet services category only captures models that meet a high impact threshold. It must be reasonably foreseeable that a service can be used to generate synthetic high impact material that would be classified as X18+ or RC. Is this threshold: (a) appropriate for differentiating high impact and therefore high risk models? (b) sufficiently clear to enable service providers to assess whether or not they meet the definition? eSafety welcomes views on alternative thresholds which may be more suitable.

WeProtect Global Alliance understands RC (refused classification) material as comprising of class 1A material (child sexual exploitation material, pro-terror material, and extreme crime and

violence material) and class 1B material (crime and violence material and drug-related material). The Alliance therefore agrees that high impact material should be classified as X18+ and RC. Any models that can be used to create synthetic child sexual abuse material and sexual deepfakes must fall under the high impact generative AI designated internet services category. It is also important to stress that generative AI is not only being used to generate abuse material, but also guides on how to offend, guides on how to groom children and guidance on how to evade detection. These criminal acts must also be covered by the high impact threshold. Even though they might not directly result in child sexual abuse imagery or videos that are classified as X18+ or RC, they can still result in very real harm.

Question 18: In relation to high impact generative AI designated internet services, do the proposed obligations (in particular, the section 21 obligation to 'detect and remove' and the section 23 obligation to 'disrupt and deter' child sexual abuse material and pro-terror material) provide appropriate safeguards? Are there specific challenges to deploying these measures in a generative AI context?

The use of classifiers will be essential in the detection and removal of AI generated child sexual abuse material, as new material can constantly be created from scratch and existing material modified to evade widely deployed hash-matching technology. As mentioned above, the challenge with detecting and reporting AI generated child sexual abuse material will be the sheer volume of new material which in turn will likely overwhelm law enforcement agencies in identifying and rescuing at-risk children and bringing offenders to justice. As AI-generated content becomes increasingly realistic and convincing, the challenge for law enforcement and hotline services will be to differentiate between "synthetic" and real abuse material in order to save children from immediate danger.

It should be feasible to build disruption and deterrence into closed source AI models, however it is likely to be much more challenging with open-source AI models since they can be downloaded, modified and disruption/deterrence measures evaded.

Question 19: In relation to machine learning model platform services, do the proposed obligations (in particular, the section 23 obligation to 'disrupt and deter') provide appropriate safeguards? Are there specific challenges to deploying these measures?

N/A

Question 20: In relation to relevant enterprise providers, do the proposed obligations (in particular, the section 23 obligation to 'disrupt and deter') provide appropriate safeguards? Are there specific challenges to deploying these measures?

N/A

Question 21: Do sections 16 to 19 effectively reflect the considerations on minimum compliance measures outlined on pages 10-11 of this discussion paper?

Yes.

Question 22: Do the obligations for detecting and removing child sexual abuse material and pro-terror material effectively reflect the considerations on minimum compliance measures outlined on pages 10-11 of this discussion paper?

Yes.

Question 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 discussion paper lists a number of technological solutions that can be used to detect and remove known child sexual abuse material. As mentioned above, automated content detection and moderation are essential elements of the response to tackle child sexual exploitation and abuse online. There are different ways in which automated technologies can be used to detect, report, remove and block child sexual abuse online. Effective detection of 'known' child sexual abuse material is made possible by two linked techniques called 'hashing' and 'hash-matching'. These techniques have significantly accelerated the identification and removal of known child sexual abuse material from the internet. In addition to these techniques, the development of AI classifiers has been incredibly useful in the detection, reporting, removal and blocking of 'unknown' or 'new' child sexual abuse material online. Such automated or semi-automated moderation systems identify illegal content by following rules and interpreting many different examples of content which is and is not harmful. In a 2021 survey of tech company practices, conducted by WeProtect Global Alliance and the Tech Coalition, 84% of the companies surveyed said they had at least partly automated processes for forwarding reports of child sexual abuse online, suggesting that report management is relatively efficient. Industry should continue to work – in partnership with safety tech experts and industry – on enhancing the accuracy of classifiers to detect 'unknown' child sexual abuse content (including livestreamed content) and grooming in both non-encrypted and encrypted video sharing environments. Open sourcing (with appropriate controls in place) should be used to encourage collaboration between relevant actors and help set consistent standards for safety technologies.

When it comes to the exception in the obligation to detect and remove known child sexual abuse material for end-to-end encrypted (E2EE) environments, a practical and solution-orientated approach should be adopted. E2EE digital services should be able to demonstrate their best efforts and intentions of detecting and removing illegal activity on their services, with clear evidence of the effectiveness of measures deployed (particularly where the technologies mentioned above cannot be deployed). The safety and privacy of child users (including survivors of abuse) must remain paramount. There are many emerging possibilities to help overcome the challenges of detecting, removing and reporting in E2EE environments. As stated in WeProtect Global Alliance's 2023 Global Threat Assessment, there is growing understanding about what might and might not be possible when it comes to detecting child sexual abuse material in end-to-end encrypted (E2EE) environments. These include but are not exclusive to:

- Client-side scanning, which involves scanning messages on devices for matches or similarities to a database of illegal child sexual abuse material before the message is encrypted and sent;

- Homomorphic encryption, which is the use of a different type of encryption that allows operations to be performed without data decryption at any point; and
- Intermediate secure enclaves, which decrypt the message at the server level by a third party and use tools to detect child sexual abuse materials.

Such technologies are still relatively nascent and require further testing and exploration. However, given that the technological landscape is fast developing – both in terms of offender behaviour and in the response – it is important to consider how such technologies may evolve and be deployed in the future.

Question 24: Do you agree with this monthly active user threshold, or are there other thresholds which can be deployed to ensure this obligation is proportionate?

The Alliance finds the user base thresholds of 1 million monthly active users in Australia for Tier 1 designated internet services and high impact generative AI designated internet services and 500,000 monthly active users in Australia for end-user managed hosting services to be proportionate.

Question 25: What are your views on the likely compliance costs for service providers and, in particular, the impact of compliance costs on potential new entrants?

As above: the likely compliance costs of the standard will vary depending on the size and complexity of the service provider. Each company is different in how it deals with hiring and training Trust and Safety staff, the type of technological systems they use to implement safety policies and how they conduct risk assessments, audits and transparency reports. For some technology companies, they are already committing to much of the above, whereas others are not as significantly invested. It is important for all industry players that provide digital services to take responsibility in ensuring that their services are safe by design, in view of the impact of child sexual abuse and exploitation on users. There are already promising examples within WeProtect Global Alliance, such as the work of the Technology Coalition, of service providers collaborating and sharing information to protect users, involving companies of different sizes and ages. It is critical that services do not inadvertently offer "safe havens" for offenders.

Question 26: Is there any additional information eSafety should consider in determining the Designated Internet Services Standard?

N/A

****

This submission is reflective of the views of the Secretariat of the WeProtect Global Alliance and does not necessarily represent the opinions and positions of any of its members.

For further information, please contact Shailey Hingorani, Head of Advocacy, Policy and Research, at WeProtect Global Alliance [redacted] *or Eleanor Linsell, Policy and Advocacy Manager, at WeProtect Global Alliance* [redacted].