TECH COALITION

Assessing Online Child Sexual Exploitation and Abuse Harms in Product Development

Table of contents

Background

Assessing Online Child Sexual Exploitation and Abuse (OCSEA) Harms in Product Development has been developed by the Tech Coalition to share considerations for how tech companies can evaluate and mitigate the risk of online child sexual exploitation and abuse (OCSEA) on their platforms.

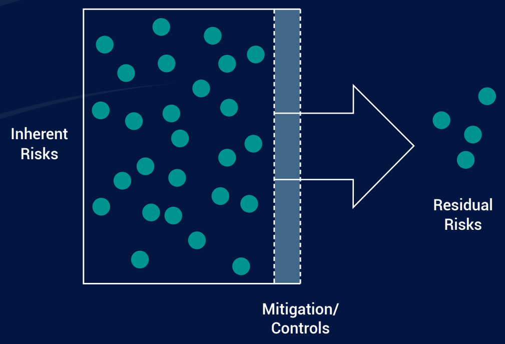
These considerations draw on the experience of Tech Coalition members and other currently available research and resources including 4C's of online risk[1], Age Assurance in the Digital World[2], the Australian eSafety Commissioner's Assessment tools[3], and the UN's Convention on the Rights of the Child[4].

In joining the Tech Coalition, member companies have demonstrated their commitment to combating OCSEA, and to their accountability for those efforts.

We hope this resource encourages all companies to embed the evaluation of potential OCSEA harms within the product development process so that together, as an industry, we can combat OCSEA.

Goals of this document

The goal of this document is twofold. First, to provide considerations to help companies assess and compare the *inherent risk*[5] of features, products and settings in regards to OCSEA. Second, to determine what and how many mitigations to put in place to reduce the *residual risk*[6] a given product or feature may pose to young people. For the purpose of this document, we will use the term "feature" to mean products, features or settings.



How to use this document

This document should be used to help encourage critical thinking and facilitate conversations during the product development process. The approach has four steps:

- Step One: Understand the specifics of the feature in question
- Step Two: Evaluate harm
- Step Three: Assess inherent risk
- Step Four: Consider mitigations and determine residual risk

Important Considerations

All companies are different: We understand all companies are different and may take different approaches to address their own unique set of features and risks. These recommendations are intended to help facilitate internal, cross functional conversations and do not intend to define a standard of care – you may wish to add or subtract from them as needed.

This was not developed to provide legal guidance or regulatory compliance: We did not overlay or compare this information with drafted legislation in regards to age assurance and safety by design.

Scope of this document is limited to evaluating risk of OCSEA: While this document could be used for other harm types (such as violence, self-harm or disinformation), this was out of scope for this exercise. To learn more about a wider range of harm types, consider reviewing The World Economic Forum's *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*[7].

Step One: Understand Feature Specifics

Evaluating features is typically not black and white; some aspects of a feature could be beneficial to a child and even enhance their safety – other aspects of a feature could be harmful. Before evaluating harm, it is necessary to understand the specifics of the feature. In order to do this, it can be helpful to have a recurring meeting between safety and feature teams – or having safety team members join product launch meetings – to understand what new features are in development.

When understanding the specifics of the feature, knowing the following will help with assessing potential harm:

- What are the different scenarios in which this feature could be used?
- Who might be the audience? Will children have access to the feature?
- What (content / personal information / etc) might be shared?
- What relationships or introductions might be facilitated?
- In what scenarios might this feature lead to less or more risk of harm?
- What next step might a child take due to this feature?
- What next step might a bad actor take due to this feature?

Step Two: Evaluate Harm

This section outlines each harm type and provides questions to help guide conversation when evaluating whether a feature may increase the risk of those online harms.

Harms are defined by leveraging the 4Cs[8]. The 4Cs is a classification that outlines online risks that may impact a child's safety, privacy, wellness and fair treatment. The classification states that these risks occur when a child:

- engages with and/or is exposed to potentially harmful content,
- experiences and/or is targeted by potentially harmful contact,
- witnesses, participates in and/or is a victim of potentially harmful conduct,
- is party to and/or exploited by a potentially harmful contract.

Contact Harms

On-platform contact

Risk of on-platform contact occurs when features facilitate introductions, or direct contact on the platform, between minors and potentially harmful people. These features may lead to child sexual exploitation, grooming, trafficking, sextortion, stalking, bullying or harassment.

Feature examples that may lead to on-platform contact include private (or semi-private) messaging, suggested friends / content, groups, visible friend lists, etc.

Potential questions to ask in order to assess whether a feature could encourage on-platform contact

Could this feature be used to...

- Introduce a child to someone they do not know?
- Promote content by a child in a public manner that could lead to new introductions?
- Build trust with a child by establishing common interests or issues?
- Have private conversations:
    - Which may encourage more harmful behavior by both minors and bad actors, and / or
    - Where there are no upstanders / alternate views which might make it harder to identify if something bad is happening or encourage the minor to get help.

Expose information that may be used for harmful purposes (Indirect Contact)

Risk of exposing information that may be used for harmful purposes occurs when a feature that may expose personally identifiable information can be used by bad actors to facilitate a crime such as child sexual exploitation, grooming, trafficking, sextortion, stalking, bullying or harassment. These harms may occur on-platform or lead to harm off-platform (either on another platform or in real life).

Feature examples that may lead to exposing information include visible friends lists, visible interest or group memberships; sharing specific PII or location information of the minor (such as location anchors), sharing information about an adult's access to a child (i.e bad actor using a vulnerable adult to get access to a child), and/ or sharing information about a child's likes or interests.

Potential questions to ask in order to assess whether a feature could expose information

Could this feature be used to...

- Share information about the child that can be used...
    - For extortion?
    - For building trust or grooming?
    - To locate the child in real life?
    - To find the child on other platforms?

Potential Considerations – additional features that may increase risk for Contact Harms

When assessing features for potential contact harms, consider how additional features, such as anonymity, encryption and ephemerality, payments and media uploads, provide benefits but may also increase the likelihood or severity of contact harms.

Anonymous interactions and unknown identities

Anonymous interactions can support at-risk minors who may feel more comfortable to discuss personal experiences if they are anonymous. However anonymous interactions may also encourage more risky or harmful behavior if a real name is not tied to the account. If harmful activity occurs, anonymous interactions may make it more difficult for the company to take action on the account (for example it may be difficult to identify all the bad actor's accounts) or external organizations (such as Law Enforcement) to follow up with legal process for evidence of abuse.

Ephemerality and encryption
Consider that design choices like ephemerality or encryption can have an impact on a feature's overall risk. For example, ephemerality can lessen the potential risk to a child as their content is harder to share and is therefore less likely to 'go viral.' On the other hand, ephemerality can also mean that the content may not be available in order to make a report to law enforcement or respond to a later subpoena.

Design choices like encryption may hinder the ability to proactively scan for child sexual abuse material, but may also provide protections to children's personal information that prevent other forms of potential identity exposure or exploitation.

Bad actors are aware of the design implications of different features and may direct victims to encrypted and/or ephemeral platforms for severe abuse and exploitation. For example: "Don't say that or send that image here. Let's move to [encrypted or ephemeral platform]." This is a common tactic in grooming and sextortion cases.

Payments
Payments can be useful between friends; however, they can also be used to facilitate financial sextortion, online trafficking / streaming and grooming by using quid pro quo (e.g. "I'll pay you X if you show me Y").

Media Upload / Consumption
Media Upload / Consumption (such as uploading photos or videos) has become a core part of an online user's experience when sending private messaging. However this is also the most common method used to send and receive CSAM, including self-generated CSAM used for sextortion.

Audio Messaging / Video Calling
Audio Messaging / Video Calling are also a core part of an online user's experience when sending private messaging. Audio messaging and Video Calling make it significantly more difficult for companies to identify, respond and preserve evidence when child abuse occurs on a company's platform. Bad actors are more frequently using video calling (i.e two-way communications with live video and audio) to conduct abuse because it's more difficult to detect. For example, a bad actor may say "Let's use [video calling company] to meet" while the users are messaging and then live abuse occurs on the video calling platform.

Content / Conduct Harms

Risky behavior

Risk of risky behavior occurs when a feature may encourage risk-seeking behavior or attention seeking. These features may encourage behaviors and activities that cause harm to young people, both as victims and perpetrators. For example – some activities may include bullying, sexting, revenge porn, trolling, threats and intimidation, peer pressure and loss of control of digital legacy/footprint.

Feature examples that may lead to risky behavior include features that show and reward engagement (number of followers / friends / likes / views / etc), or real-time interaction via livestreaming or community chat.

Potential questions to ask in order to assess whether a feature could encourage risky behavior

Could this feature…

- Encourage provocative or risky behavior that could cause harm to the child or other children?
  - Does this encourage a child to exhibit riskier behavior due to what they have seen in the media (internet, video games, television and music)?
- Encourage comparisons or competition for likes / views / subscribers / etc?

Harmful or inappropriate content

Risk of harmful or inappropriate content occurs when a feature could lead to viewing harmful content or content that may be unsuitable for minors such as pornography, violence, suicide / self-harm / eating disorder, disinformation, discrimination / hate speech. This content may be used to isolate and groom a child. Content, such as pornography, may desensitize a vulnerable minor into producing or sharing CSAM and also may be used to groom a child by creating "secrets" between the perpetrator and child.

Feature examples that may lead to harmful or inappropriate content include: suggested content, search, media content / upload and live streaming.

Potential questions to ask in order to assess whether a feature could encourage harmful or inappropriate content

Could this feature…

- Enable any user to create content that may be unsuitable for or harmful to minors?
- Enable a minor to view content that is unsuitable or harmful?

Potential Considerations – additional features that may increase risk for Content / Conduct Harms

When assessing features for potential content / conduct harms, consider how other features – such as anonymity – while providing benefits may also increase the likelihood or severity of contact harms.

Anonymous interactions and unknown identities

As noted in contact harms considerations, anonymous interactions can support at-risk minors who may feel more comfortable to discuss personal experiences if they are anonymous. However anonymous interactions may also encourage more risky or harmful behavior if a real name is not tied to the account. If harmful activity occurs, anonymous interactions may make it more difficult for the company to take action on the account (for example to identify all the bad actor's accounts) or external organizations (such as Law Enforcement) to follow up with legal process for evidence of abuse.

Contract Harms

Expose information that might be used for commercial purposes

Risk of exposing information that might be used for commercial purposes occurs when a feature may promote products / services that are harmful to minors, may encourage harmful behavior or may facilitate the purchase of items that are either illegal or unsuitable for minors.

Feature examples that may lead to exposing information for commercial purposes include geolocation data gathering, general PII data gathering, advertising, sharing personal data with 3rd parties and payments.

Potential questions to ask in order to assess whether a feature could expose information that might be used for commercial purposes…

Could this feature…

- Encourage harm through "word of mouth marketing" at scale by using names of friends in association with harmful acts or products?
- Promote illegal or harmful content, products or behaviors to children? For example (note that some of these examples are outside of OCSEA but sharing for illustrative purposes):
  - Could advertising be used for something OCSEA related (e.g. modeling, massage parlor, etc)?
  - Could the minor see harmful or illegal content (e.g. pornography, horror, violence, self-harm, suicide)?
  - Could this encourage the minor to purchase harmful, inappropriate or illegal products (e.g. drugs, cannabis dispensary, sex-related products)?
- Enable a child to click thru to a harmful business?
- Facilitate transactions of harmful products?

Additional notes: if a feature is gathering information about a minor…

- Is there a compelling safety reason to gather this information? For example – for location data, a parent may want to use geolocation to check on their child. However, sharing geolocation data can also have safety implications.
- Is there certain information that might be more harmful than others?
  - For example, if location information, what information is more risky and what information is less risky? Consider if specific information might be more risky (e.g. a child's school) than broad information (e.g. a large city)
- Note: while this is typically outside of OCSEA harms, it is worth noting that local laws and regulations may prohibit sharing user information (e.g. PII or location data) with 3rd parties or gathering data for minors (e.g. for advertising). Check with Legal and Policy teams for additional information.

Problematic overusage

Risk of exposing information that might lead to problematic overuse occurs when a feature may infringe on the minor doing other activities and/or encourages unhealthy habits or behavior.

Feature examples that may lead to problematic overuse include games, notifications, recommended content and scrolling.

Potential questions to ask in order to assess whether a feature could lead to problematic overusage

Could the feature…

- Create a fear of the minor missing out if they are not online?
- Create a fear of the minor missing out if they don't act now?
- Show that a person is available "right now" and/or when they were last online?
  - Could this be used by a bad actor to place demands on the child (e.g. sextortion – "you need to be online at this time and give me money")?

Step Three: Assess Inherent Risk

After a company has completed Step Two and evaluated the feature for potential harms, the next step is to aggregate this information to estimate potential inherent risk. To do this, some companies establish a scoring system by assigning scores.

Establish a Scoring System

For the purpose of this document, the scoring system includes three factors: Likelihood of Harm, Impact to Detecting OCSEA, and Minor Usage.

Likelihood of Harm

How likely is it that this feature will lead to a contact, content, conduct or contract harm(s)? Assign a score for each harm type depending on what is useful for your company's business (for example, some companies assign a score ranging from 0 = lowest risk to 3 = highest risk).

Impact to Detecting OCSEA

How likely is it that this feature will hinder a company's ability to detect CSAM / CSEA? This might be due to what was described in some of the harm sections as *Potential Considerations – additional features that may increase risk*. Assign a score depending on what is useful for your company's business (for example, some companies assign a score ranging from 0 = will not impact, to 3 = will impact all cases).

Minor Usage

How likely is it that this feature will be used by minors? Note that not all companies are able to determine minor usage. If a company doesn't have this information, consider using the general population statistic. Assign a score depending on what is useful for your company's business (for example, some companies assign a score ranging from 1 = no or low usage to 3 = high usage).

Estimate Inherent Risk Score

Using the above information, the company can estimate the inherent risk score by summing the likelihood of harms and impact to detect OCSEA and then multiplying that number by minor usage.

An example for illustrative purposes: a company evaluated a new feature and estimated the risk of a:

- Contact harm = 2
- Content harm = 2
- Conduct harm = 3
- Contract harm = 1

Yielding a likelihood of harm score of 8.

They then determined that the impact to detect OCSEA was minimal and estimated it to be 0 and that minor usage was high and estimated it to be 3.

The overall inherent risk score is therefore: (8+0)*3 = 24

Considerations for the Scoring System

As noted above, establish a scoring system that works best for your company. While this document used 0-3 – another company might find 0-5 or 0-10 more useful. The score of a feature on its own may be less important than establishing a system that can be used to compare inherent risk of features within the company's platform. Other ideas:

- A scoring system that ranges from 0-3 implies that a score of 0 is 25% less harmful than a score of 1. Does the company think this reflects how they may want to evaluate harm?
- Depending on a company's business model and product, a company may also want to weigh certain harms more heavily than others. For example, does a company want to double the weight of a contact harm or cases where OCSEA detection will be impacted?
- Does the impact of the harm vary by the age of the minor? If so, consider whether to have separate scoring for younger minors (13-15 years old) vs older minors (16-17 years old).

Think about and document any potential biases in the scoring. For example - what group is completing the assessment and how might their point of view bias the results? Are team members interpreting low risk or high risk differently? Is there a bias to previous risks that are less of any issue now? Also, reach out to other teams, especially teams who may have an alternate point of view, for additional input.

A scoring system is more effective if done on an ongoing basis. Keep evaluating and ask: what have we learned about how people are using this feature? Have we reduced the risk? What new or emerging harms have surfaced?

Step Four: Consider Mitigations and Assess Residual Risk

Once the inherent risk score is determined, a company can assess how to lower that risk score by adding mitigations. In some cases, it might be possible to lower the risk but it might be difficult to eliminate the risk entirely.

When assessing mitigations, determine how much risk might be removed by implementing the mitigation measures. This will be company specific however consider:

- Will the mitigation remove some of the risk by providing a means for the user to get support or report users?
- Will the mitigation remove more of the risk by proactively identifying, to high accuracy, potential bad actors or harmful content?
- Will the mitigation remove the risk entirely?

Once the company has determined mitigations, they can determine how this may decrease the inherent risk and therefore understand the residual risk. Keep in mind that residual risk scores also depend on a company's confidence level of whether youth can access the feature. Also consider reviewing the efficacy of mitigations on an ongoing basis in order to ensure they are working effectively and not being circumvented. Below are potential mitigation measures to consider. For more detailed resources and discussions about mitigation measures, please contact the Tech Coalition to discuss **membership**.

Policies

Update internal and external policies and statements, such as terms of service and/or community guidelines, to explicitly communicate that online child sexual exploitation and abuse material is prohibited. Policies can also state the action the company will take if OCSEA is identified. Also review policies for advertising, promoting of 3rd party links and data gathering to ensure policies and company products / features / settings are in alignment with regulation and industry standards. Companies may want to consider adjusting these policies, and what is allowed, by user's age.

User Reporting

Promote in-product user reporting for harmful content. Young people are more likely to report if they know reports are anonymous. Remind users about reporting options by providing in-product education especially at higher risk moments. For example, remind users about reporting options after a user has blocked another user (e.g. some young people may block, but not report, potential abusers out of fear of social repercussions).

Detection

Run proactive detection to identify suspicious activity, detect real-time potential harm in images, videos and live videos using human review, and develop machine learning classifiers, computer vision or other proactive and reactive mechanisms.

Transparency

Create and continue to enhance a company's transparency report. Transparency reports build trust, educate the public about potential harms and demonstrate accountability by providing information about a company's efforts to combat online child sexual exploitation and abuse. See the Tech Coalition's TRUST Framework[9], developed by Tech Coalition Members, for additional guidance.

In-Product User Education

Educate users about safety risks in higher risk places or higher risk moments.

Tools to Encourage Conscious Usage

Provide tools that encourage users to evaluate their usage and make conscious decisions to limit problematic usage. For example, establish screen limit times to encourage breaks and limit use per day, provide a dashboard and weekly recap of how much time is spent on the platform and/or provide an option to mute notifications.

Adjust Feature Usage or Settings based on Age and/or Online Behavior

Develop capabilities and solutions that adjust products, features and settings by online behavior and / or age. For example, if a user exhibits suspicious behavior or incurs a policy violation, consider limiting the user's access to that feature. If a company has an age estimation or verification solution in place, the company may consider tying feature usage or default settings to the user's age. For example, a company may require users to be above a certain age in order to use a feature or for users below a certain age to have more restrictive settings. When assessing which features to restrict by age, some companies may look at the likelihood for harm and minor usage scores (as discussed in Step Three).

Development & Review of the Document

This document was created by members of the Tech Coalition. It incorporates member companies' experiences and references resources and research by other entities as outlined in the resource.

The Tech Coalition acknowledges all the stakeholders who took the time to engage and to provide feedback.

The Tech Coalition will continue to receive feedback on this document and update it at regular intervals thereafter, to ensure it is keeping pace with technological and other developments.

Footnotes

[1] Livingstone, S. and Stoilova, M. (2021) 4 Cs of online risks.
https://core-evidence.eu/posts/4-cs-of-online-risk

[2] 5 Rights Foundation. (2021) But how do they know it is a child? Age Assurance in the Digital World
https://5rightsfoundation.com/uploads/But_How_Do_They_Know_It_is_a_Child.pdf

[3] Australian eSafety Commissioner. Assessment tools,
https://www.esafety.gov.au/industry/safety-by-design/assessment-tools

[4] United Nations (1989), Convention on the Rights of the Child,
https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child

[5] Computer Security Resource Center, Inherent risk, National Institute of Standards and Technology,
https://csrc.nist.gov/glossary/term/inherent_risk

[6] Computer Security Resource Center, Residual risk, National Institute of Standards and Technology,
https://csrc.nist.gov/glossary/term/residual_risk

[7] World Economic Forum. (2023), Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms,
https://www.weforum.org/reports/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/

[8] Livingstone, S. and Stoilova, M. (2021) 4 Cs of online risks.
https://core-evidence.eu/posts/4-cs-of-online-risk

[9] Tech Coalition. (2022), TRUST: Voluntary Framework for Industry Transparency,
https://www.technologycoalition.org/knowledge-hub/trust-voluntary-framework-for-industry-transparency

TECH COALITION

About Tech Coalition

The Tech Coalition facilitates the global tech industry's fight against the online sexual abuse and exploitation of children. We are an alliance of technology companies of varying sizes and sectors that work together to drive critical advances in technology and adoption of best practices for keeping children safe online. The Tech Coalition convenes and aligns the global tech industry, pooling their knowledge and expertise, to help all our members better prevent, detect, report, and remove online child sexual abuse content. This coalition represents a powerful core of expertise that is moving the tech industry towards a digital world where children are free to play, learn, and explore without fear of harm.

To learn more visit www.technologycoalition.org