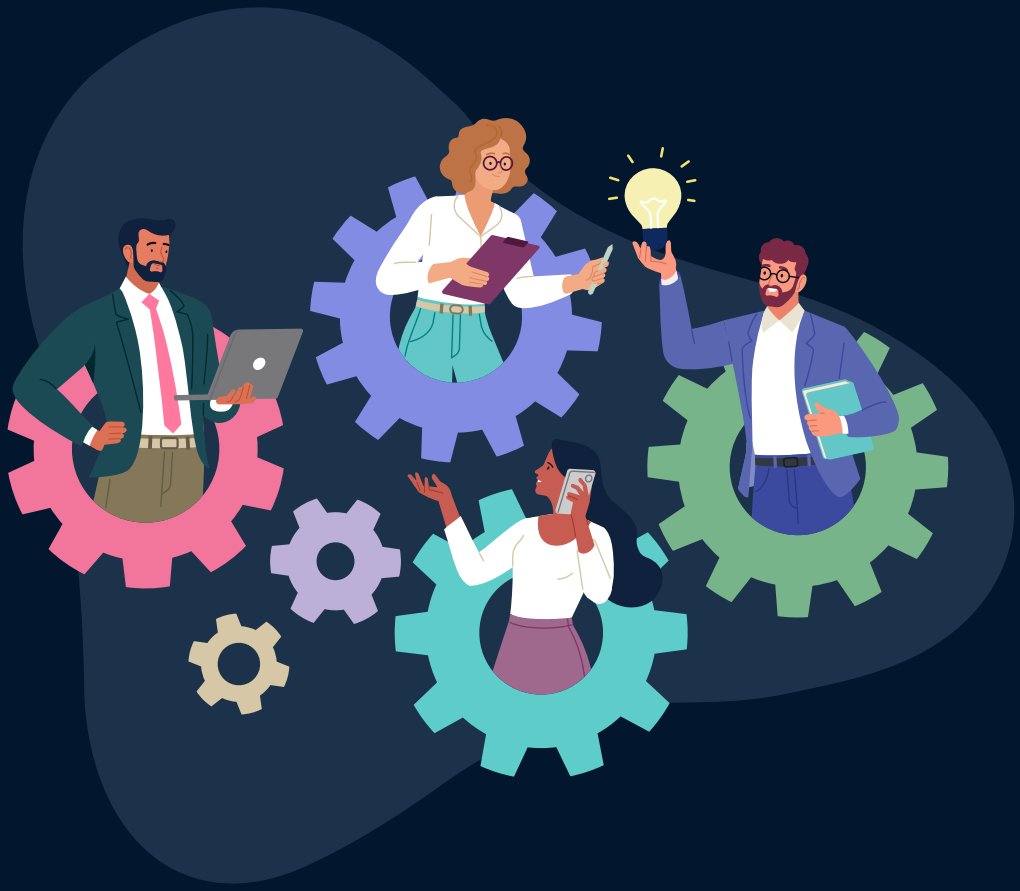
March 2026

Designing for Safety:
Preventing child sexual exploitation and abuse online

A toolkit for implementing Safety by Design strategies



Australian Government
eSafety Commissioner
eSafety.gov.au

Contents

Content warning: This toolkit discusses child sexual exploitation and abuse. Please take care when reading and consider whether this report is right for you at this time.

Foreword

Safety by Design[1] is an essential approach to addressing online harms, especially those that affect children. Across the technology sector, some of the most innovative and dedicated minds are working to embed safety into the very fabric of their platforms. Their work is not only commendable, it's critical.

Using active Safety by Design practices to combat online child sexual exploitation and abuse (CSEA) is an important step forward toward providing solutions and building upon best practice. The foundational pillars of Safety by Design – service provider responsibility, user empowerment, and transparency and accountability – ensure that the burden of safety does not fall on those most at risk, especially children. While regulation is in force in many countries to combat online CSEA, Safety by Design helps industry meet their obligations and strives to raise safety standards beyond what is required by regulators.

Ultimately, the person responsible for the harm from CSEA is the perpetrator. But we must also confront how technology is being weaponised and the role the online industry can and should play in preventing its proliferation. We call on technology companies to take these actions:

- **Embed Safety by Design principles** into core operations at the outset
- **Prioritise children's rights and safety** at every stage of the product lifecycle
- **Share tools, insights and innovations** to raise safety standards across industry
- **Invest in education and training** to build safety into product development from the start
- **Collaborate across platforms and share signals** to close the gaps exploited by perpetrators
- **Support smaller and emerging companies** to design safe, inclusive products from the outset
- **Work with governments and regulators** to build a shared understanding of harms and drive proactive solutions.

Together, we can build a digital world where children are safer, empowered, and free from child sexual exploitation and abuse online. The time to act is now.

[1] See Appendix 1: Terminology – Safety by Design is an initiative that eSafety has driven since 2018. Safety by Design seeks to shift the onus of safety from the user to service providers, encouraging providers to promote and embed safety at all stages of the technology product lifecycle.

About this toolkit

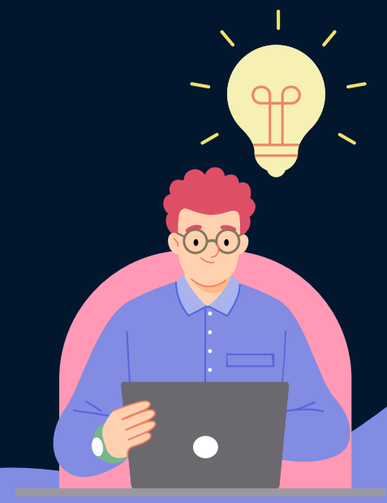
In December 2024, eSafety and the United States Department of Homeland Security co-hosted a two-day workshop with more than 20 organisations from the technology, non-government, academic and civil society sectors, across policy, trust and safety and technical domains. The goal: to share current best practice and establish new areas of collaboration to combat child sexual exploitation and abuse online.

This toolkit is an output from that workshop in the United States. It draws on current industry practices and highlights examples of good practice and innovative Safety by Design approaches. It includes five practical tools and a list of resources to help organisations of all sizes place child safety and rights at the centre of their design and development processes. It builds upon existing resources, developed in close consultation with a broad range of stakeholders across the technology ecosystem, available on eSafety's website.

The toolkit does not shy away from the challenges. It explores some of the common barriers to meaningfully adopting Safety by Design and offers strategies to overcome them. While implementing Safety by Design can seem complex, this toolkit provides the guidance to make it achievable – whether you're building from scratch or enhancing existing capabilities to prevent CSEA.

The tools will improve your capability throughout the product lifecycle:

- **Tool A:** Applying Safety by Design principles in the prevention, disruption and detection of CSEA
- **Tool B:** A checklist to help identify CSEA risk factors across your product, service or features
- **Tool C:** Common interventions to prevent, detect and disrupt CSEA across the tech industry
- **Tool D:** Guidance for start-ups: Identifying and combatting CSEA
- **Tool E:** User reporting: good practice guidelines



Note on terminology

Child sexual exploitation and abuse (CSEA) includes both:

- child sexual exploitation – material and activity that sexualises and exploits a child, but may not involve sexual abuse
- child sexual abuse – activity which involves sexual assault against a child (a subset of child sexual exploitation).

Child sexual abuse material (CSAM) refers to any representation - regardless of format – of a child engaged in real or simulated explicit sexual activity, or a representation of a child's sexual parts for primarily sexual purposes.

Throughout this toolkit we use the term 'child' to mean any individual under the age of 18.

See Appendix 3: Terminology for more guidance.

Who is this toolkit for?

This toolkit is designed for companies of all sizes – from early-stage start-ups to large-scale platforms – looking to adopt best-practice and innovative Safety by Design approaches to combat CSEA.

CSEA can occur across a wide range of digital services, often in ways that don't involve direct contact or explicit content. Whether you're building a messaging app, a gaming platform, a fintech tool, or a social network, understanding and addressing CSEA risks is essential to protecting your users and your platform.

Acknowledgements

eSafety would like to acknowledge contributions from:

- **Adobe**
- **All Tech is Human**
- **Apple**
- **Block Inc.**
- **Catherine Fitzpatrick, Flequity Ventures**
- **Common Sense Media**
- **Discord**
- **DoorDash**
- **Family Online Safety Institute**
- **GitHub**
- **Google**
- **Humane Intelligence**
- **Internet Watch Foundation**
- **k-ID**
- **Linktree**
- **Match Group**
- **Meta**
- **Microsoft**
- **Modulate**
- **National Centre for Missing and Exploited Children**
- **OpenAI**
- **Roblox**
- **ROOST**
- **Snap**
- **Stability AI**
- **Technology Coalition**
- **Thorn**
- **Vyanams Strategies (Vys)**
- **Westpac**
- **X Corp**

Understanding the problem

An estimate, published in 2024, suggests that at least 300 million children – about 8.1% of the global child population – are subjected to CSEA online each year.[2]

These figures are likely to be underestimates, with most of the iceberg hidden beneath the cold, dark surface. CSEA is significantly underreported, often hidden, and global data remains fragmented. Prevalence data is especially limited in regions with large child populations, such as the Middle East, North Africa and South and East Asia.[3]

In Australia, a survey of individuals aged 16 to 24 years found that:

- 7.6% had experienced the non-consensual sharing of a sexual image of themselves before turning 18.[4]
- 17.7% had experienced online sexual solicitation by an adult before the age of 18.

US reporting data from 2024 further illustrates the scale of the problem:

- INHOPE's worldwide network of hotlines processed nearly 2.5 million media files of suspected child sexual abuse material (CSAM) from reported URLs – a 218% increase on the previous year.[5]
- NCMEC's CyberTipline received more than 20 million reports from electronic service providers, involving more than 29 million suspected incidents of child sexual exploitation, and about 165,000 reports from members of the public.[6]
- Over 51,000 of these reports were escalated to law enforcement agencies as urgent or involving a child in imminent danger.

In the 2024-25 financial year, the Australian Centre to Counter Child Exploitation (ACCCE) Child Protection Unit received 82,764 reports of child sexual exploitation online, an average of 226 reports per day, and a 41% year-on-year increase.[7]

[2] WeProtect Global Alliance, 2024, 'World's first estimate of the scale of online child sexual exploitation and abuse', accessed 27 May 2025, URL: https://www.weprotect.org/blog/worlds-first-estimate-of-the-scale-of-online-child-sexual-exploitation-and-abuse Note: estimates of the prevalence of child sexual exploitation and abuse (CSEA) vary depending on the methodology used, the specific forms of CSEA examined, and the availability of reliable data.
[3] Fry D, Krzeczkowska A, Ren J, et al., 2025, 'Prevalence estimates and nature of online child sexual exploitation and abuse: a systematic review and meta-analysis', The Lancet Child & Adolescent Health, vol. 9, no. 3, pp. 184–193, accessed 1 June 2025, URL: https://www.researchgate.net/publication/388235447_Prevalence_estimates_and_nature_of_online_child_sexual_exploitation_and_abuse_a_systematic_review_and_meta-analysis
[4] Walsh K, Mathews B, Parvin K, Smith R, Burton M, Nicholas M, Napier S, Cubitt T, Erskine H, Thomas H J, Finkelhor D, Higgins D J, Scott J G, Flynn A, Noll J, Malacova E, Le H & Tran N, 2024, 'Prevalence and characteristics of online child sexual victimization: Findings from the Australian Child Maltreatment Study', Child Abuse & Neglect, Version of Record 6 December 2024, accessed 30 January 2025, URL: https://www.sciencedirect.com/science/article/pii/S0145213424005799
[5] INHOPE, 2024, 'Webinar Recap: The INHOPE Annual Report 2024', accessed April 12 2025, URL: https://inhope.org/articles/webinar-recap-the-inhope-annual-report-2024
[6] National Center for Missing & Exploited Children, 2024, 'CyberTipline Data', accessed 1 December 2025, URL: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

Victim characteristics vary widely and are shaped by a complex mix of factors, including:

- geographic location
- type of abuse and material
- victim's age
- method of data collection
- the role of evolving technologies.[8]

These variables affect how victims are represented in the data. However, evidence consistently shows that girls are disproportionately affected by CSEA. For example:

- In 2024, 98.7% of CSAM cases reported to INHOPE involved girls compared to 0.76% involving boys.[9]
- A 2025 meta-analysis found that 8.7% of girls had experienced CSEA in the past year, compared to 7.5% of boys.[10]
- A recent Australian study found that girls were more likely than boys to experience non-consensual intimate image sharing (10.9% of girls compared to 3.8% of boys) and online sexual solicitation by an adult (26.3% of girls compared to 7.6% of boys) before age 18.[11]

Emerging trends point to increasing risks for boys in specific contexts.

For example, the Internet Watch Foundation (IWF) reported a 25% rise in reports of self-generated CSAM coerced from boys between 2021 and 2022.[12] In addition, Australian research on sexual extortion of Australian adolescents published in 2025 shows that boys are more likely than girls to be sexually extorted (74.4% vs 58.8%).[13]

[7] Australian Centre to Counter Child Exploitation (ACCCE), 2025, '41-cent jump in online child sexual exploitation reports underscores need for whole-community approach', accessed 8 September 2025, URL: https://www.accce.gov.au/news-and-media/news/41-cent-jump-online-child-sexual-exploitation-reports-underscores-need-whole-community-approach
[8] ECPAT International, 2018, Trends in Online Child Sexual Abuse Material, p. 13, accessed 9 April 2025, URL: https://ecpat.org/resource/trends-in-online-child-sexual-abuse-material/
[9] INHOPE, 2024 'INHOPE Annual Report 2024', accessed April 12 2025, URL: https://inhope.org/articles/inhope-annual-report-2024
[10] Fry D, Krzeczkowska A, Ren J, Lu M, Fang X, Into the Light Index Study Group, 'Prevalence estimates and nature of online child sexual exploitation and abuse: a systematic review and meta-analysis', Child Abuse & Neglect, ScienceDirect website, 2024.
[11] Walsh K, Mathews B, Parvin K, Smith R, Burton M, Nicholas M, Napier S, Cubitt T, Erskine H, Thomas H J, Finkelhor D, Higgins D J, Scott J G, Flynn A, Noll J, Malacova E, Le H & Tran N, 2024, 'Prevalence and characteristics of online child sexual victimization: Findings from the Australian Child Maltreatment Study', Child Abuse & Neglect, Version of Record 6 December 2024, accessed 30 January 2025, URL: https://www.sciencedirect.com/science/article/pii/S0145213424005799
[12] Internet Watch Foundation, 2022, 'Analysis by Sex', Annual Report 2022, accessed 30 January 2025, URL: https://annualreport2022.iwf.org.uk/trends-and-data/analysis-by-sex/
[13] eSafety Commissioner, 2024, 'Sexual extortion of Australian adolescents', accessed 3 February 2025, URL: https://www.esafety.gov.au/research/sexual-extortion-of-australian-adolescents. See also: Thorn, 2024, 'Trends in Financial Sextortion: An investigation of sextortion reports in NCMEC CyberTipline data', accessed 3 February 2025, URL: https://www.thorn.org/research/library/financial-sextortion

The impact on victim-survivors

Understanding the impact of CSEA on victim-survivors – and prioritising their voices and recovery – is critical to combatting online and offline forms of harm. It is essential that the lived experience of victim-survivors guides policy, product design, and safety decisions.

Expert insights and resources on victim-survivor experiences are available at Protect Children's Experiences of Adult Survivors of CSEA Report.[14]

Who are the perpetrators?

There is no agreed and definitive profile of perpetrators of CSEA. Much of what we know about them relies heavily on data from people apprehended by law enforcement agencies or who participated in anonymous surveys – typically in English-speaking or other high-income OECD countries. As a result, these findings may not represent the full range of perpetrator characteristics or apply globally.

A common misperception is that only individuals with a sexual interest in children – such as paedophiles – engage in CSEA. In reality, motivations can vary and may include:

- sexual interest in children
- emotional self-regulation
- desensitisation to adult pornography
- financial gain (for example, in cases of financial sexual extortion).[15]

This complexity makes it difficult for the tech industry to define a standard perpetrator profile or reliably associate specific user traits with risk. Instead, platforms must focus on understanding and detecting behaviours and signals that indicate a heightened risk of CSEA on their services.

It is also important to acknowledge that CSEA is not always perpetrated by adults. For example, law enforcement agencies in England and Wales report that about half the CSEA offences they investigate are committed by children – often in the form of child-on-child or peer-to-peer abuse, including online incidents[16] and in Australia in 2024 there were several reports of AI-generated CSAM having been created and shared by students targeting classmates.[17]

[14] Protect Children, 2024, 'Experiences of Adult Survivors of Child Sexual Abuse and Exploitation Across 29 Languages', accessed 3 February 2024, URL: https://www.suojellaanlapsia.fi/en/post/our-voice-global-report
[15] Project 2Know, What Drives Online Child Sexual Abuse Offending?, Children's Rights Digital website, p. 8, accessed 30 January 2026, URL: https://www.childrens-rights.digital/background/detail/2know-report-what-drives-online-child-sexual-abuse-offending
[16] Vulnerability Knowledge and Practice Programme (VKPP), 2023, National Analysis of Police Recorded Child Sexual Abuse and Exploitation (CSAE) Crimes Report 2023, accessed 30 January 2026, URL: https://news.npcc.police.uk/releases/second-child-sexual-abuse-and-exploitation-analysis-launched
[17] McKibbin G, Isobe J & Kuruppu J, 2024, 'Big tech and porn platforms are the real perpetrators behind "sickening" deepfakes', Pursuit (University of Melbourne), 14 June 2024, accessed 10 June 205, URL: https://pursuit.unimelb.edu.au/articles/big-tech-and-pornography-are-the-real-perpetrators-behind-sickening-deepfakes

How are online platforms and services used to facilitate CSEA?

CSEA can be facilitated by almost any online platform or service. It does not always involve direct contact or sharing of material. Children are digitally connected from a young age and move fluidly between online and offline environments. CSEA often spans this continuum.

Perpetrators are constantly evolving their tactics – adapting to new technologies and finding ways to use existing systems to produce, coordinate and distribute CSEA. These are some of the most common methods:

- **CSEA-themed memes**
  Creating and sharing memes that depict or reference CSEA – often framed as humorous, edgy or provocative. These may be circulated without malice (for example, a teenager sharing a meme containing a sexual image of a peer) but still constitute harmful behaviour.
- **Grooming**
  Establishing contact and building relationships with children (or their caregivers) online, with the intent of sexually exploiting the child.
- **Link sharing**
  Posting or exchanging links to third-party hosted CSEA content across messaging platforms and forums.
- **Livestreaming CSEA**
  Broadcasting exploitation or abuse of children in real time to one or more viewers.
- **Online child sex trafficking**
  Using digital platforms to recruit, harbour, transport or receive children for the purpose of sexual exploitation.
- **Organised networks**
  Criminal or extremist groups using digital tools to coordinate large, often transnational operations. Motivations vary and may include shared interest in CSEA or in nihilistic violent ideologies (see 'sadistic online exploitation' on page 14).[18]
- **Sexual extortion ('sextortion')**
  A form of online blackmail in which someone tricks or coerces a victim into sending sexual images or videos of themselves and then threatens to share it with others unless the victim meets their demands.

[18] Institute for Strategic Dialogue (ISD), 2025, 'Terror without ideology? The rise of nihilistic violence – An ISD Investigation', accessed 9 May 2025, URL: https://www.isdglobal.org/digital_dispatches/terror-without-ideology-the-rise-of-nihilistic-violence-an-isd-investigation/

- **Sexualisation of children through digital content**
  The creation or sharing of real or synthetic visual content, text-based stories, or other material that sexualises children. Tools such as game engines may be used to create young, sexualised characters.
- **Sharing CSAM**
  Distributing images or videos through direct messaging, group chats, forums, or other online communication features. End-to-end encryption is often exploited to conceal this activity.

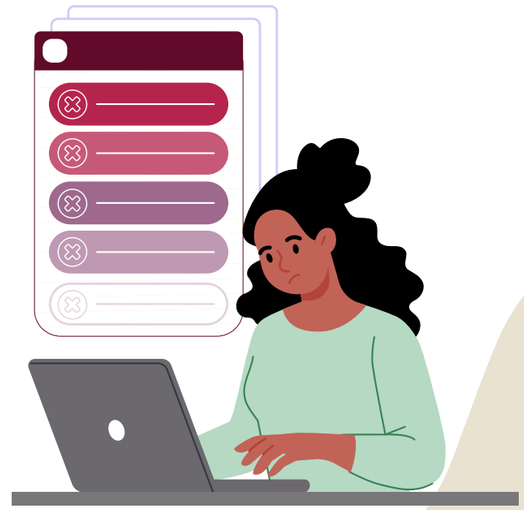
More recently, law enforcement, hotlines and other organisations have reported a rise in CSEA occurring through novel mechanisms, including these:

- **AI-generated CSAM (AIG-CSAM)**
  Generative AI technologies are increasingly exploited to create synthetic child sexual abuse material – often referred to as deepfakes or digital forgeries – featuring CSEA content of both real and AI-generated children.[19]
- **Abuse via extended reality (XR)**
  This includes virtual reality (VR) and augmented reality (AR). While research into CSEA in XR environments is still limited, perpetrators may use avatars to misrepresent themselves and groom children in immersive spaces. XR applications can also be used to share and distribute CSAM, particularly in app stores with looser content moderation policies.[20] eSafety's **immersive technologies position statement** provides more information on the risks and benefits of XR and outlines relevant Safety by Design measures.
- **Invite child abuse pyramid (ICAP) sites**
  Illegal, invite-only commercial platforms designed for the exchange, sale and reproduction of CSAM. Access to more extreme material is often granted by sharing invitation links with new users, creating a tiered system of abuse.[21]

[19] eSafety Commissioner, 2025, 'Generative AI and child safety: A convergence of innovation and exploitation', accessed 11 June 2025, URL: https://www.esafety.gov.au/newsroom/blogs/generative-ai-and-child-safety-a-convergence-of-innovation-and-exploitation
[20] WeProtect Global Alliance, 2023, Extended Reality Technologies and Child Sexual Exploitation and Abuse, accessed 2023, URL: https://www.weprotect.org/resources/library/extended-reality-technologies-and-child-sexual-exploitation-and-abuse/
[21] INHOPE, 2024 'INHOPE Annual Report 2024', accessed April 12 2025, URL: https://inhope.org/articles/inhope-annual-report-2024

- **Sadistic online exploitation**
  Violent online groups – described by the US Federal Bureau of Investigation as nihilistic violent extremist networks – use digital platforms to encourage children to harm themselves and others. Perpetrators may coerce victims to produce CSAM, sexually abuse other children (including siblings), harm animals, commit murder or even take their own lives.[22]
- **Self-generated CSAM**
  Children are increasingly creating and sharing sexually explicit content of themselves for various reasons. In some cases, the material is initially consensually shared (for example, between peers in a relationship) but later distributed without consent. In other instances, the content is produced under coercion or enticement by adults, including through grooming, sexual extortion, or commodified sexual interactions where perpetrators offer money or social opportunities in exchange for CSAM.[23]

Online forums were the most commonly used platform for sharing CSAM in 2024, according to INHOPE's annual report.[24]



[22] United States Attorney's Office, 2025, 'Leaders of 764 arrested and charged for operating global child exploitation enterprise', accessed 10 May 2025, URL: https://www.justice.gov/usao-dc/pr/leaders-764-arrested-and-charged-operating-global-child-exploitation-enterprise and National Center for Missing & Exploited Children, 2024, 2024 CyberTipline Report, accessed 10 May 2025, URL: https://ncmec.org/gethelpnow/cybertipline/cybertiplinedata
[23] Thorn, 2025, 'Commodified sexual interactions: How money and clout are obscuring risk for youth online', accessed 30 April 2025, URL: https://www.thorn.org/blog/commodified-sexual-interactions-how-money-and-clout-are-obscuring-risk-for-youth-online/
[24] INHOPE, 2024 'INHOPE Annual Report 2024', accessed 12 April 2025, URL: https://inhope.org/articles/inhope-annual-report-2024

It is important to remember that platforms and services can be exploited to facilitate CSEA at scale – either indirectly or in conjunction with other platforms. These are some examples:

- **Communications platforms:** Providing forums, group chats or similar spaces where perpetrators can communicate, share tactics and collaborate.
- **Financial services:** Enabling payments between perpetrators for CSEA-related content or commodified sexual interactions with children.
- **Internet infrastructure services:** Including cloud storage and hosting platforms used to store CSAM or support its creation – for example, by hosting AI models designed to generate abuse content.
- **Link-sharing platforms:** Hosting and sharing links to CSAM or other CSEA-related content.
- **Link shortening services:** Enabling perpetrators to mask URLs and circumvent content detection methods.
- **Machine learning model platforms:** Hosting versions of generative AI models that have been altered to bypass default safeguards, allowing them to produce CSAM.

No platform is immune. From financial services to game engines, AI models to web infrastructure – almost any digital service can be exploited to facilitate CSEA. Recognising this risk is the first step toward implementing a robust Safety by Design approach.

Case study

Westpac – Tackling sexual extortion risks in the financial sector

Banking services can become conduits for harm in cases of financially motivated CSEA. Recognising this risk, Westpac took steps to address the growing issue of sexual extortion.

Westpac in action

In March 2024, Westpac's monitoring controls flagged a suspicious transaction involving a teenage customer. The customer had shared intimate images with an individual online and was coerced into transferring $500 under threat. The alert triggered intervention by Westpac's Fraud and Scams Operations team, who engaged directly with the victim. Through empathetic customer engagement, Westpac uncovered details of the social media interactions that led to the sharing of the intimate images.

Further investigation revealed a systemic vulnerability: sexual extortion facilitated via internet banking and ATM features. Additional potential victims were identified, linked to the same product feature and suspected offenders.

Closing the gap

Westpac responded by disabling the relevant banking features for children, significantly reducing risk. The case also informed broader changes to product governance. Safety by Design principles were embedded into development processes, ensuring new and existing features undergo rigorous risk assessments to protect vulnerable or at-risk customers.

The impact of collaboration

Recognising the industry-wide nature of sexual extortion risks, Westpac shared key learnings with peer institutions. This collaborative approach strengthened collective harm prevention strategies and amplified the impact of individual interventions.

Safety by Design

This case reinforces that no platform is immune to CSEA. Westpac's proactive stance and robust Safety by Design approach show that broader systemic change can be supported by industries beyond online platforms and services.

Industry responses to CSEA

Collaboration and knowledge sharing

Stakeholders across the technology, government, academic and civil society sectors have committed to combatting CSEA through some notable coalitions and initiatives. These are some examples:

Table 1: Example initiatives to counter CSEA

Developer Good Practices: Combating Online Child Sexual Exploitation and Abuse Assessing OCSEA Harms in Product Development Trust: Voluntary Framework for Industry Transparency	**Tech Coalition**
Safety by Design for Generative AI: Preventing Child Sexual Abuse	**Thorn, All Tech is Human**
Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse	**Five Country Ministerial**
AI and Child Protection: A Collaborative Approach to a Safer Future	**Safer AI for Children Coalition**
Combatting Online Child Sexual Exploitation and Abuse in Gaming	**Digital Thriving**
Real Time Threats: Analysis of Trust and Safety Practices for Child Sexual Exploitation and Abuse (CSEA) Prevention on Livestreaming Platforms	**Centre for Democracy and Technology**
Combating the Sexual Exploitation of Children for Financial Gain – Financial Crime Guide	**AUSTRAC**
CyberTipline Reporting API Technical Documentation	**NCMEC**

Platform-level interventions

At the product, service or feature level, technology companies use a range of technology-led tactics to detect and identify CSEA. Many of these tactics are commonly deployed by companies to help fight and prevent CSEA. They include these:

- **Age assurance**
  Age assurance measures help deliver age-appropriate experiences. This may include one or more approaches, such as biometric age estimation (for example, facial analysis), AI tools that assess behavioural patterns, or age verification against external sources like credit card or ID data.
- **AI-driven text descriptions of suspected CSAM**
  Tools that detect CSAM by analysing the language used to describe the material (for example, written captions, text-based metadata, or language embedded in audio content).
- **Authentication or verification requirements**
  These require users to authenticate or verify their identity through methods such as linking a credit card, phone number or social media account to the platform or service.
- **Classifiers**
  Machine learning classifiers play a vital role in detecting potential CSAM across various content types. File name and path classifiers assess risk based on naming conventions without opening files. Image classifiers analyse pixel data to identify ‘known’ (previously identified) and new CSAM, including through nudity detection.. Text classifiers categorise open-ended content like messages and posts to flag grooming behaviours and other high-risk interactions. Similarly, video classifiers evaluate visual content in streams to detect and track concerning activities, helping surface previously unknown CSAM.
- **Content moderation tools (general)**
  This is software that provides content moderation capabilities for detecting, disrupting, and responding to CSEA. These tools typically integrate multiple components – such as classifiers, filters, and reporting workflows – into a single system to support end-to-end moderation.
- **Device-level interventions**
  These are technical measures that operate at the device level to prevent the creation or consumption of CSEA. For example, SafeToNet’s HarmBlock can be embedded into a device’s operating system to block explicit content, including CSEA, from being displayed on-screen or captured by the device’s camera. Human Mobile Design (HMD) has partnered with SafeToNet to roll out HarmBlock in its FusionX1 model.

- **Direct message interventions**
  These include prompts such as ‘Are you sure you want to send this?’ when suspected CSEA material is being shared in direct messages. These prompts introduce a moment of hesitation that can interrupt and prevent harmful behaviour. This is particularly relevant in cases involving child-on-child CSEA.
- **Hash-matching**
  A form of digital ‘fingerprinting’ that allows CSEA (or other) photos and videos to be rapidly compared against a database of previously identified (‘known’) material so they can be removed without being viewed. Databases include the Internet Watch Foundation (IWF) image hash list or NCMEC’s CSAM and exploitative file hash lists.

  Different types of hashing serve different purposes. Cryptographic hashing detects exact matches while perceptual hashing can identify visually similar content (such as resized, cropped or rotated images). Hashing techniques can also be applied to other types of multimedia content.

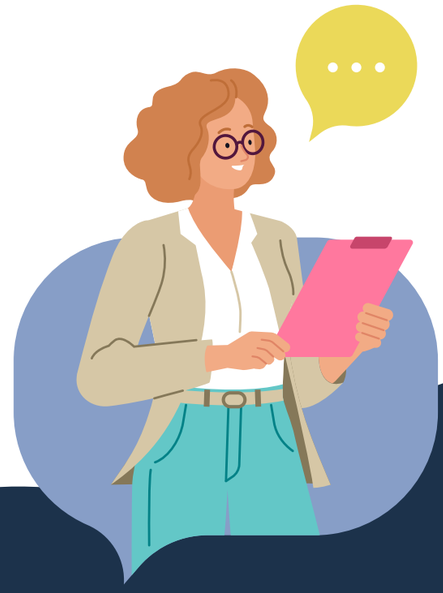
  Examples of image hashing technologies include (but are not limited to): PhotoDNA, MD5, Internal Hash, Thorn’s Safer Hash, PDQ, SHA1/SHA2, and CLIP.
- **Keyword detection**
  Tools that scan text against databases of keywords associated with CSEA – such as those from the IWF keyword list – and flag matches for review or action by the platform or service.
- **Parental tools**
  Tools that enable parents and carers to link to their children’s accounts and manage privacy and safety settings. These help provide an age-appropriate level of visibility into their child’s activity (for example, offering greater oversight for younger children and more autonomy for older teenagers).
- **Redirection through pop-up messages and chatbots**
  Tools that detect CSEA-related activity – such as searching for CSAM or attempting to access URLs previously removed for hosting CSAM – and automatically respond with deterrents. These may include pop-up messages or chatbots that direct users to help-seeking resources.

- **Safety/privacy by default measures for children**
  Child accounts can be set to the highest privacy and safety settings by default (for example, blocking messages from unknown senders or requiring approval to receive messages from someone who is not a contact). These measures are most effective when combined with robust age assurance.
- **URL detection**
  Tools that detect and compare URLs against lists of known CSEA-hosting websites (for example, the IWF URL list). Matches are flagged for appropriate action by the platform or service.
- **Link analysis (financial transactions and networks)**
  A fraud detection technique used by financial service institutions to analyse payment activity and identify patterns of potentially violative activity, such as fraud, money laundering, or sexual exploitation.

  Not every technology company may be set up to deploy these tactics. Some measures may not apply to all sectors, services, products or platforms for a range of reasons, including these:

  - **Legislative context:** Different legal and regulatory schemes may apply depending on a provider’s jurisdiction and service type.
  - **Service type:** The trust and safety approaches of a social media platform may differ significantly from those of a financial services provider.
  - **User age range**: Platforms may be designed for children of various ages (under 18), specifically for teenage children (13 to 17), or for adults (18+). Regardless of intended users, providers must account for the likelihood that younger users may access their services.

Whatever the approach, interventions must be implemented with care – ensuring that human rights, privacy and security are safeguarded at every step.



Common barriers to addressing CSEA

Despite there being many tactics available to address CSEA, significant barriers still hinder the implementation of effective safety measures. Table 1 summarises some key challenges.

Table 2. Barriers faced by industry in combating CSEA

Barrier type	Examples
Resourcing	• Limited human and technical capacity to respond quickly to emerging forms of harm. • Financial investment required to implement trust and safety measures (particularly challenging for small to medium enterprises), with trust and safety often considered a cost centre. • Inadequate resourcing to support collaboration between platforms, service providers, and safety tech companies. • Limited access to tools and expertise, especially for small and medium enterprises. • Resourcing often prioritised toward privacy over safety, rather than addressing both in parallel.
Technical	• Some content moderation tools may be inefficient or unable to meet business needs at scale. • Challenges detecting and disrupting repeat perpetrators who intentionally evade traditional tactics (for example, IP bans, device blocks). • Difficulty detecting and responding to child-on-child CSEA compared to adult-on-child cases. • Challenges moderating ephemeral content, such as livestreams. • Difficulty applying content moderation and CSEA detection in end-to-end encrypted (E2EE) environments. • A limited number of specialist technical tools that can operate at the required scale • Where tools exist – for example, text and video classifiers – they often rely on limited training data available and still need human oversight. • Antiquated hardware and systems used by law enforcement and other agencies in the reporting and response ecosystem.[25] • Detection and disruption tools can negatively affect user experience – for example, causing delays in loading – which can impact user satisfaction and business performance. • Automated moderation solutions can show inadvertent bias – for example, more likely to detect CSAM of light skin children than dark skin children – if not properly evaluated and corrected. • Difficulties detecting and disrupting cross-platform abuse.
Legislative and regulatory	• Concern companies may face legal liability for engaging in good-faith red-teaming if tools or models surface CSEA content. • Lack of harmonisation across legislation – both within regions and globally. • Fragmented reporting obligations for CSEA and other unlawful content across jurisdictions. • Persistence of a false dichotomy between privacy and safety, despite the fact both can often be supported simultaneously. This includes balancing privacy and security requirements with safety expectations.
	• Image classifiers remain 'data poor' because training them to detect CSEA requires access to illegal content in most countries. • Incentives that render regulatory compliance a checkbox exercise, rather than a meaningful commitment to protecting children.
Research and knowledge	• Lack of an agreed standardised taxonomy of harms within and across industry sectors. • Inconsistent naming and definitions of mitigation measures across research and industry. • Shortage of trust and safety professionals, particularly those with expertise in CSEA. • Lack of community education, understanding, and buy-in. • Trust and safety – and Safety by Design – are largely absent skillsets within engineering education pathways, unlike privacy and security, which are introduced early in the curriculum. • Gendered perceptions of the field – child safety is often seen as the responsibility of women. • Limited philanthropic interest and investment to help drive and support action on CSEA, which weakens the incentives for organisations to prioritise and take meaningful action on this issue. • The need to design and implement measures that account for the diverse needs of children globally. Children living in different cultural contexts – or experiencing conflict, poverty and displacement – may require tailored approaches to fully realise their rights and safety online.
Intra-industry	• Limited availability of established and mature cross-platform solutions for addressing CSEA. • Concerns that child safety initiatives may be used to justify censorship or government overreach. • Perception that acknowledging the issue of CSEA could expose companies to liability, creating reluctance to collaborate. • Challenges in reaching and securing buy-in from industry stakeholders who do not prioritise CSEA. • Difficulty creating safe environments – such as regulatory sandboxes – for testing and sharing technical solutions, particularly given legal constraints around handling CSEA-related material.
Cross-platform	• Lack of consistent policies and processes across platforms creates issues such as a 'threshold gap,' where the same conduct or content may be actionable on one platform but not another. In some cases, behaviour may not violate any platform's policies at any stage of the CSEA cycle. • Individual platforms often lack visibility into the full scope of perpetrator activity, making it difficult to detect and disrupt abuse. Most companies are unable to track signals across platforms. • Cross-platform investigations are costly and resource intensive. • Limited availability of centralised databases where information and signals can be securely shared to support coordinated responses. • No standardised formats or systems for sharing hashes across platforms. • The sector struggles collectively to keep pace with evolving tactics used by perpetrators. • Detection capabilities for novel forms of CSAM (for example, AI-generated CSAM) remain limited. • Competitive concerns can make companies hesitant to cooperate across platforms. • Industry-wide responses are undermined by disengaged companies. Perpetrators often exploit platforms with fewer CSEA countermeasures – for example, by distributing CSEA links across services to see where they gain traction.

[25] Stanford Internet Observatory, 2024, 'How to Fix the Online Child Exploitation Reporting System', accessed 10 January 2025, URL: https://cyber.fsi.stanford.edu/publication/how-fix-online-child-exploitation-reporting-system

Driving systemic change in online safety demands a coordinated response to a wide array of interrelated challenges. This list of barriers makes clear that there is no single solution – industry must be equipped with a diverse and adaptable toolkit. Tool C contributes to this effort by offering a targeted set of tactics to combat CSEA.

There are recurring themes across the barriers and initiatives are underway to not only address immediate challenges but also build long-term resilience and alignment across sectors. These are some examples:

Tackling cross-platform abuse

One of the most challenging barriers to tackling CSEA is cross-platform exploitation and abuse. Perpetrators often initiate contact with children on services such as social media and online gaming platforms, then shift to less detectable environments such as end-to-end encrypted messaging services.[26]

Example 1: A perpetrator contacts a child through the chat function of an online game. After a period of grooming, the perpetrator shifts the conversation to an encrypted messaging service where private messages are not scanned for CSEA-related conduct or content.

Example 2: A perpetrator repeatedly attempts to generate CSAM using Service A's text-to-image generator. Service A detects the activity and bans the user account. The perpetrator then registers with Service B using the same basic subscriber information. As Service B is a smaller company with less robust safeguards, the perpetrator succeeds in generating CSAM.

This tactic, often referred to as 'platform hopping', creates several challenges:

- Different services hold separate pieces of the puzzle, and no single entity has full visibility into the abuse lifecycle.
- Variation in terms of service, community guidelines, and enforcement approaches allow perpetrators to 'window shop' across platforms to exploit loopholes.
- Data sharing efforts are often siloed within specific sectors, excluding key players like payment providers – frequently due to regulatory barriers such as privacy obligations.
- There is little consistency in how child safety issues are defined, addressed, or reported across platforms.
- The volume of 'noise' in signals collected by services makes it difficult to extract and share actionable insights.
- Signal sharing remains largely manual, resulting in inefficiency and inconsistency due to a lack of automation.

[26] van der Spuy A, Livingstone S, Byrne J, et al., 2024, Guiding principles for addressing technology facilitated child sexual abuse and exploitation (CSEA), LSE Research Online, p. 34, accessed 10 January 2025, URL: https://researchonline.lse.ac.uk/id/eprint/126219/

Industry is aware of the challenges posed by cross-platform CSEA and many companies are working together to detect and disrupt it.

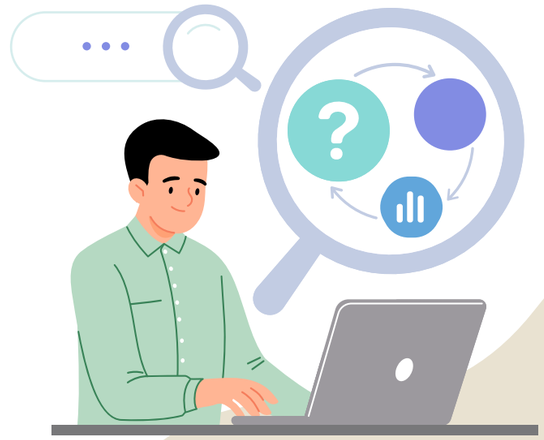
Several initiatives led by child safety organisations and industry associations aim to improve prevention, detection, and disruption of cross-platform CSEA. These include, but are not limited to:

- NCMEC's Take It Down program[27]
- NCMEC's NGO hash lists
- text and content classifier products adopted by major companies
- open-source tools and resources, such as ROOST,[28] which are particularly valuable for companies seeking scalable, modular solutions to meet basic child safety needs.

Signal sharing

A key measure in tackling cross-platform exploitation and abuse is signal sharing – the practice of collecting and analysing signals (such as communication, subscriber details, metadata) to identify and understand threats. In the context of CSEA, signal sharing helps detect, analyse and share relevant indicators of harm to enhance industry-wide awareness of perpetrator behavior, and improve response capabilities. When signal intelligence is shared effectively between platforms, it can prevent perpetrators evading detection by platform hopping.

For example, if a unique account handle is flagged for CSEA-related activity on one platform, that signal can be shared with collaborating companies as a potential risk. This enables platforms to respond more proactively and make better-informed decisions about content and conduct on their services.



[27] National Center for Missing & Exploited Children, 2024, Take It Down, accessed 1 August 2025, URL: https://takeitdown.ncmec.org/
[28] Roost, 2024, Roost Tools, accessed 30 January 2026, URL: https://roost.tools/

Case study

Lantern – Advancing child safety through signal-sharing

The Tech Coalition developed Lantern to address a persistent challenge in the fight against CSEA; the fragmented and siloed nature of detection efforts across the tech industry.

While individual companies often respond effectively to abuse within their platforms, CSEA can be perpetrated across multiple platforms. Previously, any one company could only see a fragment of the harm facing a victim, allowing perpetrators to avoid detection. As the first industry cross-platform signal-sharing program, Lantern addresses these gaps to enhance industry collaboration against CSEA.

Lantern in action

In 2024, a company newly onboarded to Lantern sought to proactively identify previously undetected CSEA threats. By leveraging cross-platform signals – such as indicators of online enticement or solicitation of child sexual abuse material (CSAM) sourced from other companies – the platform was able to uncover multiple instances of abuse that had previously gone undetected.

These discoveries included:

- Users appearing to discuss the sexual abuse and production of CSAM involving their own children (including instances of original content),
- Adults uploading nude images of themselves in exploitative contexts,
- An adult user attempting a contact offence with a self-identified child.

Investigations can fall short without a more complete, multi-platform view of abusive behaviour, Lantern aims to fill these critical gaps by enabling secure signal-sharing, allowing companies to strengthen ongoing investigations, prioritise user reports, and escalate potential harms appropriately.

The impact of cross-platform collaboration

In 2024 alone, as a result of signals shared in Lantern, participating companies:

- actioned over 100,000 user accounts for CSEA-related violations
- removed more than 7,000 pieces of CSAM content
- blocked or removed over 135,000 UR Ls linked to abusive behavior.

Each metric is in addition to actions taken by the original signal uploader, and represent new outcomes that would not have been possible without cross-industry collaboration through Lantern.

Balancing privacy and safety in encrypted environments

End-to-end encryption (E2EE) presents significant challenges for content moderation and enforcement. While it's often framed as a trade-off between privacy and safety, the two are not mutually exclusive. With thoughtful, intentional design, platforms can uphold both values.

Industry should consider leveraging approaches that support both goals, including these:

- **Client-side scanning:** Detect CSAM before it is encrypted or transmitted. On-device scanning enables proactive detection while preserving user privacy.
- **Homomorphic encryption:** Moderate encrypted content without decrypting it. This approach keeps data secure while enabling safety checks.[29]
- **Risk-based E2EE deployment:** Not all platforms are equal. Tailor the use of E2EE based on platform risk profiles, user behaviour and specific use cases.
- **Conditional encryption policies:** Set clear, transparent rules for limiting E2EE when there is a high level of confidence that unlawful activity is occurring. Consider which elements of a platform may not need encryption under any circumstances – for example, profile pictures.
- **Interventions:** Use pop-ups, nudges or warnings in encrypted environments to disrupt harmful behaviour and encourage users to pause or report.
- **Signal-based detection:** Go beyond message content. Analyse metadata, user activity and behavioural patterns to detect potential abuse. For example, signals such as usernames, profile pictures, group memberships and interactions with children can indicate risk.
- **Limit viral spread:** Restrict the ability to mass-forward or rapidly share content to prevent rapid dissemination of harmful material.

[29] Anri van der Spuy, Sabine Witting, Patrick Burton, Emma Day, Sonia Livingstone and Kim R. Sylwander, 'Guiding principles for addressing technology facilitated child sexual abuse and exploitation (CSEA)', Security and Safety, vol. 4, accessed 30 January 2026, URL: https://researchonline.lse.ac.uk/id/eprint/126219/3/van_der_Spuy_et_al_2024_Guiding_principles_for_addressing_technology-facilitated_CSEA.pdf

Helping law enforcement agencies with better information-sharing

Law enforcement agencies are often overwhelmed by the sheer volume of CSEA reports and often lack the critical detail needed to identify perpetrators or protect children at risk.[30] This leads to gaps in information-sharing, missed opportunities to identify victims, and delays in stopping perpetrators.

The tech industry is uniquely positioned to lead a coordinated response to this issue. By embracing collaborative strategies, companies can help close critical gaps in the current prevention ecosystem. Here's how the industry can step up and drive meaningful change:

- **Break down silos:** Tech companies, hotlines and law enforcement agencies should work together – not in isolation. Shared protocols and feedback loops can dramatically improve the quality and usefulness of reports.
- **Standardise what matters:** Align on what information is most useful to investigators. Help internal teams understand what meets the threshold for reporting and what doesn't.
- **Streamline reporting workflows:** Build smarter, more efficient systems that reduce noise, improve triage and prioritise high-risk cases. Better data leads to faster responses.
- **Educate your users:** Raise global awareness of what CSEA is, why it's illegal and how to report it. Empower users to play an active role in prevention.
- **Design for impact:** Make sure every report is actionable. Capture the right metadata, context and signals – without overloading the system.

The result? Greater consistency, faster response times and better protection for children.

Safety by Design in practice

The Safety by Design Principles (Tool A) highlight key actions to improve reporting systems. Consistently applying principles 1.5, 1.6, 3.5.c – on reporting pathways, triage processes, and collaboration – can help reduce bottlenecks and make reporting more effective.

[30] Stanford Internet Observatory, 2024, 'How to Fix the Online Child Exploitation Reporting System', accessed 10 January 2025, URL: https://cyber.fsi.stanford.edu/publication/how-fix-online-child-exploitation-reporting-system

Case study

Snap Inc. and the National Center for Missing and Exploited Children (NCMEC) recalibration

A global rise in sextortion in 2023 and 2024 was one of a confluence of factors that prompted Snap Inc. to approach NCMEC in early 2024 to recalibrate its CyberTip reporting efforts. Additional factors included yet another jump in annual reports to NCMEC from electronic service providers in 2023, feedback about Snap's own CyberTips, and a closer examination of the U.S. Federal reporting statute. In short, many technology companies, including Snap, have come to report **everything** related to child sexual exploitation to NCMEC for fear of missing **something**. However, this has resulted in mounting CyberTips, increased frustration among global law enforcement, and a distorted portrayal of the overall landscape of online risks – not to mention the unintended impacts on child victims.

The goal of Snap's recalibration was to increase the actionability and value of Snap's CyberTips to law enforcement by refining the company's platform policies and reporting protocols. In the year subsequent to that recalibration Snap saw a 49% reduction in NCMEC escalations; more accurate report-labelling; and stricter, more prescriptive internal actions for Snap, particularly regarding imagery of young people whose estimated age cannot be readily determined. (Snap's internal policies, which were amended in conjunction with the recalibration, were fully implemented in September 2024.)

Over time, Snap expects NCMEC and law enforcement will see reduced volumes of reports from Snap, but the ones they do receive will be of higher quality with clear actionability. At the same time, Snap will continue to address violating imagery on its platform at the content, account, and device level – some of this material, however, may simply not warrant reporting to NCMEC.

Opportunities to prevent, detect and disrupt CSEA

Emerging technologies are reshaping the digital landscape, bringing both risk and opportunity. Rapid innovation without robust safeguards has opened new pathways for CSEA. However, these same technologies, when used responsibly, can be powerful tools to detect, prevent and disrupt harm at scale.

As perpetrators adapt their tactics, industry responses must evolve too. Through continuous innovation, adaptability and collaboration, technology companies and regulators can stay ahead of emerging threats and help keep digital spaces safer for children online.

AI is a game-changer

Artificial intelligence is already transforming how CSAM is detected and reported. Companies can create, share and refine models to enable the automatic detection, review and reporting of CSAM at scale.[31] This has the potential to streamline operations and reduce both caseloads and psychological risks for content moderators.

AI-driven nudges can interrupt harmful behaviour before it escalates[32] and some companies are now deploying pre-upload scanning models to detect CSAM before it ever reaches the platform.

AI in content moderation

Generative AI systems, including large language models (LLMs), are creating new possibilities for smarter, more efficient content moderation. These technologies offer innovative capabilities for industry and law enforcement responding to CSEA.[33]

LLMs can assist in CSAM detection by scanning and processing images reported by users and generating text descriptions that capture the content and its severity. This helps moderators identify and prioritise content that may need human review or referral to hotlines or law enforcement.

AI tools – especially machine learning classifiers – also help law enforcement manage growing caseloads and reduce exposure-related harms. They can automate the categorisation of CSAM, prioritise cases based on urgency and risk, and identify patterns in grooming tactics or perpetrator behaviour.[34]

[31] Wolbers H, Cubitt T & Cahill M, 2025, Artificial intelligence and child sexual abuse: A rapid evidence assessment, accessed 22 January 2025, URL: https://www.aic.gov.au/publications/tandi/tandi711
[32] eSafety Commissioner, 2025, 'Generative AI and child safety: A convergence of innovation and exploitation', accessed 11 June 2025, URL: https://www.esafety.gov.au/newsroom/blogs/generative-ai-and-child-safety-a-convergence-of-innovation-and-exploitation
[33] Diaz Garcia J A & Carvalho J P, 2025, 'A survey of textual cyber abuse detection using cutting edge language models and large language models', accessed 9 January 2025, URL: https://arxiv.org/abs/2501.05443
[34] International Centre for Missing & Exploited Children (ICMEC) Australia, 2024, SaferAI for Children Coalition: Discussion Paper on AI and Child Protection in Australia, accessed 26 November 2025, URL: https://icmec.org.au/saferai-for-children-coalition-discussion-paper/

Case study

Safer Predict Classifier by Thorn

Thorn is a technology nonprofit that transforms the way children are protected from sexual abuse and exploitation in the digital age. Its technology solutions are designed to help equip digital platforms with purpose-built solutions to detect CSAM and text-based interactions that could lead to child sexual exploitation (CSE).

Thorn launched 'Safer', a CSAM detection solution, in 2019. Initially, the product relied solely on widely adopted hashing techniques to detect previously reported and verified CSAM. In 2020, Thorn additionally released its CSAM image classifier, which gave platforms the ability to detect previously unknown CSAM. Since then, in response to emerging threats such as sexual extortion and a rise in self-generated CSAM, Thorn developed 'Safer Predict' to detect these harms. Safer Predict leverages machine learning classification models (including text, image and video classifiers) to detect new CSAM and text-based CSE.

The Safer Predict CSAM classifier uses deep learning (convolutional neural networks) to analyse thousands of attributes within images and videos to predict the presence of CSAM. The model classifies images and video into three categories: CSAM, adult pornography and other. Potential CSAM is queued for human review.

The Safer Predict CSE text classifier works by identifying exploitation at the line-by-line and conversation level. The classifier identifies text-based harms, including discussions of sexual extortion, self-generated CSAM, and potential offline exploitation. Trust and safety teams can 'stack' multiple labels to prioritise and escalate high-risk interactions.

By combining cutting-edge machine learning with a mission-driven approach, Thorn's Safer Predict empowers platforms to proactively detect and disrupt child sexual exploitation and abuse.

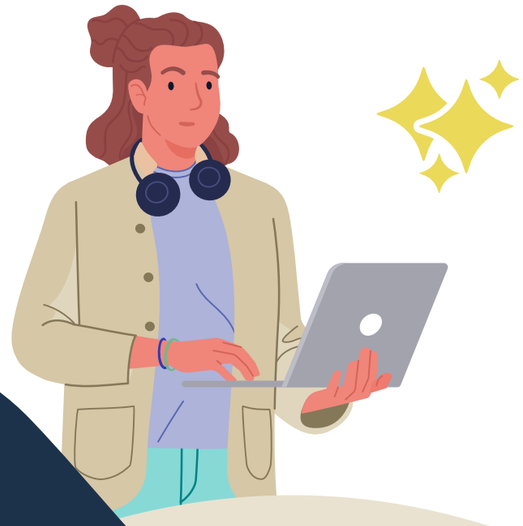
AI challenges

While there is great potential, challenges remain. Legal constraints around processing and training on CSAM limit how AI systems – especially LLMs – can be safely and lawfully developed and deployed.

It's also important to note that innovation without accountability can be dangerous. For example, AI deepfake and nudify apps have made it easy to create CSEA material, by altering images and videos of real children, or by generating images or videos of children who look real but are not (the latter is commonly referred to as 'synthetic' CSEA material). The high volume and rapid spread of AI-generated CSEA material poses a serious threat. Apart from potentially normalising CSEA material, it challenges the capacity of law enforcement agencies to determine when a real child is shown, so they can attempt to identify them and stop the abuse.

There are also emerging risks. Perpetrators may attempt to weaponise AI or exploit automation in moderation systems to mislead law enforcement agencies and evade detection. For example, while AI-generated watermarking can help distinguish synthetic from real CSAM, perpetrators might place an AI watermark on non-synthetic CSAM in the hope that authorities dismiss the material as synthetic and fail to identify a child in need of rescue.[35]

[35] Stanford Cyber Policy Center, Freeman Spogli Institute, 2025, AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims, accessed 29 November 2025, URL: https://cyber.fsi.stanford.edu/news/ai-csam-report



Maximising benefits

These are some actions we need to take now to maximise the benefits of evolving technologies:

- **Closing the oversight gap:** Technology is moving faster than regulation. Companies must lead by example, setting internal safety standards that go beyond compliance.
- **Embedding Safety by Design in AI:** Safety must be integrated at every stage of the AI lifecycle – from training data to deployment.[36] Developers must audit datasets to make sure they are free of CSEA material and mitigate the risk of generating synthetic CSEA material.[37]
- **Tackling immersive platform risks:** Avatars and virtual spaces can be used for exploitation and abuse.[38] Safety must be built into the core architecture of immersive environments.
- **Anticipating adversarial misuse:** Emerging technologies evolve quickly – and perpetrators are agile. Build systems that can adapt and respond to new threats before harm occurs, not just after.
- **Supporting start-ups to build for safety:** New entrants often lack the resources or specialist teams to prioritise safety. Industry leaders must help them integrate safety early through mentorship, shared tools and funding incentives.
- **Boosting digital literacy:** Children, parents and educators need the skills to navigate emerging technologies safely. AI and immersive environments require new capabilities – and cross-sector collaboration is needed to deliver them.

[36] SeSafety Commissioner, 2024, 'Generative AI position paper', accessed 29 November 2025, URL: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai
[37] eSafety Commissioner, 2025, 'Generative AI and child safety: A convergence of innovation and exploitation', accessed 11 June 2025, URL: https://www.esafety.gov.au/newsroom/blogs/generative-ai-and-child-safety-a-convergence-of-innovation-and-exploitation
[38] NSPCC Learning, 2023, Child Safeguarding and Immersive Technologies, p. 22, accessed 11 June 2025, URL: https://learning.nspcc.org.uk/research-resources/2023/child-safeguarding-immersive-technologies

Case study

Safety by Design for generative AI by Thorn and All Tech is Human

In 2022, it became very apparent that generative AI models were actively being used to harm children. Thorn identified this growing harm landscape as a place for proactive intervention, specifically ensuring safeguards are in place to prevent future misuse and harm, alongside reactive measures to mitigate harms that we are already seeing. Thorn also recognized that there were tangible, practical, technical, and policy solutions that could be put in place to prevent the misuse of generative AI in furthering child sexual exploitation and abuse.

In 2023, Thorn and All Tech is Human (ATIH) started the Safety by Design for Generative AI initiative. They brought together industry stakeholders in the generative AI technology space – across open and closed source, AI developers and providers, as well as other key players in technology distribution (such as search engines and social platforms). The goal was to define, align on, and commit to a set of generative AI safety by design principles and mitigations to prevent its misuse in furthering child sexual exploitation and abuse. A working group met three times a week from July to November 2023. The group was able to establish these fundamental principles and mitigations building off of existing research, standards, and guidelines.

If comprehensively followed, the guidelines will:

- reduce the ability of generative models to produce AIG-CSAM and other sexually exploitative content of children
- ensure that any CSEA material that is produced can be more reliably detected
- limit the distribution of models, services, and apps that are used to produce CSEA material.

These principles and mitigations are documented in a paper published by Thorn and ATIH. Since the public launch in April of 2024, the following companies have committed to acting on these principles and to transparently sharing progress: Amazon, Anthropic, Civitai, Google, Invoke, Meta, Metaphysic, Microsoft, Mistral AI, OpenAI and Stability AI.

Also since that launch, Thorn has pursued adoption of these principles and mitigations through multiple avenues. This includes collaborating with standard-setting institutions, informing policymakers on what is technically feasible and impactful in this space, and working directly with builders of technology to implement Safety by Design interventions.

A cultural shift toward safety

CSEA is one of the most urgent and complex challenges facing the digital ecosystem today. While the tech industry has made progress in addressing CSEA, progress alone is not enough. We need a cultural reset where safety is a core value, not a compliance exercise. The approach should be holistic.

- **Building safety into product DNA:** Safety failures are product failures. Treat them like bugs. Child safety must be non-negotiable from day one – baked into product design, development and deployment.
- **Aligning and leading:** Collaborate across industry to define, adopt and enforce best practices. Set the standard and lead by example.
- **Backing safety-first start-ups:** Innovation and safety must go hand in hand. Investors and tech leaders should champion start-ups that embed safety from the ground up – by sharing tools, mentoring founders and helping them scale responsibly.
- **Training for Safety by Design:** Equip engineers, designers and developers with the skills to build safe systems. Make Safety by Design part of core tech education and onboarding.
- **Open sourcing safety:** Share frameworks, tools and lessons learned. Make it easier for everyone to build safer products. Collaboration beats competition when it comes to protecting children.
- **Leveraging brand power:** Advertisers should reward platforms that prioritise safety. Brand investment can help drive safer digital ecosystems.
- **Disincentivising harmful models:** Platforms built on unregulated, uncensored content should face real consequences. Safety must be a competitive advantage, not an afterthought.
- **Designing for age-appropriate experiences:** Standardise protections based on age assurance measures to ensure children are empowered to have rewarding and safe online experiences.

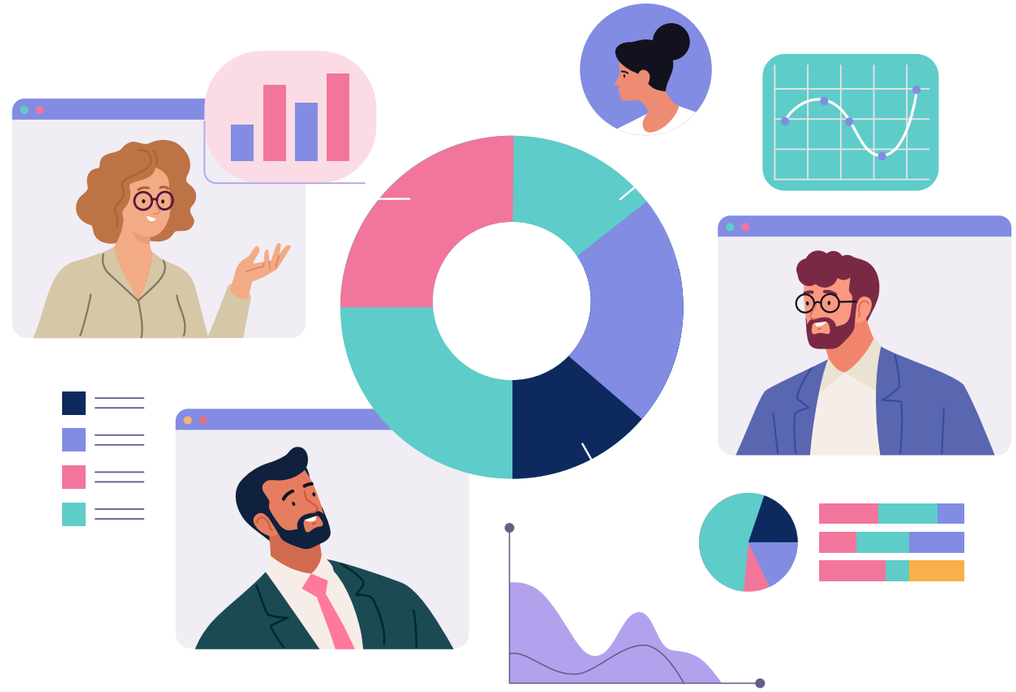
The following set of tools introduces a Safety by Design approach to help the technology industry make this cultural shift:

- **Tool A:** Applying Safety by Design principles in the prevention, disruption and detection of CSEA
- **Tool B:** A checklist to help identify CSEA risk factors across your product, service or features
- **Tool C:** Common interventions to prevent, detect and disrupt CSEA across the tech industry
- **Tool D:** Guidance for start-ups: Identifying and combatting CSEA
- **Tool E:** User reporting: good practice guidelines

There is also a list of referenced and additional resources.

These tools are just the beginning. Safety by Design is not a one-time fix. It is a continuous, evolving commitment to embedding safety into every layer of product development, platform governance and user experience. It demands cross-sector collaboration, operational transparency and a cultural shift that puts child safety ahead of competition and convenience.

eSafety remains committed to listening, learning and evolving. That includes engaging a broader range of stakeholders – particularly smaller and emerging companies – whose voices and innovations are vital to progress.



Appendix 1: Tools

Tool A: Applying Safety by Design principles in the prevention, disruption and detection of CSEA

The Safety by Design principles – developed through extensive research and consultation – are grounded in three core commitments:

1. Service provider responsibility

2. User empowerment and autonomy

3. Transparency and accountability

These principles offer practical, actionable guidance to help platforms and services embed safety into their design, assess their current practices, and continuously improve. They are designed to be realistic and achievable, regardless of a company's size or stage of development.

At their heart, these principles take a human-centric approach. They prioritise the safety, rights and expectations of users, positioning user safety as the third foundational pillar of digital technology development – alongside privacy and security.

Workshop participants collaborated to develop an updated set of Safety by Design sub-principles specifically tailored to addressing CSEA. These reimagined sub-principles provide clear, implementable measures that technology providers of all sizes can adopt to strengthen their response to CSEA.

The following pages outline practical steps for applying each principle to the detection, prevention and disruption of CSEA.

Principle 1. Service provider responsibility

The burden of safety should never fall solely on the user, especially children. Technology companies must take proactive steps to understand, assess and address online harms in the design and delivery of their platforms and services. This means anticipating potential risks during online interactions and designing features to prevent misuse and reduce people's exposure to harm.

1. Nominate individuals or teams and make them accountable for creating, evaluating, implementing and operating user safety policies.

- **a.** Invest in child safety expertise, proportionate to the size of the company and the nature of the platform.
  - **i.** For smaller companies, this may involve regular touchpoints with external experts.
  - **ii.** For larger companies, this may involve embedding child safety experts within the organisation.

2. Develop community guidelines, terms of service and moderation procedures that are fairly and consistently implemented.

- **a.** Write community guidelines, codes of conduct, and user policies in a child-friendly manner – particularly where the platform is likely to be used by, or aimed at, children and young people. Consider creative formats to support engagement and understanding.

Case study

DoorDash's 2025 community guidelines – a plain language approach to platform standards

In 2025, DoorDash took a significant step toward improving user experience and trust by launching a unified set of community guidelines. This move was in direct response to user feedback requesting clearer, more accessible standards for behaviour on the platform. Previously, expectations for dashers, merchants and customers were scattered across various legal documents and webpages, making it difficult for users to understand what was expected of them.

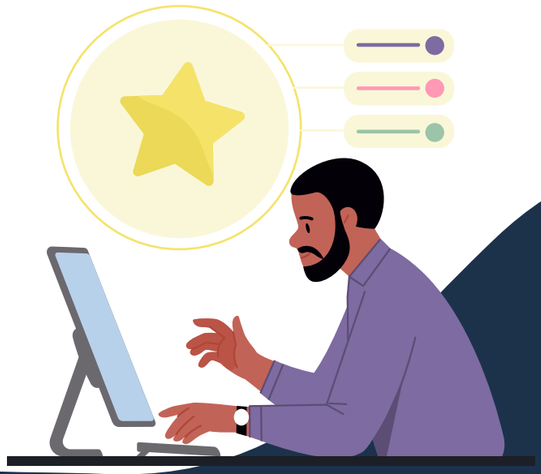
DoorDash addressed this by consolidating its policies into a single, easy-to-navigate framework titled "**Introducing Our Community Guidelines: One Platform, One Set of Standards for Everyone.**" The guidelines are grounded in four high-level principles – **Respect, Safety, Authenticity and Integrity** – each designed to be understandable across a wide range of ages and capacities.

- **Respect** emphasises kindness and prohibits harassment, discrimination and misuse of personal information.
- **Safety** prioritises physical and emotional well-being, banning actions that could cause harm or compromise food safety.
- **Authenticity** requires users to accurately represent themselves and use their own accounts, with tailored verification processes depending on their role.
- **Integrity** calls for honest behaviour and discourages manipulation of the platform for personal gain.

These principles serve as a foundation for more detailed guidelines, which include practical examples and links to related policies. By using plain language and a values-based structure, DoorDash ensures that its expectations are not only transparent but also relatable and actionable.

This approach reflects a broader trend in digital platform governance: simplifying complex legal and behavioural standards into accessible formats that foster inclusivity and accountability. DoorDash's 2025 guidelines demonstrate how clear communication and thoughtful design can enhance user trust and safety while maintaining operational integrity.

3. **Put processes in place to detect, surface, flag and remove illegal and harmful conduct, contact and content** with the aim of preventing harms before they occur. These processes should be regularly reviewed and updated to ensure they remain fit for purpose.
   - **a.** Regularly update senior leadership on process changes and their impact.
   - **b.** Introduce and embed technical interventions where appropriate (such as in-house or external content moderation tools equipped to detect and prevent CSEA).
   - **c.** Implement other effective prevention layers that make CSEA more difficult for perpetrators



Case study

ROOST delivers open-source trust and safety infrastructure for the AI era

Small-scale and start-up platforms can lack access to essential safety tools, leaving users exposed to harm. To close this gap, ROOST launched two open-source tools in 2025: Coop and Osprey. These foundational technologies enable trust and safety teams to detect, triage and respond to harmful content – without the cost or complexity of building enterprise systems in-house.

Coop equips platforms with content review capabilities, including routing flagged material for expert human review with context to help them take action. It integrates directly with the National Center for Missing and Exploited Children's (NCMEC) API, ensuring mandatory reporting of child sexual abuse material and compliance with regulations.

Osprey is a powerful investigation and incident response tool that gives safety teams visibility into platform activity and the ability to act at scale. Designed for platforms of all sizes, Osprey delivers enterprise-grade functionality without enterprise-level infrastructure.

These tools were made possible through strategic partnerships. Coop builds on technology originally developed by Cove, whose intellectual property ROOST acquired to open-source as a public good. Osprey was created by Discord and donated to ROOST, underscoring a shared commitment to safer online spaces.

Zooming out, Coop and Osprey are more than tools, they are strategic interventions closing critical safety gaps where resource constraints leave smaller and emerging platforms exposed. By making these solutions open-source, ROOST is levelling the playing field and driving an industry-wide shift toward higher safety standards.

4. **Establish infrastructure to support internal and external triaging, clear escalation pathways, and readily accessible user reporting mechanisms**, ensuring users can flag concerns or violations at the point they occur.
   - **a.** Establish minimum service level agreements and define success metrics for reviewing and actioning user reports and detected violations.
   - **b.** Provide clear, in-product reporting mechanisms that are easy to access and navigate, including for children and young people.
   - **c.** Provide portals for non-users to report violative content without putting undue burden on the reporter (for example, where a victim, survivor or parent may not be a registered user).
   - **d.** Ensure reports are actioned and responded to swiftly. Companies should strive to continuously improve action and response times, without compromising accuracy.
5. **Establish clear internal protocols for engaging with law enforcement, support services and illegal content hotlines.**
   - **a.** Set up referral protocols that do not overwhelm law enforcement with reports of non-CSEA content. Seek feedback from law enforcement and hotlines to ensure reporting methods are useful and actionable.
6. **Prepare documented risk management and impact assessments** to assess and remediate any online harms that could be enabled or facilitated by the product or service.
   - **a.** Conduct a child safety risk assessment.
   - **b.** Conduct red teaming of AI models.
7. **Implement social contracts at the point of registration.** These outline the duties and responsibilities of the service, the user and any third parties for keeping all users safe.
8. **Balance security by design, privacy by design and user safety considerations** to maintain the confidentiality, integrity and availability of personal data and information.
   - **a.** Ensure child safety is central to decisions that weigh up security, privacy, and safety needs. Smaller companies may draw on existing resources or consult experts when considering how child safety risks may manifest on their platform. Larger companies should ensure their child safety experts are part of the decision-making process.
9. **Embed child safety into the product development lifecycles throughout the company**, not just in trust and safety teams.
   - **a.** Implement specific action items, checks and unit tests for product managers, engineers, quality assurance and other relevant teams. Encourage continuous improvement.
   - **b.** Educate internal teams about child safety.
   - **c.** Prepare taxonomies and materials that are easy to understand across all teams, including engineering and technical staff.

Principle 2. User empowerment and autonomy

The dignity of all users – particularly children and other vulnerable groups – is of central importance. Tech products and services should prioritise user wellbeing and uphold fundamental consumer and human rights. This is critical in the context of CSEA, which causes profound and far-reaching harm to victims, their families and communities.

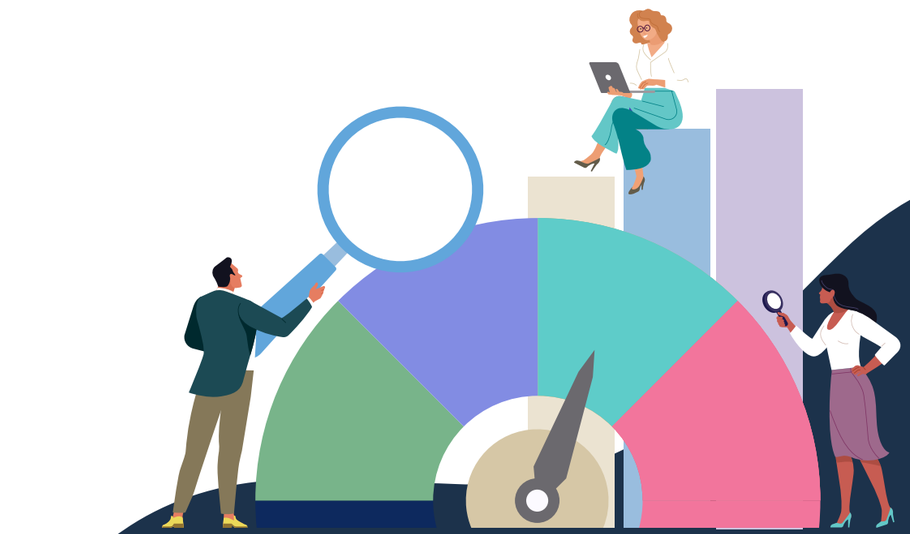
1. **Provide technical measures and tools that enable users to manage their safety, with the highest privacy and safety levels set by default.**
   - **a.** Enable stronger default settings for child accounts, embed age-appropriate experiences, and/or allow parents or guardians to tailor a child's online experience.
   - **b.** Encourage and enable users to be honest about their age by designing and implementing low friction but reliable age assurance measures.
   - **c.** Implement age-appropriate onboarding pathways that align with a user's developmental level.
2. **Establish clear protocols and consequences for service violations** that serve as meaningful deterrents and reflect the values and expectations of users.
3. **Use technical features to mitigate risk and prompt safer interactions** by flagging concern to users at relevant points in the service.
   - **a.** Where appropriate, use technological tools to detect CSAM and CSEA on a platform or service.
4. **Provide built-in support functions and feedback loops** so users know the status and outcome of their reports and have the opportunity to appeal.
   - **a.** Establish minimum expectations for service level agreements and success metrics for response to reports.
   - **b.** Provide trauma-informed responses where users are reporting their own imagery – for by offering additional support resources.
   - **c.** Implement robust appeals processes for both reporters and those who are reported against.
5. Evaluate all product designs and features before release to ensure risks are mitigated, particularly for users with distinct characteristics or capabilities.
   - **a.** Consider edge cases and how the service might be misused, conduct unit testing for child safety.
   - **b.** Conduct red teaming of AI models.
6. Provide trauma-informed resources for anyone affected by CSEA content, including staff reviewing CSAM and users reporting it.
   - **a.** This may include referrals to external support services, such as Kids Helpline (Australia) and Childhelp (U.S.).

Principle 3. Transparency and accountability

Transparency and accountability are hallmarks of a robust approach to safety. They give users confidence that platforms and services are upholding their safety commitments. They also help educate and empower users to address safety concerns. In the context of CSEA, transparency and accountability help victim-survivors know their rights and protections and encourage industry to lift safety standards.

1. **Embed safety into everyday practice:** In addition to dedicate Trust and Safety employee, make user safety part of the responsibilities, training and workflows of everyone who works with, for or on behalf of the product or service.
2. **Ensure safety policies are accessible and clear:** Safety-related policies – such as terms and conditions, community guidelines and user-safety processes – should be easy to find, easy to understand, and updated regularly. Remind users of these policies periodically and inform them about changes through targeted in-service communications.
   - **a.** Clearly prohibit CSEA in terms of service, blogs and other communications. Include strong statements of zero tolerance.
   - **b.** Be transparent about the measures taken to address CSEA, including any deterrent effects. (See 4a).
   - **c.** Apply CSEA-related policies consistently and ensure enforcement actions are proportionate to the harm.

Internally, make it clear which teams and systems are responsible for enforcing safety policies.



Case study

Linktree's zero tolerance policy for CSEA and CSAM

Linktree has publicly committed to a zero-tolerance policy for CSEA and CSAM, reinforcing its role in safeguarding online spaces. This commitment is outlined in its community standards and further supported by its membership in the Tech Coalition.

Linktree's approach includes:

- **Strict content moderation and enforcement:** The platform prohibits any content that depicts, promotes or facilitates the abuse or exploitation of children. This includes digitally altered or animated media that may be harmful.
- **Immediate removal and reporting**: Any suspected CSAM or CSEA content is swiftly removed and reported to the appropriate authorities, including the Australian Centre to Counter Child Exploitation (ACCCE) and international partners.
- **Permanent bans:** Accounts found to be involved in uploading or sharing CSAM or CSEA-related content are permanently banned to prevent recurrence.
- **Proactive collaboration:** Linktree works with industry peers through the Tech Coalition to share best practices, improve detection technologies, and strengthen global child safety standards.
- **User education and support:** Linktree provides resources and guidance to help users identify and report harmful content, aligning with broader efforts such as those led by the eSafety Commissioner.

Linktree's public stance and active participation in cross-sector initiatives demonstrate its accountability and commitment to creating a safer digital environment for children. By embedding child safety into its platform governance, Linktree sets a strong example for responsible tech leadership.

3. **Engage openly with users and experts:** Consult a broad range of users and stakeholders in the development, interpretation and application of safety standards – and in assessing their effectiveness.

4. **Publish meaningful safety data and analysis** Each year, release a summary of reported abuses on the service, along with a meaningful analysis of the underlying data, moderation efforts, and the effectiveness of enforcement.

   a. Regularly publish transparency reports that show how the platform or service is enforcing its CSEA policies. Where different parts of the service carry different risks (for example, newsfeeds compared to direct messaging), break down the information to show how the service responds to different risks. Include:

   i. Qualitative data – descriptions of the technologies, systems and processes used to prevent, detect and remove CSEA.

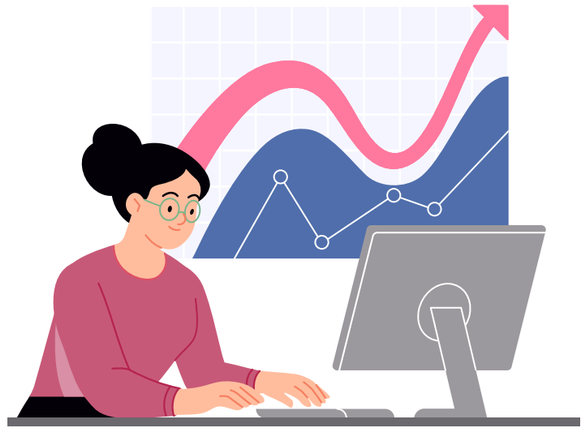
   ii. Quantitative data – number of complaints received, response times, reports to law enforcement and approved hotlines (such as NCMEC), account bans for CSEA violations, and performance metrics on detection and reporting processes.

5. **Commit to ongoing innovation and collaboration:** Continue to invest in safety-enhancing technologies, and collaborate with others by sharing with tools, best practice.

   a. Stay up to date about emerging risks and ensure safety interventions reflect new and evolving threats.

   b. Regularly review and evaluate child safety measures to ensure they remain effective as technologies and platforms evolve.

   c. Take part in industry-wide and cross-sector collaborative initiatives.



Tool B: A checklist to help identify CSEA risk factors across your product, service or features

Who is this checklist for?

This checklist is for professionals across the tech sector – including product managers, engineers, designers, data scientists and trust and safety teams – who build, deploy or maintain digital products and services.

Whether you're launching a new feature, scaling a platform, or refining user engagement strategies, this checklist can help you proactively assess potential vulnerabilities.

It is as a practical starting point to help identify and mitigate risks related to CSEA, even in contexts where such risks may not be immediately apparent.

Features and products involved in the CSEA cycle may not be immediately obvious. These include:

- sharing URLs that link to CSAM
- shortening URLs to evade detection by URL-blocking tools
- hosting websites that share CSAM, grooming guides or other perpetrator tactics
- facilitating payment for the exchange of CSAM or other CSEA-related content.

The checklist

Activity-based risks

Does the platform, product or feature enable any of the following activities:

- ☐ Hosting, storing, or sharing files (such as documents, images, videos)?
- ☐ Messaging between users via direct messages, chat rooms or discussion forums?
- ☐ Sharing or shortening of URLs?
- ☐ Sharing of user-generated content (such as text posts, images, videos, links)?
- ☐ Generating text, images and/or videos?
- ☐ Sharing of ephemeral content (such as livestreams, disappearing messages)?
- ☐ Use of user rewards (such as in-game currencies, digital gifts)?
- ☐ Sharing or generating nude, pornographic or sexually explicit adult content?
- ☐ Storing or training on large volumes of data (such as for AI model development)?

Children's access and interactions

- ☐ Is your platform, product or feature intended for or likely to be used by children?
  - **o** If yes, can you demonstrate that the best interests of the child are a primary consideration in its design and operation?
- ☐ Are age-appropriate experiences underpinned by age assurance systems that enable differentiated access based on developmental needs?
- ☐ Does your platform, product or feature enable social interactions (such as friend requests, follows, direct messages, live chats)?
  - **o** If yes, are mitigations in place to limit interactions between children and unknown adults?
- ☐ Are community guidelines and user policies written in child-friendly language?
- ☐ Can children easily find and use in-app reporting tools?
- ☐ Are users delivered age-appropriate experiences (such as stricter default content settings for younger users)?

Content moderation

- ☐ Have you implemented the best available – or most appropriate – technical tools to detect CSEA on your platform, product or feature?
  Several tools are available for free or at low cost.
  - **o** If yes, are you continuously measuring the impact and effectiveness?
- ☐ Do you have processes and tools to **effectively** moderate content in languages other than English?
- ☐ If you use community moderation, are there rules and safeguards to ensure your staff are notified of any CSEA-related activities discovered or actioned by community moderators?
- ☐ Does your content moderation system include cross-checking with existing databases of known CSAM hashes and URLs?
- ☐ Are there established timelines and processes for reviewing and evaluating your content moderation tools and policies to ensure they remain effective?
- ☐ Do you proactively moderate for both known and unknown CSAM?
- ☐ Have you conducted red-teaming exercises to test for CSEA-related risks and weaknesses? (Ensure all red-teaming activities remain lawful).
- ☐ Is there a risk that your platform, product or feature has been developed using training data that may contain CSAM?
- ☐ Do you have triaging and reporting mechanisms to escalate appropriate content to law enforcement and child safety hotlines?
  - **o** Have you consulted with law enforcement on thresholds for reporting content and what information or formats are most effective?

Organisational practices

- ☐ Are child safety risk assessments conducted before launching – or making changes to – any service, product or feature? Is there mandatory, ongoing training for all staff, including engineers, developers and policymakers, about CSEA and how it manifests?
- ☐ Does your company proactively participate in industry initiatives that focus on addressing CSEA?
- ☐ Does your company engage with reputable child safety organisations (such as through partnerships with organisations such as the Internet Watch Foundation)?
- ☐ Do you publish regular transparency reports that include information about child safety efforts (such as number of reports, violations, banned accounts, and response times)?
- ☐ Have you assessed CSEA risks in your supply chain (such as ensuring the tools you use do not contain or facilitate CSEA)?

Next steps

This checklist may have helped you identify CSEA-related risks that require further assessment and the development of appropriate mitigation measures. You should consider conducting a thorough child safety risk assessment, seeking expert advice, and implementing robust safeguards.

Tool C: Common interventions to prevent, detect and disrupt CSEA across the tech industry

Tool C includes a repository of common and achievable interventions that organisations should consider deploying.

It is not enough to rely on a single intervention. Companies should adopt a multifaceted and ongoing approach to combatting CSEA. Research by the Australian Institute of Criminology suggests that detection tools are most effective when multiple methods are used together, such as combining deep-learning algorithms with multi-modal image or video descriptors.

Measure	Description	Type
Safeguards for staff wellness	Implement and routinely review safeguards to protect staff exposed to CSEA-related content, including text-based materials such as conversations between perpetrators and victims. Supporting staff wellbeing strengthens organisational capacity to identify and respond to online harms. Safeguards may include mental health literacy training, regular check-ins with staff, access to counselling services, and using AI to filter the most harmful material before human review.	Organisational
Trust and safety staff trained in CSEA	Appoint dedicated trust and safety personnel to create, implement, and evaluate child safety policies and operations. These roles are critical to preventing harm before it occurs. Staff should have a clear understanding of CSEA and how it may manifest on the platform or service.	Organisational
Understand legal and regulatory obligations	Proactively understand child safety laws and regulatory requirements in all jurisdictions where the company operates. This includes complying with relevant industry standards and frameworks, and implementing policies and processes that, at a minimum, meet legal obligations.	Organisational
Threat intelligence	Adapt threat intelligence practices – commonly used in cybersecurity and fraud prevention – to the CSEA context. This involves gathering, analysing and applying actionable information about past, current, or emerging threats (such as IP addresses or behavioural signals linked to perpetrators or coordinated attacks). Used effectively, threat intelligence helps organisations understand the CSEA threat landscape and proactively mitigate risks.	Organisational, technical
Terms and conditions and/ or service level agreements addressing CSEA	Establish clear terms of service and user policies that define acceptable conduct and explicitly prohibit CSEA-related activity, such as the creation, possession, or sharing of CSAM. These should be supported by robust enforcement measures. Minimum age requirements must also be clearly stated and consistently applied.	Policy
Community guidelines and/or user policies explicitly addressing CSEA	Publish clear community guidelines and user policies that define acceptable and unacceptable behaviour, with specific reference to conduct involving children. These policies typically provide more detailed guidance than terms and conditions or service level agreements. In many cases, a standalone CSEA policy may be appropriate. Providing a child-friendly version of these guidelines can help children and young people understand what conduct and content is acceptable, and what they can expect from the platform or service.	Policy
Limiting features available to accounts	Restrict certain account features – such as link or image sharing – when there are indicators of violative behaviour related to CSEA or other platform policies, even if the behaviour does not reach the threshold for confirmed CSEA. This can serve as a preventative enforcement measure to reduce risk while further review takes place.	Policy, technical
Limiting higher risk features by age group	Restrict access to features known to have a higher risk of facilitating CSEA, such as livestreaming for users under 18.	Policy, technical
Specific rules for user interaction	Establish clear policies and enforcement procedures governing interactions between adult and child users. These should include stricter safeguards than those applied to child-to-child interactions and must be reinforced by appropriate technical controls to prevent misuse.	Policy, technical
Suspension or closure of account	Suspend or close user accounts – automatically in clear-cut cases or manually in edge cases – when CSEA-related policies or procedures are violated. Penalties and enforcement actions should be proportionate and supported by clear appeal mechanisms to ensure procedural fairness and to reduce the risk of overenforcement.	Policy
Transparency reporting on CSEA	Publish transparency reports that clearly explain the measures taken to proactively prevent, detect, disrupt and respond to CSEA at the platform or service level. Reports should include CSEA-specific data, such as the number of user reports, referrals to law enforcement, and emerging content trends.	Policy, organisational
Child safety risk assessments	Ensure that child-specific risks are identified, assessed, and addressed through appropriate mitigation strategies. This may be integrated into standard risk and safety assessments during the development of new products, platforms, or services, or conducted as a standalone assessment, depending on the context.	Procedural, organisational
Red-teaming for CSEA	Incorporate CSEA-specific scenarios into standard red-teaming processes to identify flaws or weaknesses that could be exploited to facilitate harm. Any vulnerabilities uncovered should be addressed promptly to reduce risk. Ensure testing is lawful and, where appropriate, consult relevant legislation and government agencies before conducting red-teaming activities.	Procedural
User reporting pathways	Ensure users can report CSEA to the platform or service provider through clearly defined reporting categories dedicated to this type of harm. Establish infrastructure to support effective triage, escalation, and user feedback – informing users of the status of their report, any actions taken, and providing opportunity to appeal decisions.	Procedural, technical

Measure	Description	Type
Pathways for reporting CSEA to law enforcement and/or hotlines	Establish clear procedures for reporting CSEA to law enforcement and/or child safety hotlines. These pathways should be developed in line with legal requirements and, where possible, based on feedback from law enforcement or hotline agencies to ensure reports are actionable and effective.	Procedural, organisational
Age assurance	Use age assurance measures to help deliver age-appropriate experiences. This may include one or more approaches, such as biometric age estimation (such as facial analysis), AI tools that assess behavioural patterns, or age verification against external sources like credit card or ID data.	Technical, procedural
AI-driven text descriptions of suspected CSAM	Use tools that detect CSAM by analysing the language used to describe the material (such as written captions, text-based metadata, or language embedded in audio content[39]).	Technical
Authentication or verification requirements and similar methods of friction	Require users to verify or authenticate their identity through methods such as linking a credit card, phone number, or social media account to the platform or service.[40]	Technical, procedural
Classifiers (file name/path)	Use machine learning classifiers to analyse file names and file paths, automatically categorising them into risk-based categories. These tools help assess the likelihood that a file contains CSAM without needing to open it.[41]	Technical
Classifiers (image)	Use machine learning models to analyse pixel values and automatically categorise images based on predetermined labels. This is particularly useful for detecting unknown CSAM. Nudity detection classifiers are a specific type designed to detect nudity in images.	Technical
Classifiers (text)	Use machine learning models to automatically categorise open-ended text – such as direct messages or posts – into predetermined categories (such as 'CSAM' and 'non-CSAM'). This is particularly useful for detecting grooming behaviours and other high-risk interactions.	Technical
Classifiers (video)	Like image classifiers, video classifiers use machine learning to analyse and categorise visual content in video streams. These tools can identify and track specific activities within videos and are particularly useful for detecting previously unknown CSAM.	Technical
Content moderation tools (general)	Software that provides content moderation capabilities for detecting, disrupting and responding to CSEA. These tools typically integrate multiple components – such as classifiers, filters, and reporting workflows – into a single system to support end-to-end moderation.	Technical, procedural
Device-level interventions	Technical measures that operate at the device level to prevent the creation or consumption of CSEA. For example, SafeToNet's HarmBlock can be embedded into a device's operating system to block explicit content, including CSEA, from being displayed on screen or captured by the device's camera. Human Mobile Design (HMD) has partnered with SafeToNet to roll out HarmBlock in its FusionX1 model.[42]	Technical
Friction measures in direct messages	Deploy prompts such as 'Are you sure you want to send this?' when suspected CSEA material is being shared in direct messages. These prompts introduce a moment of friction that can interrupt and prevent harmful behaviour. This is particularly relevant in cases involving child-on-child CSEA.	Technical
Hash-matching	Use tools that generate a hash for an image and compare it against a database of known CSAM, such as the Internet Watch Foundation's (IWF) image hash list, or NCMEC's CSAM and exploitative file hash lists, to identify matches. Different types of hashing serve different purposes. Cryptographic hashing detects exact matches; and perceptual hashing can be identified visually similar content (for example, resized, cropped or rotated images). Hashing techniques can also be applied to other types of multimedia content.[43] Examples of image hashing technologies include, (but are not limited to): PhotoDNA, MD5, Internal Hash, Thorn's Safer Hash, PDQ, SHA1/SHA2, and CLIP.[44]	Technical
Keyword detection	Use tools that scan text against databases of keywords associated with CSEA – such as those from the IWF keyword list – and flag matches for review or action by the platform or service.	Technical
Parental tools	Enable parents and carers to link to their children's accounts and manage privacy and safety settings. Provide an age-appropriate level of visibility into their child's activity (such as offering greater oversight for younger children and more autonomy for older teenagers).	Technical, procedural
Redirection through pop-up messages and chatbots	Deploy tools that detect CSEA-related activity – such as searching for CSAM or attempting to access URLs previously removed for hosting CSAM – and automatically respond with deterrents. These may include pop-up messages or chatbots that direct users to help-seeking resources.[45]	Technical

[39] Wolbers H, Cubitt T & Cahill M, 2025, Artificial intelligence and child sexual abuse: A rapid evidence assessment, accessed 22 January 2025, URL: https://www.aic.gov.au/publications/tandi/tandi711
[40] Gorwa R & Thakur D, 2024, Real Time Threats: Analysis of Trust and Safety Practices for Child Sexual Exploitation and Abuse (CSEA) Prevention on Livestreaming Platforms, Center for Democracy & Technology, p. 19, accessed 22 January 2025, URL: https://cdt.org/insights/real-time-threats-analysis-of-trust-and-safety-practices-for-child-sexual-exploitation-and-abuse-csea-prevention-on-livestreaming-platforms/
[41] Al Nabki M W, Fidalgo E, Alegre E & Alaiz Rodriguez R, 2023, 'Short text classification approach to identify child sexual exploitation material', Scientific Reports, accessed 22 January 2025, URL: https://www.nature.com/articles/s41598-023-42902-8 and Anderson P, Zuo Z, Yang L & Qu Y, 2019, 'An intelligent online grooming detection system using AI technologies', 2019 IEEE International Conference on Fuzzy Systems (FUZZ IEEE), accessed 22 January 2025, URL: https://ieeexplore.ieee.org/document/8858882

[42] HarmBlock,, AI powered protection against sexual imagery, accessed 30 January 2026, URL: https://harmblock.com/
[43] Child Rights International Network (CRIN), 2023, 'Explaining the technology for detecting child sexual abuse online', accessed 30 January 2026, URL: https://home.crin.org/readlistenwatch/stories/explainer-detection-technologies-child-sexual-abuse-online
[44] Tech Coalition, 2024, Tech Coalition 2024 Annual Report, accessed 30 January 2026, URL: https://technologycoalition.org/resources/2024-annual-report/
[45] Price S, McKillop N, Scanlan J, Rayment McHugh S, Christensen L & Prichard J, 2024, 'A Review of Digital Interventions as Secondary Prevention Measures to Combat Online Child Sexual Abuse Perpetration', Journal of Child Sexual Abuse, Taylor & Francis Online, accessed 30 January 2026, URL: https://www.tandfonline.com/doi/full/10.1080/10538712.2024.2415549 Prichard J, Wortley R, Watters P, Spiranovic C & Scanlan J, 2024, 'The effect of therapeutic and deterrent messages on Internet users attempting to access "barely legal" pornography', Child Abuse & Neglect, Version of Record 1 August 2024, accessed 30 January 2026, URL: https://www.sciencedirect.com/science/article/pii/S0145213424003454; Hunn C, Watters P, Prichard J, Wortley R, Scanlan J, Spiranovic C & Krone T, 2023, How to implement online warnings to prevent the use of child sexual abuse material, Trends & Issues in Crime and Criminal Justice, no. 669, Australian Institute of Criminology, Canberra, accessed 30 January 2026, URL: https://www.aic.gov.au/publications/tandi/tandi669

Measure	Description	Type
Safety/privacy by default measures for children	Ensure that child accounts are set to the highest privacy and safety settings by default (such as blocking messages from unknown senders or requiring approval to receive messages from someone who is not a contact. These measures are most effective when combined with robust age assurance.	Technical, policy
URL detection	Use tools that detect and compare URLs against lists of known CSEA-hosting websites (such as the IWF URL list). Matches are flagged for appropriate action by the platform or service.	Technical
Link analysis (financial transactions and networks)	A fraud detection technique used by financial service institutions to analyse payment activity and identify patterns of potentially violative activity, such as fraud, money laundering or sexual exploitation.	Technical, industry
Industry collaboration	Participate in alliances, working groups, projects, and other activities that promote information and skill sharing across industry stakeholders. These efforts support collective improvement of practices to address CSEA.	Industry
Open-source tooling	Provide or use publicly available, ready-to-deploy tools that enable companies of all sizes to implement CSEA countermeasures more easily and cost-effectively (for example, open-source safety tools provided by Robust Online Open Safety Tools 'ROOST').	Industry, technical
Signal sharing	Securely share CSEA-related signals – such as IP addresses, usernames and metadata – with other companies and child safety organisations to help detect and disrupt cross-platform CSEA activity (such as Project Lantern).	Industry, technical
Voluntary principles	Collaborate across industry, child safety organisations, and other stakeholders to establish shared commitments and guiding principles for combatting CSEA. Examples include: • Safety by Design for Generative AI Principles • Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse • Guiding principles for addressing Tech-facilitated CSEA	Industry, organisational

Tool D: Guidance for start-ups: Identifying and combatting CSEA

This guidance is designed for anyone building or leading a digital product or service in a start-up environment.

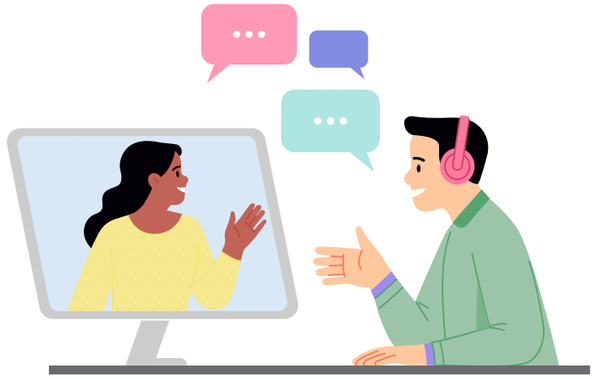
CSEA can occur on almost any digital platform – even those not typically associated with messaging or image sharing. From fintech apps that may unknowingly facilitate illicit transactions, to social platforms that can be exploited for grooming, the risks are real and far-reaching.

Start-ups move fast. In the race to launch – while navigating product-market fit, funding rounds and scaling – child safety can be unintentionally sidelined. That's a risk no company can afford to take.

Embedding safety from day one isn't about bureaucracy – it's about building smart, intentional guardrails that protect your users, your brand and your mission. A safety-forward approach doesn't slow you down, it sets you apart.

Benefits for start-ups

- **Cost efficiency:** Addressing CSEA risks early helps avoid costly fixes, legal exposure and reputational damage.
- **Brand trust:** Users and partners trust platforms that take safety seriously. That trust generates loyalty, growth and long-term success.
- **Regulatory readiness:** Proactive safety measures help you stay ahead of compliance requirements and demonstrate accountability to regulators and investors.
- **Innovation edge:** Safety by Design is not just ethical – it builds better products, creates stronger communities, and drives continuous improvement.



Start-ups have a unique advantage

Unlike legacy platforms, start-ups are not burdened by outdated systems or reactive policies. They have the opportunity to build safety into their DNA from day one, learning from the hard lessons of others and avoiding costly missteps.

The tech industry increasingly recognises that safety is not a competitive advantage – it's a shared responsibility. Larger platforms are sharing insights and tools to help smaller players succeed safely.

Industry-wide initiatives are helping start-ups build capacity in child safety. For example, the Tech Coalition's **Pathways Program** provides resources, mentorship, and access to child safety technology for companies new to trust and safety.

No matter your stage of growth, you can also engage with the **National Center for Missing & Exploited Children (NCMEC)** to access services such as hash-sharing initiatives.

Your next step

Whether you're building a social app, marketplace, or new AI tool, ask yourself: What are we doing to protect children on our platform now? And, how will we continue to protect children as our platform scales?

If the answer isn't clear, now is the time to act. A guided self-assessment is a good place to start and there are many free and paid tools and resources available to support you.

Resource	Application; harm type
eSafety's **Safety by Design self-assessment tools**	Broader; all online harms
Thorn and All Tech Is Human's **Safety by Design for Generative AI: Preventing Child Sexual Abuse**	Tech sector using or providing generative AI; CSEA
Digital Thriving's **Combatting Online Child Sexual Exploitation and Abuse in Gaming** guide	Gaming; CSEA
The Tech Coalition's **Developer Good Practices: Combating Online Child Sexual Exploitation and Abuse**	Developers in the tech sector; CSEA
AUSTRAC's **Combating the Sexual Exploitation of Children for Financial Gain – Financial Crime Guide**	Financial tech; CSEA
Centre for Democracy and Technology's **Real Time Threats** research report	Platforms with livestreaming; CSEA

Avoiding the pitfalls: lessons start-ups can't afford to miss

In the race to build and scale, risks that don't feel immediate are easy to overlook. But when it comes to CSEA, the cost of inaction is too consequential. Here's how to stay ahead:

- **Pitfall 1:** Plan for the future. Ignoring CSEA risks during early development leads to costly retrofits, reputational damage, and legal exposure. Safety is not a 'later' issue – it's a now priority. Bake it into your roadmap from day one and prepare for a future state when your platform scales.
- **Pitfall 2:** Avoid one-and-done safety investments. CSEA threats evolve. Your safety strategy must evolve too. A one-off investment won't cut it. Routine reviews and updates are essential to keep your platform resilient and responsive.
- **Pitfall 3:** Steer clear of late-stage tool adoption. Waiting to implement content moderation and other technical tools can create overwhelming backlogs and missed opportunities to intervene early. If you're not ready to deploy tools yet, plan ahead so you can manage the surge when the time comes.
- **Pitfall 4:** Don't over-report to law enforcement agencies. Flooding law enforcement or hotlines with vague or low-quality reports can dilute the impact of genuine cases. Define clear thresholds and criteria for reporting and focus on providing actionable intelligence.

What next?

Your next move depends on the findings of your self-assessment. These should reflect your operational reality – your product, your users and your specific risk profile. But one principle applies across the board:

- **Safety must be part of your culture**. Embed it into internal processes, empower your team to take ownership, and educate your users to navigate your platform safely.

At the operational level, there are practical, low cost measures any start-up can take. For example:

Action	Guidance
Clear policies and protocols for users that address CSEA	Clearly communicate a zero-tolerance policy toward CSEA, enforcing penalties as required.
User reporting systems	Provide a simple and accessible way for users to report suspected CSEA, and ensure reports are promptly actioned by staff, automated systems, or both.
Reporting flow to law enforcement/hotlines	Establish a clear process for referring or reporting CSEA to law enforcement or CSEA hotlines in your jurisdiction.
Moderation tools	Use technical tools to help prevent CSEA on your platform or service – for example, hash matching tools and age assurance systems such as facial age estimation and ID-verification. While content moderation systems require dedicated resources – particularly at the outset – there are free and low-cost tools available. For example: • Microsoft's PhotoDNA Cloud Service • Google's Child Safety Toolkit, consisting of its Content Safety API and CSAI Match • Thorn's CSAM Keyword list • IWF's Image Intercept • IWF's member resources (membership fees vary based on company size) • Discord's Visual Safety Technology – CLIP.

The IWF's Image Intercept tool is specifically designed to help smaller, eligible companies and start-ups proactively detect and stop known CSAM.

Taking a Safety by Design approach from the outset can prevent major challenges later. When safety is embedded early, it becomes a foundation – not a fix. It should be part of your company's culture: integrated into internal processes, championed by your team, and reflected in how users interact with your platform.

Empower your team to take ownership of safety. Equip your users with the knowledge and tools to navigate your service confidently. When safety is built in, not bolted on, it protects your product, your people and your purpose.

Tool E: User reporting: good practice guidelines

Why it matters

Effective user reporting mechanisms are essential for maintaining platform integrity, protecting at-risk users, and enabling rapid responses to harmful content, conduct and contact – including CSEA. Poorly designed reporting tools can deter users from reporting, delay critical interventions and erode trust in your service.

Design principles for reporting functions

Discoverability

- Reporting functions should be easy to find, clearly visible, and have available directly on the platform. Navigation and functionality should be consistent across all devices and access channels, providing a seamless experience throughout the reporting process.[46]
- Users should be able to report directly from the material or activity they are concerned about – such as a post, comment, account, direct message, livestream or video call.
- Reporting options should not be in hard-to-find locations – for example, buried behind multiple clicks or taps, or within complex help articles about seeking support.

Usability

- Reporting should be simple, fast, and easy to use. Users are less likely to report if the process feels ineffective or cumbersome.[47]
- Use clear categories (such as 'Child Sexual Abuse or Exploitation') and pre-filled workflows to guide users through each step.
- Allow users to add free-text descriptions. This supports more accurate and authentic reporting – particularly when an incident does not fit neatly into a predefined category. It can provide feedback to guide the design of reporting functions. It also allows users to report material they know exists but can't directly locate (for example, an intimate image they know is being shared on the platform, but don't know where or by whom).
- Do not require users to upload or recreate harmful content to report it.
- Enable people to report without needing to register or log in. This is important where a person affected – such as the parent of a victim – is not an account holder on the service.[48]

[46] Many of the eSafety industry codes and standards include these specifications on the location of reporting mechanisms.
[47] eSafety's research into young gamers found most thought that reporting to the platform had little effect. Research from the ADL Centre for Technology and Society found that of people who were physically threatened online, 48% didn't report it to the platform because they didn't think the platform could or would do anything about it.
[48] This is a requirement under section 13 of the Basic Online Safety Expectations.

Context sensitivity

- Consider the **social and psychological barriers** that may discourage people from reporting – especially children and marginalised groups.
- Ensure anonymity is meaningful. Avoid including contextual information that could unintentionally identify the person making the report.
- Provide **feedback loops** to build trust and transparency. Acknowledge receipt of the report, share status updates and clearly communicate the outcome.

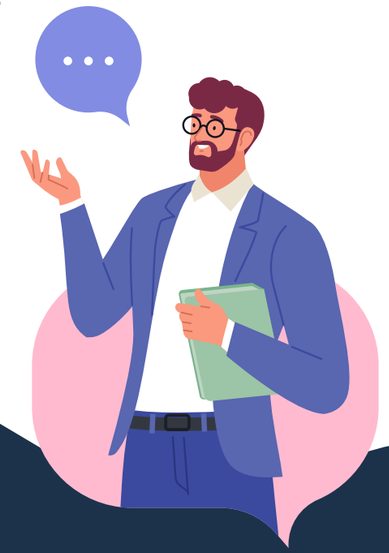
Common reporting deterrents

Avoid reporting designs that:

- make reporting hard to find or access
- require multiple steps or information the reporter may not have – for example, the name of a child in an image or video
- require personal information to submit a report
- require the user to create or log into an account
- use vague or ambiguous reporting categories
- force users to revisit harmful material (such as uploading screenshots or reviewing content)
- fail to provide feedback or clear timelines for resolution
- expose the reporter's identity or increase the risk of social reprisal

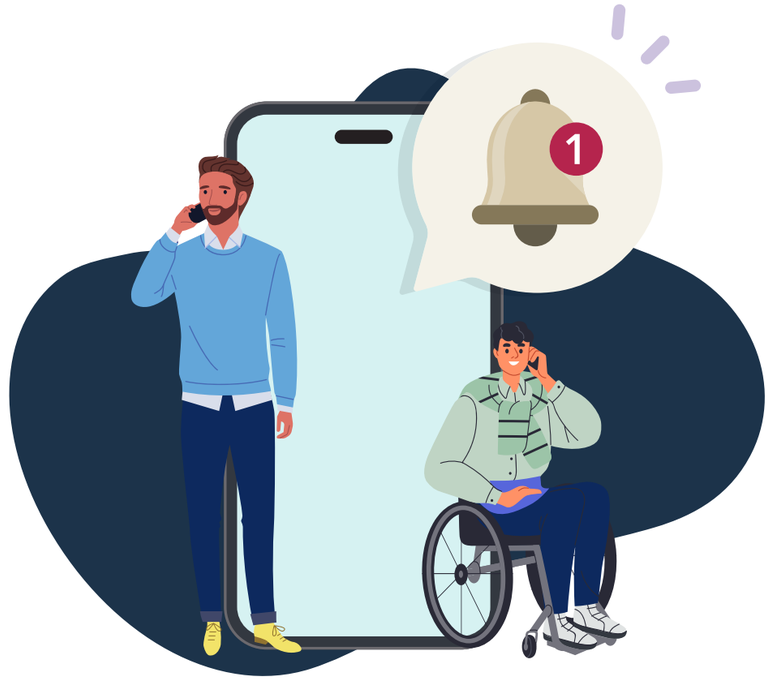
Implementation checklist for tech teams

1. Embed reporting tools directly into platform features (for example, in-app, in-chat, in-video).
2. Ensure consistent user experience in reporting across all access points – app, mobile, desktop and web.
3. Prioritise visibility and ease of use, especially for children and other at-risk or vulnerable users.
4. Simplify reporting processes – particularly for those in vulnerable or high-risk groups.
5. Use in-product nudges to remind users they can report and include direct pathways to do so.
6. Support both structured categories and open-ended input fields.
7. Use visual guides (such as screenshots, videos) to support users with varying literacy levels.
8. Allow anonymous and account-free reporting.
9. Enable reporting across all service elements – such as content, accounts, messages and livestreams.
10. Assess the need for mandatory personal information fields in reporting forms.
11. Build feedback loops with clear timelines and status updates. Monitor and review these loops regularly to ensure they remain effective.
12. Provide users by making reports with appropriate contact information for law enforcement, hotlines, regulatory bodies or other relevant authorities. Make it clear that reporting to the platform is still necessary so the service can take appropriate action, such as banning users or removing violating content. The role of law enforcement and other authorities is distinct.
13. Offer information about third-party referral services (such as mental health providers) at the point of reporting. Where possible, these services should be localised.



Appendix 2: List of referenced and additional resources

Resource	Developed by
Safety by Design framework and tools including, but not limited to: • Overview • Safety by Design Foundations • Start-up self-assessment tool and enterprise tool • Business model canvas • Free short course developed with RMIT University and eSafety • Investor checklist Regulatory guidance: • Online Safety Codes and Standards • Basic Online Safety Expectations • Image-Based Abuse Scheme • Online Content Scheme	eSafety Commissioner
• Developer Good Practices: Combating Online Child Sexual Exploitation and Abuse • Assessing OCSEA Harms in Product Development • Trust: Voluntary Framework for Industry Transparency	Tech Coalition
Safety by Design for Generative AI: Preventing Child Sexual Abuse	Thorn, All Tech is Human
Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse	Five Country Ministerial
AI and Child Protection: A Collaborative Approach to a Safer Future	Safer AI for Children Coalition
Combatting Online Child Sexual Exploitation and Abuse in Gaming	Digital Thriving
Real Time Threats: Analysis of Trust and Safety Practices for Child Sexual Exploitation and Abuse (CSEA) Prevention on Livestreaming Platforms	Centre for Democracy and Technology
Combating the Sexual Exploitation of Children for Financial Gain – Financial Crime Guide	AUSTRAC
CyberTipline Reporting API Technical Documentation	NCMEC
Typology of Online Harms	World Economic Forum
Terminology guidelines for the protection of children from sexual exploitation and abuse	End Child Prostitution, Child Pornography and Trafficking of children for sexual purposes (ECPAT)



Appendix 3: Terminology

Age assurance: An umbrella term covering technologies used to determine a user's age. The methods used offer different levels of certainty. Some confirm a person's actual age – often verified against an external source – while others give an age range or estimate to support an age-appropriate experience.

Age estimation: Measures that infer an approximate age or age range without other confirmed sources of information about the individual. This can involve the use of biometric data such as facial scans or other information such as behavioural patterns to estimate a person's age or age range.

Age verification: Processes that determine a person's age to a high level of certainty, typically by confirming data against an external source. An example of age verification is using physical or digital government identity documents to verify a person's age.

AI-generated child sexual abuse material (AIG-CSAM): Visual depictions of sexually explicit conduct involving a child, created or facilitated by generative AI technologies.[49]

Children: Individuals under the age of 18.

Child-on-child sexual abuse: Sexual activity between children that occurs without consent, without equality (whether mental, physical, or age-based), or as a result of physical or emotional coercion.[50]

Child sexual abuse material (CSAM): Any representation – regardless of format – of a child engaged in real or simulated explicit sexual activity, or a representation of a child's sexual parts for primarily sexual purposes. While the laws of many countries continue to use the term 'child pornography', there is a global shift toward the term 'child sexual abuse material' (CSAM) to more accurately reflect that creating or sharing such material is a form of child sexual abuse.[51]

Child sexual exploitation and abuse (CSEA) includes both:

- **child sexual exploitation** – material and activity that sexualises and exploits a child, but may not involve sexual abuse
- **child sexual abuse** – activity which involves sexual assault against a child (a subset of child sexual exploitation)

[49] Thorn, Safety by Design for Generative AI, accessed 30 January 2026, URL: https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf
[50] Saprea, n.d., 'Child on child sexual abuse', accessed 30 January 2026, URL: https://saprea.org/blog/child-on-child-sexual-abuse/
[51] World Economic Forum, 2023, Typology of Online Harms, accessed 2023, URL: https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf

End-to-end encryption (E2EE): A specific method for securing communications from one device, or 'end point', to another. E2EE transforms text, images, or audio into an unreadable format on the sender's system or device, which can only be decrypted by the recipient's system or device.

Grooming (in the context of CSEA): Predatory behaviour intended to manipulate a child into trusting the perpetrator so sexual exploitation or abuse can occur.

Hash: A process that breaks a digital file into small parts and combines them to produce a unique numerical value, often called a 'digital signature' or 'digital fingerprint'. This value can be used to identify or match the original file.

Hash database: A database that stores hashes of images or videos. In the context of CSEA, these databases contain hashes of confirmed ('known') CSEA material.

Hash-matching: The use of digital technologies that compare photos or videos against a database of previously identified ('known') material that has been marked with a hash ('digital fingerprint'), allowing copies or versions to be rapidly confirmed and removed without being viewed.

Homomorphic encryption: A form of cryptography that allows computations on encrypted data without needing to decrypt it first.

National Center for Missing and Exploited Children (NCMEC): Based in the US, NCMEC maintains hash lists of confirmed CSAM and exploitative imagery, which are made available to service providers to help detect this content. NCMEC also operates the CyberTipline, which receives CSEA reports from the public and industry and shares them with law enforcement agencies around the world.

Self-generated CSAM: Sexually explicit content of a child created by that child. It is often created under coercion or enticement by adults (including through grooming, sexual extortion or offers of money). In some cases, the material is initially consensually shared (for example, between peers in a relationship) but later distributed without consent.

Red-teaming: The practice of stress testing systems (physical or digital) to identify flaws, weaknesses, gaps and edge cases (rare, unexpected or extreme user behaviours and experiences).

Safety by Design: Safety by Design is a key online safety initiative which eSafety has driven since 2018. Safety by Design seeks to shift the onus of safety from the user to service providers, encouraging providers to promote and embed safety at all stages of the technology product lifecycle through three principles:

- service provider responsibility
- user empowerment and autonomy
- transparency and accountability.

These principles guide providers to anticipate, detect and eliminate online harms before they occur by incorporating, assessing and enhancing user safety.

Sexual extortion ('sextortion'): A form of online blackmail in which someone tricks or coerces a victim into sending sexual images or videos of themselves and then threatens to share it with others unless the victim meets their demands. These demands are often to send payment (sometimes known as 'financial sextortion'), or to send more sexual content or interact sexually online (such as in a video call).