



Best Practice Guide: Implementing Warning Messages for CSAM Prevention

About this guide

Technology platforms fight a constant battle against illicit content. Static content blocking leaves defensive gaps. Proactive behavioural interventions offer a stronger approach, by disrupting attempts to access content and redirecting users to support that enables them to change their behaviour.

This guide provides a set of best practice principles for designing and deploying automated warning messages to disrupt and deter the viewing and distribution of child sexual abuse material (CSAM). These principles are based on the latest research being conducted worldwide with a range of technology companies.

The guide describes: when and where messages can be deployed; how they should be designed; the phrasing and wording to use; and, finally, several key pitfalls to avoid.

Implementing these guidelines requires industry collaboration. The [CSAM Deterrence Centre](#) (based in Australia) and [Project Intercept](#) (based in the United Kingdom) invite technology platforms globally to partner directly with our research and practice teams.

When and where messages should be deployed

Effectiveness depends on delivering the message precisely when a user initiates a high-risk action. Situations when the warning should be triggered extend beyond simple keywords to capture a broad range of situational contexts.

- **Diverse Trigger Points:** Implement triggers for flagged search terms, attempts to access known CSAM URLs, uploading material which matches known CSAM hashes, and the activation of grooming detection algorithms.
- **Conversational & Contextual Analysis:** Use natural language processing (NLP) and multimodal analysis to identify risk signals in text, images, or metadata exchanges within interactive environments like messaging platforms and comment threads.
- **Layered Defence:** Deploy warnings as part of a defence in depth strategy that includes content blocking, search-term suppression, and friction mechanisms (e.g. delay till next request following trigger). Deploying a warning message alone is not a complete strategy.
- **Escalation:** Instead of static triggers, implement a sequenced intervention pathway. If risky behaviour persists, the system should trigger progressively more assertive responses rather than repeating the same message.
- **Differentiate Outcomes:** When setting success criteria, distinguish between immediate outcomes (e.g., abandoning a search) and intermediate outcomes observed over time (e.g., longer delays between repeat attempts).

How they should be designed

A warning is only effective if it is noticed, understood, and acted upon.

- **Visual Salience:** Use high-contrast colours (e.g. standard safety colours), bold typography, and

recognisable hazard symbols (e.g., "!") and signal words (e.g. Warning or Caution) to capture immediate attention.

- **Interruptive Formats:** Messages are best paired with a friction mechanism, requiring an exit or a delay before repeating actions. This can include full-page interstitials or pop-ups that require a user response (e.g. clicking "Exit") before proceeding, or a delay in their ability to attempt similar actions again.
- **Clarity:** Keep messages clear, concise, and direct using simple language. Research suggests that excessively long messages are a major deficiency and likely to be ignored.
- **Message Variation:** Periodically vary the visual design, wording, and framing of messages to counter "warning fatigue" and habituation. Dynamic messaging increases the novelty and can capture user attention. Monitoring message performance continuously to rotate and change messages to mitigate warning fatigue.

What wording to use

The tone and perceived authority of a message are as critical as its visual appearance. The context in which the message will be deployed affects how it is received. Partnering with us enables your team to access the latest research on how to frame wording for different online environments.

- **Source Credibility:** Where appropriate, attribute messages to authoritative sources, such as law enforcement agencies, regulators, or reputable NGOs (e.g. Internet Watch Foundation).
- **Signpost to Support:** Include non-punitive, anonymous links to help-seeking resources and therapeutic services for users concerned about their behaviour.
- **Legal Framing:** Clearly state the illegality of the behaviour and mention the potential for consequences, when contextually appropriate.
- **Proportionate Response:** Ensure that the tone matches the risk level to preserve user trust and avoid unnecessary escalation for accidental or one-off triggers.
- **Interactive Engagement:** Integrating an interactive chatbot can increase behavioural interruption and provide a bridge to support services.

Key Pitfalls to Avoid

- **Avoid Sensationalism:** Do not use ambiguous, eroticized, or sensationalist imagery/language, as this can inadvertently trigger a "forbidden fruit effect" and increase curiosity.
- **Avoid Excessive Punishment:** While legal warnings are effective, overly punitive or shaming language can trigger defensiveness and reduce help-seeking behaviour.
- **Minimise False Positives:** Use context-aware filtering (e.g. Bayesian classifiers) rather than simple keyword lists to preserve user trust and avoid over-blocking benign content.
- **Avoid Deception:** Users respond negatively to messages that attempt to deceive them. If using a chatbot, be transparent about its automated nature.
- **Static Deployment:** Avoid "set and forget" implementations. Failing to adapt to evolving user behaviours can lead to declining effectiveness and the emergence of avoidance tactics.

Partner With Us

To effectively deter CSAM, we need to share knowledge and work together across the industry. When we act alone, our impact is limited. The [CSAM Deterrence Centre](#) (based in Australia) and [Project Intercept](#) (based in the United Kingdom) work directly with technology platforms on a pro bono basis to implement these behavioural interventions safely and effectively. If you are interested in partnering with us, please reach out. Together, we can work globally to create custom warning systems and measure how well they work in real-world situations.

Contact the CSAM Deterrence Centre at: contact@csamdeterrence.com

Contact Project Intercept at: Intercept@lucyfaithfull.org.uk

Learn more about the evidence for warning messages and how to design them at csamdeterrence.com