Protect Children

Research Report

Preventive Pathways

How warning messages interrupt and prevent online child sexual abuse at scale

Findings and recommendations from a large-scale experiment

Protect Children 2026

Acknowledgements

This report was prepared for Protect Children by Tegan Insoll, Caoilte Ó Ciardha, Joel Scanlan, and Juha Nurmi.

Overall supervision and strategic oversight were provided by Anna K. Ovaska and Nina Vaaranen-Valkonen.

Funding

This report was produced as part of a project led by Protect Children and funded by the Home Office.

Part of Caoilte Ó Ciardha's contribution to the study was funded by the Tech Coalition Safe Online Research Fund (Grant No. 23-EVAC-0015.2-University of Kent).

Part of Juha Nurmi's contribution to the study was funded by the European Commission under the Horizon Europe funding programme, as part of the project SafeHorizon (Grant Agreement 101168562).

About Protect Children

Protect Children is a non-governmental organisation driven by the mission to create a world free from child sexual abuse and exploitation. The organisation, based in Helsinki, works internationally to prevent harm before it happens, support victims and survivors, and advocate for stronger protections for children.

Learn more: www.protectchildren.fi/en

Cite this report

Protect Children. (2026). Preventive Pathways: How warning messages interrupt and prevent online child sexual abuse at scale. www.protectchildren.fi/en/post/preventive-pathways-csam-warnings-research-report.

Cover photo by Siarhei Nester from Pexels.

About the authors

Tegan Insoll is the Head of Research at Protect Children, leading innovative research on the prevention of online child sexual abuse and exploitation. Her work focuses on understanding the impact of abuse on victims and survivors, understanding offending behaviours, and developing evidence-based prevention and early intervention strategies.

Dr Caoilte Ó Ciardha is an independent forensic psychology researcher and consultant, and an Honorary Reader (Associate Professor) in the School of Psychology at the University of Kent. His work focuses on understanding and preventing sexual harm, particularly in online environments, including help seeking, deterrence, and emerging risks linked to generative AI misuse. He serves as an Associate Editor of the journal Sexual Abuse and collaborates with technology companies, NGOs, policymakers, and academic partners on evidence-informed approaches to child safety.

Associate Professor **Joel Scanlan** is the academic co-lead of the CSAM Deterrence Centre (an initiative led by Jesuit Social Services, which operates Stop It Now! Australia), working with the technology industry and other stakeholders to prevent the access and sharing of child sexual abuse material (CSAM). His disciplinary background is in cybersecurity, but he collaborates closely with multidisciplinary teams spanning psychology, law, and criminology to prevent and disrupt child abuse online.

Security researcher **Dr. Juha Nurmi**, a Postdoctoral Research Fellow at Tampere University, works to tackle online child sexual abuse. He particularly studies online users who access child sexual abuse material. Furthermore, he maintains the main Tor search engine, Ahmia.fi, enabling searches for onion websites while filtering CSAM and redirecting CSAM users to mental health resources.

Anna Ovaska is the Deputy Director of Protect Children. Anna contributes to the comprehensive prevention of all crimes of sexual violence against children by engaging in research, legislative analysis, and the drafting of statements and papers.

Nina Vaaranen-Valkonen, M.Soc.Sc, is the Executive Director of Protect Children and a Social Psychologist and Cognitive Psychotherapist. She brings over 30 years of experience in the health sector, dedicated to protecting children from all forms of sexual violence, particularly online child sexual abuse and exploitation.

Contents

At a glance:

- This report presents **new evidence from a large-scale study on child sexual abuse material (CSAM) warning messages**, showing how they interrupt and prevent online child sexual abuse at scale.
- Over a 140-day period in 2025, there were **more than 3 million blocked CSAM-related searches** on the Ahmia.fi dark web search engine.
- Each blocked search **triggered a warning message** linking to an anonymous help resource.
- Overall, 12% of warning messages resulted in a click to a help resource, equating to more than **380,000 clicks** in total.
- We tested multiple warning messages and found that **all outperformed a neutral message**.
- The **average click-through rate increased** from around 9% to around 16% after the messages were redesigned.
- We **surveyed over 6,000 CSAM seekers** and found that **the impact of warnings extends beyond immediate clicks**.
- Among CSAM seekers who had encountered a warning, three in four said it had some **effect on their behaviour over time**, and one in five engaged with the linked resources.

Executive Summary

Warning messages are increasingly used by online platforms to deter individuals from accessing child sexual abuse material (CSAM), and direct users to support. By disrupting engagement with CSAM and encouraging help seeking, these interventions help reduce harm to children, victims, and survivors. However, there is limited real-world evidence on how the content of these messages shapes behaviour, particularly in high-anonymity environments.

This report presents findings and recommendations from a large-scale CSAM warning message experiment on Ahmia.fi, a search engine operating on the dark web, drawing on data from more than 3 million searches for CSAM and subsequent warning message exposures. The study examined whether changes to warning message content influenced engagement with an anonymous self-help program, using a combination of randomised message testing and platform-level analysis over time. A separate self-report survey, conducted after the study period, asked individuals who searched for CSAM on Ahmia.fi about their responses to warning messages.

Across all analyses, warning message content had a clear and consistent effect on behaviour. Every message tested increased engagement with help resources compared to a neutral baseline. At the platform level, the redesign of warning messages was associated with a substantial increase in help seeking: the proportion of users clicking through to the support program rose from 8.7% to 15.7%. This corresponds to over 115,000 additional help clicks during the campaign, demonstrating that even small changes to message design can translate into meaningful behavioural shifts at scale.

Survey responses corroborated and extended these findings. Among CSAM searchers who recalled previously seeing a warning, 17.3% reported clicking on help or information links. This indicates that warning message interactions were not merely passive or incidental but reflected deliberate engagement with support resources. Effects extended beyond the immediate interaction: 39.4% reported reflecting on their behaviour, 16.4% reported stopping in the short term, and 15.2% reported stopping in the long term, and 6.4% sought help. Many participants who initially continued searching still reported later reflection or behaviour change.

The scale of effect produced here, from a single search engine, means that investment in deterrence pathways and investment in the services that receive their referrals are connected policy decisions.

These results identify practical and scalable opportunities for the design of online safety interventions, with implications for governments and regulators, industry, and civil society:

- Warning messages should be recognised as core prevention infrastructure within online safety and child protection frameworks, providing a low-cost, scalable method for reaching those at risk of causing harm.
- Platform expectations should move from presence to performance, beyond whether a warning is displayed to whether its content has been evaluated.
- Warnings should be treated as dynamic interventions and be regularly tested, refined, and rotated to reduce habituation and maintain effectiveness.
- The warning interface should make the support pathway the dominant visible action at the point of intervention, while suppressing competing or distracting content.
- Warning messages should be deployed across contexts, including in encrypted and privacy-focused environments as part of a layered prevention strategy, recognising that they remain effective even where user identification is not possible.
- Classifier precision should be improved to support more proportionate and tailored responses to different forms of risk and intent.

Embedding evidence-based warning messaging across digital platforms at scale could create millions of opportunities to interrupt harmful behaviour, reduce demand for CSAM, and prevent harm to children, victims, and survivors.

Full design, analyses, and results of the Ahmia.fi message intervention study are available in the accompanying research paper.

www.protectchildren.fi/en/post/preventive-pathways-csam-warnings-research-report

1. Introduction

The scale and accessibility of CSAM is a severe and growing threat, with serious consequences for victims and survivors. Children whose abuse is recorded and shared often suffer enduring trauma, as images remain accessible and public over time.[1] Today, CSAM is accessible across a wide range of online environments, from mainstream platforms to privacy-focused networks such as the dark web.[2] This has contributed to a substantial and persistent level of online offending, which demands urgent attention.

CSAM is increasingly understood as a preventable public health problem. While traditional approaches focus on identifying and removing illegal content, there is growing recognition of the need to intervene earlier, at the point where individuals are actively searching for or attempting to access such material. Deterring CSAM-seeking behaviour and diverting individuals towards support at this stage have the potential to interrupt escalating offending pathways and limit the revictimisation of victims and survivors through repeated circulation of abuse material.

Recent studies have pointed to a rise in CSAM-seeking behaviour among young adults and adolescents. This is reflected in a recent survey by Protect Children, in which 45% of respondents (9,239 of 20,592) were aged 18 to 24. In addition, 11,206 individuals under the age of 18 attempted to participate but were screened out of the study.[3] These findings highlight the importance of intervening early in the offending pathway. Preventing and redirecting CSAM seeking behaviour before it becomes more entrenched or escalates is an important opportunity to prevent harm and reduce the risk of further offending.

Digital platforms provide a unique opportunity to interrupt harmful behaviour in real time and direct users towards support. One increasingly used approach is the

---

[1] Finkelhor, D., Turner, H., Colburn, D., & Mitchell, K. J. (2025). Persisting concerns about image exposure among survivors of image-based sexual exploitation and abuse in childhood. *Psychological Trauma: Theory, Research, Practice, and Policy*, 17(Suppl 1), S88–S93. https://doi.org/10.1037/tra0001815; Hanson, E. (2017). The impact of online sexual abuse on children and young people. In J. Brown, Online risk to children: Impact, protection and prevention (pp. 97–122). *Wiley Blackwell*. https://doi.org/10.1002/9781118977545.ch6; Canadian Centre for Child Protection. (2024) Experiences of child sexual abuse material survivors: How technology companies' inaction leads to fear, stalking, and harassment https://protectchildren.ca/en/resources-research/experiences-of-child-sexual-abuse-material-survivors-report/.

[2] Protect Children. (2024). Tech Platforms Used by Online Child Sexual Abuse Offenders. https://www.suojellaanlapsia.fi/en/post/tech-platforms-child-sexual-abuse.

[3] Protect Children. (2026). CSAM Perpetrator Research Report: Findings from a Survey of CSAM Perpetrators on Digital Platform Use and Design (Tell Me More About Tech). https://www.protectchildren.fi/en/post/tmat-csam-perpetrator-research-report.

deployment of warning messages.[4] These messages are triggered when users enter high-risk search terms or attempt to access prohibited content. In addition to signalling that the behaviour is harmful or illegal, warning messages can provide immediate pathways to help resources. In doing so, they create opportunities to redirect individuals away from harmful behaviour and towards prevention-focused interventions.[5]

By increasing engagement with perpetration prevention and support services, warning messages may contribute to longer term reductions in offending behaviour. Evidence from digital perpetration prevention initiatives has shown promising results, with randomised controlled trials demonstrating that online interventions can significantly reduce CSAM viewing and related risk factors.[6] The impact of warning messages may therefore extend beyond the immediate interruption of harmful searches by increasing access to programmes that support sustained behaviour change over time. Given that some CSAM seekers may also pose a risk of contact offending against children,[7] increasing engagement with effective prevention services may help reduce the likelihood of future abuse, protect potential victims, and prevent further harm before it occurs.

Despite the widespread implementation of CSAM warning messages across digital platforms, there is limited large-scale evidence on how the content of these messages influences behaviour, particularly in high-anonymity environments such as the dark web. Most existing deployments have been implemented without rigorous or transparent evaluation, leaving important questions about effectiveness, optimisation, and best practice unanswered.

This report presents findings from a large-scale intervention on a dark web search engine. The intervention builds on existing efforts to prevent access to CSAM and to promote help seeking among users at risk. By systematically testing how different warning messages influence engagement with an anonymous self-help program, this study provides new evidence on how digital interventions can support prevention pathways in high-anonymity online contexts.

---

[4] Protect Children. (2025). Online Warning Messages for CSAM Prevention: Evidence and Practice Mapping Report. https://www.suojellaanlapsia.fi/en/post/online-csam-warnings-report; Watters, P. A., Scanlan, J., Prichard, J., & Wortley, R. (2026). CSAM Desistance via AI, Chatbots and Automated Warnings. *Electronics*, 15(11), 2281. https://doi.org/10.3390/electronics15112281.

[5] Lucy Faithfull Foundation. (2026). Project Intercept Impact Report. Trigger for Change: The power of warnings to prevent online child sexual abuse. https://www.lucyfaithfull.org.uk/project-intercept/.

[6] Hillert, J., Haubrock, L. S., Dekker, A., & Briken, P. (2024). Web-Based Initiatives to Prevent Sexual Offense Perpetration: A Systematic Review. *Current psychiatry reports*, 26(4), 121–133 https://doi.org/10.1007/s11920-024-01489-1; Lätth, J., Landgren, V., McMahan, A., Sparre, C., Eriksson, J., Malki, K., Söderquist, E., Öberg, K. G., Rozental, A., Andersson, G., Kaldo, V., Långström, N., & Rahm, C. (2022). Effects of internet-delivered cognitive behavioral therapy on use of child sexual abuse material: A randomized placebo-controlled trial on the Darknet. *Internet interventions*, 30, 100590. https://doi.org/10.1016/j.invent.2022.100590.

[7] Insoll, T., Ovaska, A., Nurmi, J., Aaltonen, M., & Vaaranen-Valkonen, N. (2022). Risk Factors for Child Sexual Abuse Material Users Contacting Children Online. *Journal of Online Trust and Safety*. https://doi.org/10.54501/jots.v1i2.29.

2. Methods

This report brings together two complementary studies to understand how warning message content influences behaviour. The first is based on real-world behavioural data from a large-scale field experiment, while the second draws on self-reported survey data from users searching for CSAM. Together, they provide insight into both observed behaviour and user-reported responses.

Study 1: Message test experiment

The first study is a large-scale field experiment combined with a platform-level evaluation, implemented on Ahmia.fi, a search engine operating on the dark web, to examine how warning message content influences behaviour.

Study design

The study covered a 140-day period, consisting of a baseline period and a campaign period. Across this period, all searches containing blocked CSAM-related terms[8] triggered a warning message linking to an anonymous help resource, the ReDirection program.

70-day
Baseline period
an existing warning message was shown
23 June 2025 - 31 August 2025

70-day
Campaign period
new messages were introduced and tested
1 September 2025 - 9 November 2025

Warning messages

Prior to the campaign, users were shown a single, static warning message linking to a support resource. During the campaign, this was replaced with eight redesigned warning messages, varying in content and framing, alongside a neutral informational message.

[8] Ahmia.fi maintains a list of 1,281 banned CSAM-related search terms, based on existing CSAM detection lists as well as monitoring of platform-specific search patterns.

The redesigned messages drew on four key themes:

- **Legality and consequences:** Highlighting the legal risks and potential personal consequences associated with accessing CSAM.
- **Harm to children:** Emphasising the harm inflicted on children by viewing CSAM.
- **Behavioural control and self-efficacy:** Focused on the individual's ability to take control of their situation and access support.
- **Psychological distress:** Highlighting the potential emotional and psychological difficulties that may arise from viewing CSAM.

Each theme was presented in both a positive and a negative framing. Messages within each category shared the same heading but used differently framed content. All messages included the same closing sentence: "You can take the first step to change your behaviour by accessing support through the ReDirection program", and a link to the support program. The messages were displayed in English and Spanish. See Table 1 for messages.

Message testing

The study used two complementary approaches to assess how messages performed both under controlled and real-world conditions.

- **Randomised testing days:** To isolate the effect of message content, the platform included randomised testing days throughout the campaign. On these days, users triggering a warning were shown different messages at random. This allowed for direct comparison between message types under the same conditions, reducing the influence of external factors such as time or user behaviour.
- **Daily rotation:** Outside of the randomised days, messages were rotated on a daily schedule, with each day corresponding to a specific message condition. This ensured that all messages were shown across different days of the week. This approach allowed us to assess how messages performed in the context of a live platform over time.

Data and outcomes

The analysis used aggregated daily data from Ahmia.fi. No user-level identifiers were available, in line with the platform's privacy-preserving design. The primary outcome was engagement with support, measured as the proportion of warning messages that resulted in a click through to the self-help program. This data combines engagement with both the English and Spanish warning messages. We also examined changes in search behaviour at the platform level, based on the proportion of searches that triggered a warning message.

Table 1. Warning message content and framing used in the campaign

Category	Heading	Negative Framing	Positive Framing
Neutral	ReDirection Self-Help Program	ReDirection is a self-help program which aims to help you stop viewing sexual images of children.	
Legality	Child sexual abuse imagery is illegal.	Accessing sexual images of children puts you at risk of arrest and may cost you your relationships, your job, or your freedom.	Getting professional help may reduce the risk of arrest and help you keep your relationships, your job, your freedom.
Harm	Child sexual abuse imagery causes harm to children.	Searching for and viewing sexual images of children adds to that harm.	Getting help to stop viewing sexual images of children is one way you can stop that cycle of harm.
Control	Getting help for child sexual abuse imagery starts with a single click.	Don't dwell on the barriers stopping you from getting anonymous help.	It's easier than you think to get anonymous help to stop viewing sexual images of children.
Distress	How is searching for child sexual abuse imagery affecting you?	Take a moment to think about how searching for sexual images of children is likely to cause you feelings of shame, guilt, and anxiety.	Getting help to stop searching for sexual images of children can take away your feelings of shame, guilt, and anxiety.

The **ReDirection program** is an anonymous online self-help resource to prevent child sexual abuse perpetration developed by Protect Children. The program is available globally in multiple languages.

ReDirection has undergone two randomised control trials, with preliminary findings indicating that ReDirection is effective in reducing the use of CSAM. Results are pending publication.

Learn more: https://www.protectchildren.fi/en/redirection-program

Analytic approach

Three complementary analyses were used to assess the impact of the intervention:

- **Randomised-day analysis**, comparing message performance under random assignment
- **Campaign-period analysis**, examining patterns across the full deployment period
- **Interrupted time series analysis**, assessing changes in behaviour before and after the introduction of the redesigned messages

Together, these approaches provide a robust assessment of how message content influences behaviour, both at the individual interaction level and at the level of the platform as a whole.

Ethics

The message test experiment received ethical approval from the Ethics Committee of the Tampere Region (Finland) on 16 June 2025 (Statement 68/2025).

Full design, analyses, and results of the Ahmia.fi message intervention study are available in the accompanying research paper.

www.protectchildren.fi/en/post/preventive-pathways-csam-warnings-research-report

Study 2: Searcher survey

The second study is a self-report survey conducted after the 140-day message test experiment. This survey asked individuals searching for blocked CSAM-related terms on Ahmia.fi about their responses to and views on warning message content.

Recruitment

Participants were recruited between 4 March and 27 April 2026 via a survey link displayed in response to blocked CSAM-related search queries on Ahmia.fi.

The survey provided information about the study, requested informed consent to participate, and required confirmation that participants were aged 18 or over before proceeding. The survey was offered in English and Spanish. No compensation was offered for participation. To maximise participant benefit, the questionnaire provided links to relevant help resources.

Sample

The analytic sample included all respondents who provided substantive survey data (N = 6,172). The sample was predominantly male (77.1% of valid responses), mostly aged 18–29 (63.0% of valid responses), and geographically diverse, with respondents reporting being from 139 different countries. The largest proportions reported being from the United States (23.5%), India (8.9%), Germany (5.8%), France (5.2%), and the United Kingdom (4.4%). See Table 2 for the sample characteristics.

Ethics

The survey study received ethical approval from the Ethics Committee of the Tampere Region (Finland) on 4 March 2026 (Statement 23/2026).

Table 2. Sample characteristics

	N	%
Age		
18 to 29 years old	3471	63.0
30 to 44 years old	1518	27.6
45 to 64 years old	359	6.5
65 plus years old	161	2.9
Gender		
Man	4205	77.1
Woman	642	11.8
Non-binary or other	605	11.1
Language		
English	5703	92.4
Spanish	469	7.6
Country (top 10)		
United States	1007	23.5
India	384	8.9
Germany	249	5.8
France	223	5.2
United Kingdom	190	4.4
Russia	137	3.2
Brazil	130	3.0
Japan	118	2.7
Pakistan	111	2.6
Spain	111	2.6

3. Results

Across the 140-day study period, more than 3 million searches on the Ahmia.fi search engine included banned CSAM-related terms and triggered a warning. These warnings resulted in more than 380,000 clicks through to help resources.

→ 3 million

blocked CSAM-related searches on Ahmia.fi dark web search engine

2,200 per day

→ 380,000

clicks on the ReDirection program, anonymous help resource

2,700 per day

↑ 15.7%

redesigning the messages nearly doubled the click-through rate to the help resource

from 8.7% to 15.7%

↑ 115,000

additional clicks attributable to the redesigned messages

1,600 per day

Message content and help-seeking

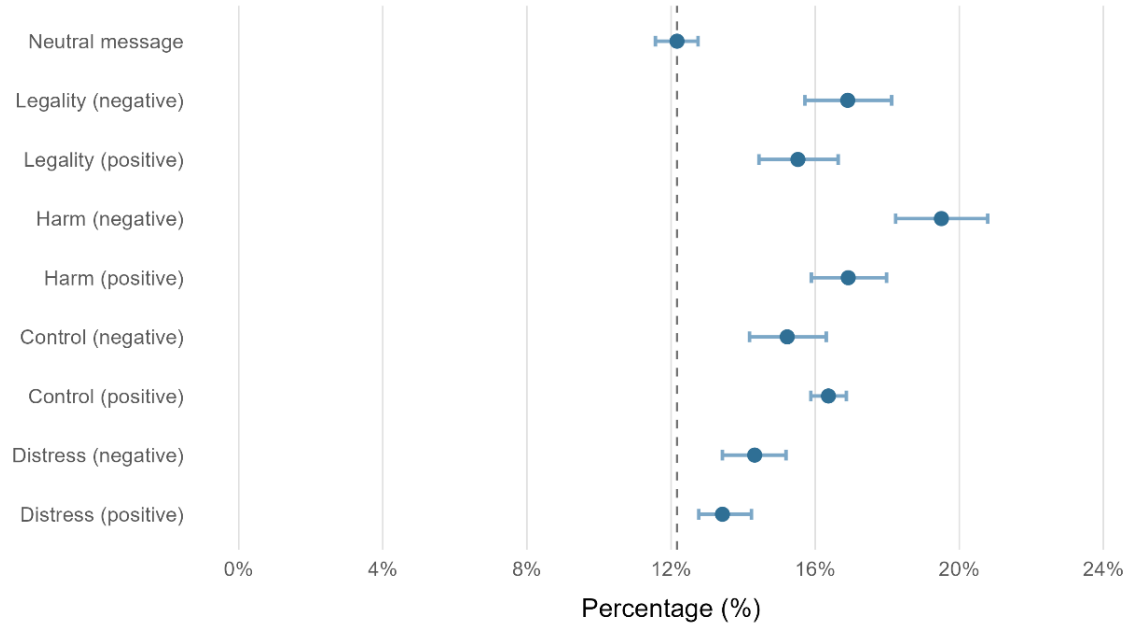
Message content affected whether users clicked through to support. Both on days where messages were assigned at random, and across the full campaign period, all tested messages resulted in higher click-through rates than the neutral message. The results were consistent between clicks on the English and Spanish messages.

On the **randomised days**, the negatively framed harm message had the highest click-through rates. The remaining messages—including the positively framed harm message—performed at a similar level to each other. Legality and control messages were within this range, while distress-focused messages showed smaller effects. Differences between messages were present but not large. See Figure 1.

Across the **campaign period**, the relative pattern—with the largest click-through rate associated with the negatively framed harm message—was consistent with the results from randomised days.

Figure 1. Click-through rates by message type on randomised days

Average daily click-through rates to the support program for each message type on days where messages were assigned at random. Points show the observed rates, and error bars indicate uncertainty. The dashed line shows the neutral message for comparison.



Platform-level change

As shown in Figure 2, click-through rates increased from 8.7% before the campaign to 15.7% during the campaign. This corresponds to over 115,000 additional clicks to support resources during the campaign period. The increase in help seeking was observed immediately following the introduction of the redesigned messages, with no evidence that it reflected a pre-existing trend, and was maintained over time. The neutral condition alone showed a click-through rate of 12.2%, exceeding the pre-campaign baseline of 8.7% (a relative increase of approximately 40%).

70-day
Baseline period

an existing warning message was shown

23 June 2025 - 31 August 2025

8.7% average click-through rate

126,053 clicks to support resources
Average 1,801 clicks per day

70-day
Campaign period

new messages were introduced and tested

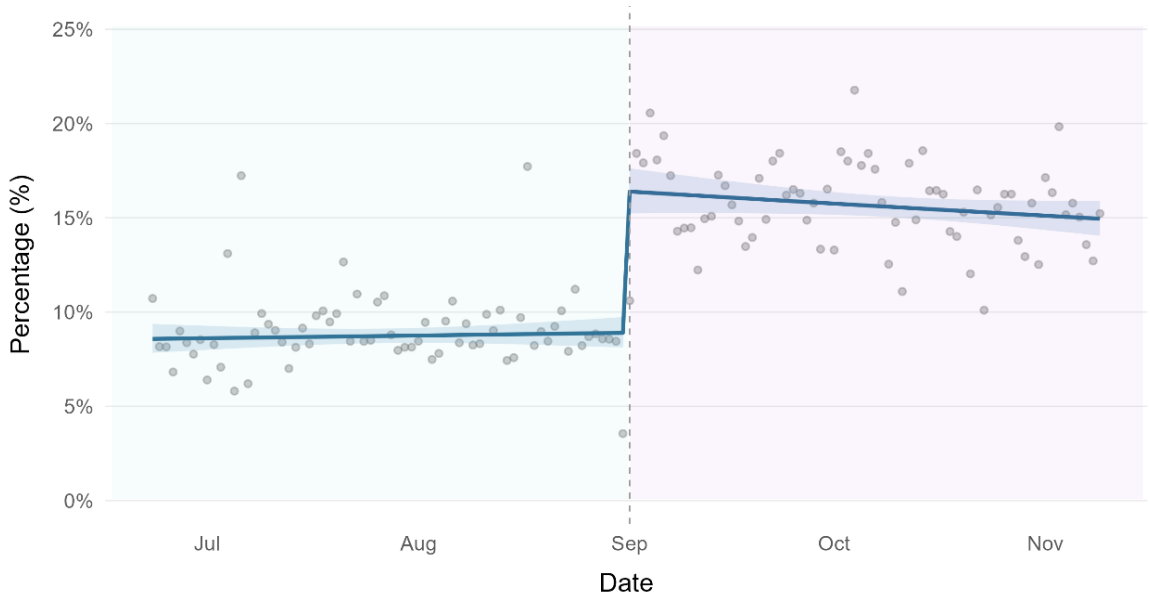
1 September 2025 - 9 November 2025

↑ 15.7% average click-through rate

↑ 260,444 clicks to support resources
Average 3,721 clicks per day

Figure 2. Increase in help seeking following warning message redesign

Daily click-through rates to the support program before and during the campaign period. The dashed vertical line marks the introduction of redesigned messages. The solid line shows the estimated trend, with shading indicating uncertainty.



Search behaviour

The campaign was not associated with a statistically significant change in the overall rate of CSAM-related searches triggering a warning message. However, the direction of the change was toward a decrease rather than an increase.

Survey results

Recalled exposure to warning messages: When asked whether they recalled seeing a warning message when searching for CSAM, 32.0% of respondents reported that they had seen one. This included 9.6% who recalled seeing one or two messages and 22.5% who recalled seeing many. The remaining 68.0% reported not recalling seeing any warnings.

Immediate response to warning messages: Among those who had seen a warning, immediate responses were mixed. While a majority (61.2%) reported ignoring the warning and continuing to search—either on the same platform (33.6%) or elsewhere (27.5%)—many respondents (38.8%) reported a positive response, including stopping searching for CSAM (21.5%) or clicking a help or information link (17.3%).

Long-term behavioural response: Longer-term patterns suggested greater reflection and disengagement over time. A majority of respondents (77.4%) reported that seeing a warning message had some effect on their behaviour over time: 39.4% reported that seeing a warning made them reflect on their behaviour, 31.6% reported that it made them stop searching for CSAM (16.4% in the short term and 15.2% in the long term), and 6.4% reported that it encouraged them to seek help. Only 22.7% reported that seeing a warning had no effect on their behaviour. Importantly, many participants who initially continued searching after seeing a warning still reported later reflection or behaviour change, suggesting that the effect of warnings is not limited to immediate responses.

Warning message effectiveness: When asked about the types of messages that would work to either deter behaviour or to encourage help seeking, preferences for message content were fairly evenly spread, although “harm” and “legality” were slightly more commonly selected. In terms of framing, there was a consistent but modest preference for positively framed messages.

Message testing: When participants were shown example messages and asked which would be most likely to influence their behaviour, legality-focused messages were selected most often (34% for stopping searching; 30% for encouraging help seeking). However, there was a clear mismatch between what people said they preferred in general and which messages they actually selected as most influential. Most participants chose a message that did not match their earlier stated preference, indicating that stated preferences alone are not a reliable guide to which messages people believe would influence their behaviour.

Table 3. Survey results

Percentages are based on valid responses for each variable; totals may vary due to item non-response.

	N	%
Recall seeing a warning message		
No, none	3698	68.0
Yes, one or two	522	9.6
Yes, many	1222	22.5
Immediate response to warning message		
Click help or information link	196	17.3
Stop searching	244	21.5
Continue searching elsewhere	312	27.5
Ignore and continue searching	381	33.6
Long-term response to warning messages		
No change	250	22.7
Reflected on behaviour	433	39.4
Stopped searching short term	180	16.4
Stopped searching long term	167	15.2
Sought help	70	6.4
Effective framing for warning messages to deter users		
Positive	2240	54.5
Negative	1868	45.5
Effective framing for warning messages to encourage help seeking		
Positive	2306	56.7
Negative	1758	43.3
Message test: warning message most likely to stop searching		
Legality	981	34.1
Harm	824	28.6
Control	565	19.6
Distress	510	17.7
Message test: warning message most likely to encourage help seeking		
Legality	889	30.2
Harm	759	25.8
Control	654	22.2
Distress	641	21.8

4. Discussion

This study shows that warning message design can interrupt and redirect users during CSAM-related searches.

Across more than 3 million warning exposures, messages with targeted content led to higher engagement with support than a neutral message. These effects were observed under randomised conditions and during live deployment. At the platform level, the redesign resulted in over 115,000 additional clicks to support resources.

In the survey, 17% of CSAM searchers who recalled seeing a warning message reported clicking on help or information links. This demonstrates that warning interactions were not merely passive or incidental clicks, but reflected deliberate initial engagement with support resources.

Message content makes a difference

Message content is not interchangeable. Small changes in wording produced consistent differences in behaviour, and these differences scaled across the platform.

At the same time, the results do not support a single “best message”. Our negatively framed harm-focused message had higher engagement, but the remaining messages performed at similar levels and all improved on the neutral baseline. Multiple approaches were effective. Whether combining message themes produces interactive effects is a question for future research, as is testing alternative themes.

For platforms, this means message design should not be treated as a one-time decision. It is a component that can be adjusted and improved.

Survey responses further underline the importance of empirical testing. There was a clear gap between participants’ general stated preferences for warning message themes and the specific messages they identified as most likely to influence their behaviour. Participants often endorsed one type of message in principle, but selected examples from different thematic categories when presented with actual warning messages. This indicates that stated preferences alone are not a reliable guide to message design, and empirical behavioural testing remains essential.

Primary impact was on help seeking

On Ahmia.fi, content blocking and warning messages have been in place for a number of years, already reducing the CSAM search volume by 57.7% since 2018.[9] The campaign was not associated with a statistically significant change in the rate of CSAM-related searches triggering a warning message, though the direction of the change was toward a decrease rather than an increase. The primary effect was therefore increased engagement with support. In this setting, the remaining opportunity was to influence what users do next rather than whether they search.

Platforms and policymakers should treat deterrence and diversion as distinct but complementary objectives. Where access to harmful content is already reduced, as in this case, further gains may come from strengthening pathways to support.

Effects extend beyond the immediate response

Survey responses suggest that the observed click-through rate reflects meaningful engagement. They also suggest that the observed rate understates the intervention's broader effect. Among respondents who had seen a warning, 17.3% reported clicking a help or information link in response. However, effects also extended beyond the immediate interaction: 38.8% reported reflecting on their behaviour, 16.4% reported stopping searching for CSAM in the short term, and 15.2% in the longer term. Additionally, 6.4% reported seeking help after seeing a warning message. Many participants who initially continued searching still reported later reflection or behaviour change. This is consistent with warning messages acting along a longer behavioural pathway than is captured by the click-through metric, and supports evaluation frameworks that look beyond the immediate response.

Importantly, the survey was visible only to users who had searched for a CSAM-related term triggering a warning, meaning that even participants who reported stopping behaviour or help seeking had not necessarily ceased CSAM-related searching altogether. This highlights the need for messaging and wider prevention efforts to recognise that behaviour change may be gradual and non-linear, including periods of reduced searching, continued searching, or relapse.

The findings emphasise that the contribution of warning messages is not only the immediate interruption of harmful searches, but also the increased engagement with support services at scale. This increased uptake supports longer term behaviour change by connecting more individuals to interventions with demonstrated effectiveness in reducing harmful behaviour over time.

---

[9] Nurmi, J., Paju, A., Arroyo, D., Scanlan, J. & Patsakis, C., Illicit Search Behavior and Responses to Filtering and Deterrence on Tor: Evidence from Ahmia. http://dx.doi.org/10.2139/ssrn.6329359.

Warnings are effective in high-anonymity environments, but effectiveness depends on context

Results show that messages can change behaviour even in high-anonymity environments. The intervention was deployed on a single search engine that indexes content on the dark web, but that has a clear anti-CSAM ethos. As a result, CSAM searchers on that platform may differ from people seeking CSAM on the clear web or more accessible platforms, as well as those seeking CSAM on less moderated dark web spaces. However, end-to-end encryption is now increasingly common on the clear web and in mainstream mobile applications, meaning that the distance between these populations has possibly narrowed in recent years. Nonetheless, users encountering warnings on Ahmia.fi are not a uniform group. They may differ in their level of persistence, familiarity with the platform, and willingness to change.

This has direct implications for other platforms, which operate in different parts of the ecosystem and reach users at different points. For example, messages that rely on detection or surveillance are less likely to be credible to privacy-aware users. On large mainstream platforms, users may be earlier in the pathway, where messages that highlight consequences may have a stronger deterrent effect. Where users are struggling or ambivalent, messages that emphasise their capacity to change or make support easier to access may increase engagement with support resources.

Underlying this is a more basic question about how confidently platforms can identify which users to intervene with. On Ahmia.fi, warning messages operate in a high-signal environment where CSAM-related searches can often be identified with relatively low ambiguity. On clear web platforms, however, false positives become a greater concern for both user experience and proportionality, particularly where excessive false positives may reduce trust in warnings, create unnecessary friction for legitimate users, or discourage platforms from deploying stronger interventions. Improving the precision of classifiers used to trigger warning messages can therefore enable platforms to intervene more confidently and support more tailored responses to different forms of risk and intent.

The findings have particular relevance for encrypted and privacy-focused environments. Ahmia.fi operates on the dark web with a privacy-preserving design, and yet warning messages produced consistent and substantial increases in engagement with support. This indicates that warning messages remain effective even when users deliberately operate outside conventional moderation and when the platform cannot rely on user identification to tailor or escalate the response.

This is complemented by recent results from Project Intercept, which highlight a successful partnership deploying warning messages on Mega, an end-to-end

encrypted cloud storage provider.[10] The warning messages meet millions of users each year and lead to meaningful engagement with the Stop It Now resource, showing that effective prevention interventions can still be implemented in high-anonymity and end-to-end encrypted environments.

The warning message interface itself is part of the intervention

On most clear-web search engines, warning messages appear above or alongside moderated search results. On Ahmia.fi, the warning replaces the search results entirely upon triggering. The click-through rates observed during the campaign (15.7% overall, with individual messages performing higher) are several times those typically reported on platforms that retain visible search results. Message content is part of what produced this effect, but the surrounding interface design likely also contributed. Even the neutral message increased click-throughs from 8.7% (baseline) to 12.2%, a 40% relative increase, indicating that a substantial part of the campaign's effect came from the redesigned warning presentation rather than message content alone. Where competing pathways are suppressed at the moment of warning, the support pathway becomes the primary available action. This has implications for platforms whose current implementations layer warnings over other content rather than replacing it.

Implications for policy and practice

The findings identify a practical and scalable opportunity for online safety interventions. Warning messages can increase engagement with support at scale. Message design can be improved without new infrastructure, and effects can be tested and refined using existing platform data.

For policymakers and technology companies, this supports investment in systematic testing of warning message content, integration of clear and accessible support pathways, and ongoing optimisation rather than static deployment.

The scale of the effect also has downstream implications. The 70-day campaign produced more than 115,000 additional click throughs to a single help resource, from one search engine. Increasing the quality and range of deployments of warning messages into mainstream platforms will substantially increase referral volume to support services. Investment in deterrence pathways and investment in the services that receive their referrals are therefore connected policy decisions.

---

[10] Lucy Faithfull Foundation. (2026). Project Intercept Impact Report. Trigger for Change: The power of warnings to prevent online child sexual abuse. https://www.lucyfaithfull.org.uk/project-intercept/.

5. Whole System Recommendations

1. Warning messages can effectively interrupt harmful behaviour at low cost and at scale.

Warning messages should be recognised as effective, scalable, and low-cost prevention measures. They could be required in any context and on any digital platform where content is searched for, accessed, or shared.

Governments and regulators

- Recognise CSAM warning messages as a core, scalable prevention measure within online safety and child protection frameworks.
- Consider requiring the deployment of warning messages in any context where users search for, attempt to access, share, or otherwise engage in behaviours indicative of CSAM-seeking activity.
- Position warning messages alongside detection and removal systems as part of a layered CSAM prevention strategy.

Industry

- Treat warning messages as a high-impact, low-cost intervention that can meaningfully change behaviour at scale.
- Ensure warnings are deployed consistently across relevant user journeys, including search, sharing, and access points.

Civil society

- Advocate for warning messages to be embedded across platforms as standard prevention infrastructure rather than optional add-ons.
- Engage at consultation stages of industry codes to ensure warning message requirements are explicit, comparable, and enforceable across jurisdictions, rather than relying on voluntary "best efforts" framings.

2. Platforms should prioritise evaluation of warning message effectiveness.

In addition to continuous improvement and tailoring messages to user populations, evaluations provide insight into the level of offending on a given platform to assist in broader prevention efforts. Platform expectations should move beyond "do you display a warning?" to "have you evaluated the warning's content?".

Governments and regulators

- Shift from presence to performance-based regulatory expectations.
- Encourage minimum standards for evaluation, including behavioural outcomes such as engagement with support services.

Industry

- Implement routine evaluation of warning message effectiveness, including message content, framing, and user response. Use findings to refine messaging and improve prevention outcomes over time.
- Publish aggregated evaluation outcomes (i.e., exposure volumes, click-through rates to support, and changes following design iteration) in transparency reports, in a form that supports cross-platform comparison without compromising user privacy or revealing information that would aid evasion.

Civil society

- Champion independent evaluation as the standard, not the exception.
- Convene cross-sector forums to establish shared evaluation methodologies and outcome metrics so findings are comparable across platforms and jurisdictions.
- Translate technical results into accessible briefings for policymakers, regulators, and the public, and ensure the voices of victims and survivors are reflected in how "effective" is defined.
- Partner with academic institutions to provide independent analysis of platform-published evaluation data and publicly recognise platforms that meet a high evaluation standard.

3. Warning messages should be treated as dynamic interventions which are updated and improved over time.

To mitigate habituation to a message, it is simple to change and rotate messages to maximise engagement and redirection to support services. Warning messages are not a set-and-forget, but a proactive intervention.

Governments and regulators

- Build iteration into regulatory expectations explicitly. In transparency frameworks specify that evidence of ongoing message testing, refinement, and rotation forms part of demonstrating compliance with "disrupt and deter" obligations.
- Avoid prescribing specific message wording; instead, require regular, structured reporting on what was tested, what changed, and the measured effect.
- Co-fund independent research and shared infrastructure so smaller platforms without in-house testing capacity can still meet the standard.

Industry

- Treat warning messages as dynamic, not static interventions.
- Regularly rotate and test warning message content to reduce habituation and maintain effectiveness over time.
- Tailor messages to different contexts and user groups where possible.

Civil society

- Provide input on message design to ensure content remains ethically grounded and avoids unintended harms. For example, avoid framings that risk reinforcing shame in ways that suppress help seeking.
- Partner with researchers to build evidence on how messages should be tailored to different user populations (those at different points in the pathway to harm, on different platform types, with differing levels of awareness) and provide those findings back into industry practice.

4. User interface design should maximise the visibility and accessibility of warning messages.

The visual prominence of the warning on Ahmia.fi is a sizable divergence from clear web environments where moderated search results are typically still shown—this is a likely reason for the high click-through rate in this study. This indicates that interface design, not only message wording, is likely to be an important part of the intervention. Consolidating links and information and minimising other content shown can improve warning effectiveness.

Governments and regulators

- Endorse design principles for warning interfaces in industry codes and standards so that warnings are the dominant element at the point of intervention and pathways to support are clear, accessible, and easy to act upon.
- Encourage platforms to minimise competing or distracting content, particularly advertising and unrelated navigation pathways, while allowing flexibility for curated safeguarding or educational content where appropriate.
- Frame these as outcome-based principles rather than prescriptive layouts, so platforms can adapt to their own product surfaces while meeting the underlying behavioural requirement.

Industry

- Treat the warning interface as a dedicated intervention environment, not a notice layered over the existing experience.
- Minimise competing or distracting content at the point of warning, including recommendations, unrelated navigation pathways, and commercial content. Where search results must remain visible, prioritise curated safeguarding, educational, or help-seeking resources.
- Consolidate harm framing, legal information, and the support link into a single message with one clear primary action.

Civil society

- Conduct and publish usability research on warning interfaces, including with people with lived experience of help seeking.
- Develop reference designs and pattern libraries that platforms (particularly smaller ones) can adopt or adapt, lowering the bar to high-quality implementation.
- Hold platforms publicly accountable when warning interfaces are buried, ambiguous, or surrounded by competing content that materially undermines their effect.

5. Warning messages should be deployed across encrypted and privacy-focused environments.

The findings show that warnings can be effective and can change behaviour even in high-anonymity and privacy-focused environments, such as the dark web.

Governments and regulators

- Recognise warning message interventions as a privacy-compatible component of CSAM prevention in encrypted contexts.
- Treat them as distinct from purely detection-based measures, whose technical capability and rights trade-offs remain contested.

Industry

- Where content scanning is limited, move the intervention upstream to the point of search query, link sharing, or intent signal, rather than abandoning the warning approach altogether.
- Test message designs that are credible to privacy-focused users, particularly those that do not rely on framings of detection or surveillance.

Civil society

- Build evidence on the effectiveness of warning messages in encrypted and privacy-focused environments, where user populations and platform constraints differ from those of mainstream search and content services.

6. Classifier precision should be improved to support proportionate and tailored interventions.

Improving the accuracy of systems that trigger warning messages, with more precise and behaviourally informative classifiers, can help ensure that interventions are better matched to the type and level of risk.

Governments and regulators

- Recognise that the effectiveness and proportionality of warning interventions depend partly on the quality and precision of the classifiers used to identify high-risk searches or behaviours.
- Encourage investment in classifiers that reduce false positives and support more context-sensitive interventions, particularly on platforms where legitimate and harmful searches may coexist.

Industry

- Routinely refine the classifiers used to trigger warning messages, with particular attention to reducing false positives in mixed-intent environments.
- More precise classifiers can support stronger interventions where appropriate, while also enabling more tailored responses to different patterns of activity, including information-seeking, help seeking, tentative CSAM-related searching, and more entrenched, sophisticated, or high-severity forms of CSAM-seeking behaviour.

Civil society

- Support independent research and scrutiny around how classifiers are used to identify and respond to potentially harmful searches and behaviours.
- Advocate for approaches that are proportionate, transparent, and facilitate tailored intervention in high-risk contexts.