

# Chatbots and warning messages: innovations in the fight against online child sexual abuse

By Sarah Smith



# Contents

- 
- 3 **Who we are**
  - 4 **Executive summary**
  - 6 **Introduction: why a chatbot and why now?**
  - 8 **A history of our deterrence messaging and campaigning**
  - 10 **The reThink chatbot experiment**
  - 21 **The future: Chatbot 2.0, Project Intercept and more**
  - 23 **Recommendations for the tech sector**
  - 24 **Appendices**
  - 28 **References**

# Who we are

---

## How we help keep children safe

We work to reach adults and young people to prevent abuse from happening in the first place – and, if it already has, to prevent it from happening again.

Where abuse has already taken place, we work with all those affected, including adult men and women who have abused; young people with harmful sexual behaviour; children with concerning sexual behaviours; and victims of abuse and other family members. But we also work with families and with adults and young people where there has been no abuse, to help them keep themselves and others as safe as possible.

We run the Stop It Now UK and Ireland helpline. A confidential service available to anyone with concerns about child sexual abuse, including adults worried about their own or someone else's sexual thoughts, feelings or behaviour towards children. And we run Shore, a website that provides a safe space for teenagers worried about their own or a friend's sexual behaviour.

---

## The Faithfull Papers

We research and evaluate our work to make sure what we do protects children, and we share the evidence with professionals and the public. We want to make best use of our expertise, our data and our insights, independently and in partnerships, to develop new strategies and interventions that help keep children safe.

We advocate for a greater focus on preventing abuse before it happens and for a public health approach to the prevention of child sexual abuse.

The Faithfull Papers are a series of reports showcasing our understanding of what works to protect children to the widest possible audience – to policymakers, journalists, researchers and partner organisations in the UK and overseas.

**Around one in 6 children will be sexually abused.**

**Around one-third of this is carried out by under-18s.**

**And around 9 in 10 children who are sexually abused know their abuser.**

**At the Lucy Faithfull Foundation, we work to stop child sexual abuse before it happens.**

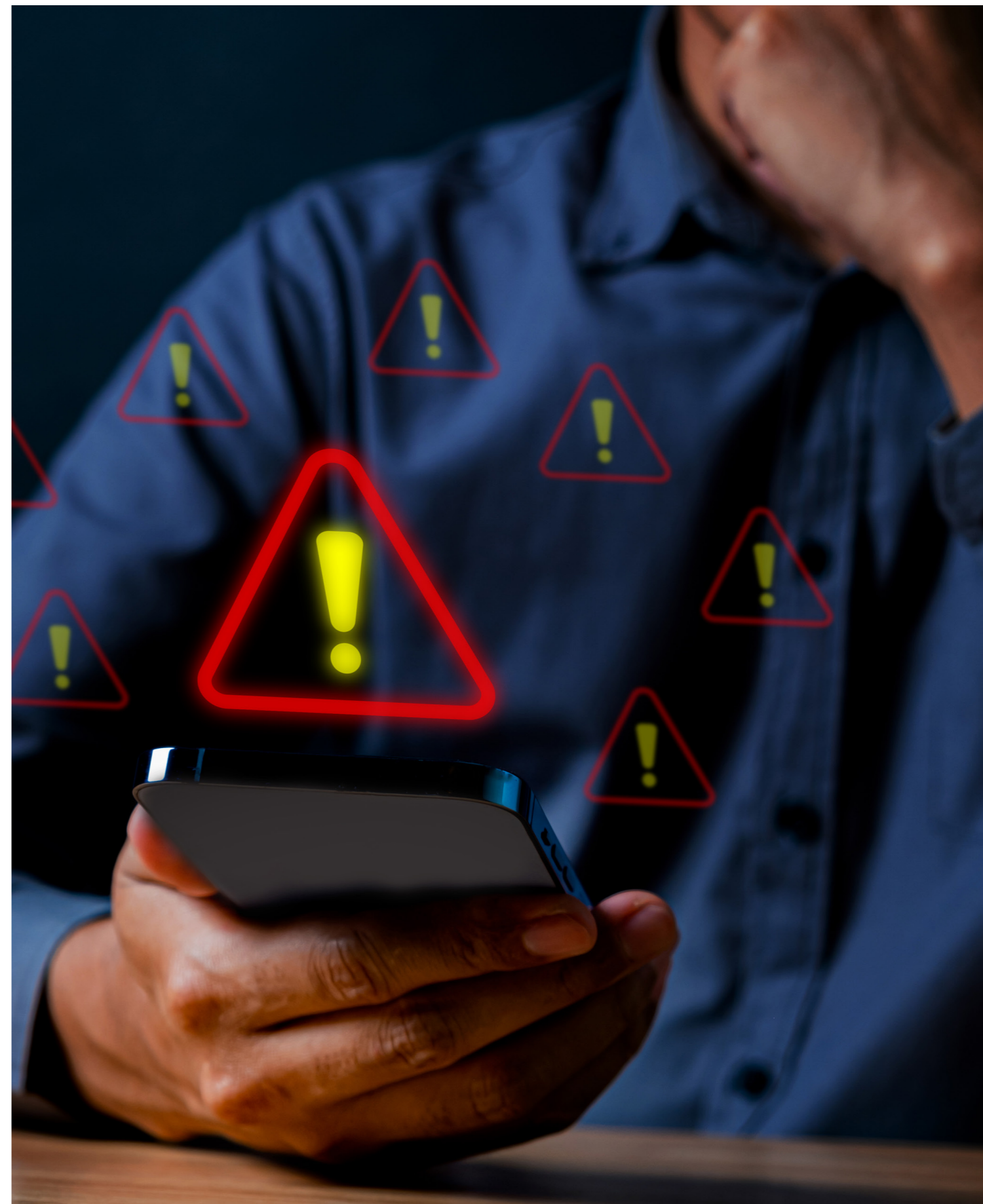
**We're here for everyone who needs us.**

# Executive summary

- The reThink chatbot experiment was an innovative collaboration between Lucy Faithfull Foundation, the Internet Watch Foundation (two child protection charities) and Aylo (formerly Mindgeek), an online adult pornography company. The aim was to see whether warning messages and a chatbot could reduce the number of searches for sexual images of children on the Pornhub UK website. It ran from March 2022 to August 2023.
- The results showed that the warning message and chatbot did have a deterrent effect on people looking for sexual images of under-18s on the site.
- One message was enough for people in 82% of the sessions that triggered a warning to stop searches for illegal images in that session. In some sessions users continued to search for legal pornography; some halted their online session completely; some left the site.
- The experiment has also shown that the warning message and chatbot can engage and encourage people to seek help and support from our Stop It Now services.
- When the chatbot was turned off for a month, searches for sexual images of children rose again on Pornhub UK.
- A small but persistent group of users were not swayed by the intervention and continued to search for sexual images of children. However, this number was a tiny proportion of those who were diverted away from their searches for illegal material.
- The chatbot experiment sits within our wider online deterrence work. From it we can see evidence of our deterrence approach working in two ways: for some people, instructing them that what they are doing might be illegal is enough to get them to stop their immediate behaviour, while others need more in-depth support from our Stop It Now helpline advisors and online self-help.
- Future directions for chatbot include refining and updating its dialogue abilities; rolling it out across further pornography sites; and developing a model that can be used in contexts other than adult pornography, for example, on social media or file-sharing platforms.
- Combining our expertise in preventing sexual offending with the expertise a tech company has about its own platform creates an effective intervention.
- The sharing of crucial data by Aylo has been key to the success of this experiment. This collaboration provides a blueprint for any other online tech company that wishes to improve safety on their platforms through sharing data and experimenting with the most effective ways of deterring those on a path to offending.

## Impact of the project by numbers

- **There was a statistically significant reduction in searches for sexual images of under-18s over the 18-month project**
- **The chatbot and warning message were shown together 2.8 million times**
- **99.8% of searches over the 18-month project did not trigger the chatbot/warning**
- **1,656 people requested details of our help services after seeing the chatbot/warning**
- **490 visits to our Stop It Now website after seeing a warning message or chatbot**
- **68 Stop It Now helpline callers were identified as having interacted with the chatbot**



# Introduction: why a chatbot, and why now?

Online child sexual abuse, which involves making, viewing or sharing sexual images of under-18s as well as online grooming, is a huge and growing problem both in the UK and globally.

In 2023 the Internet Watch Foundation (IWF) [identified and removed 275,652 sexual images of children](#) (1) from the internet, while the CyberTipline (run by the National Center for Missing and Exploited Children) received more than [36 million reports](#) of suspected sexual images or videos of children (2). As the numbers grow and with the rise of [AI-generated](#) material (3), the outlook is increasingly alarming.

Law enforcement agencies have long made it clear that this is not a problem that they alone can manage. The number of people believed to be viewing and sharing this material dwarfs the number who are actually arrested; the [NCA estimates](#) that there are up to 830,000 adults in the UK that pose some degree of sexual risk to children (4). Tackling the issue has to start with prevention, that is, deterring people from starting down a pathway to offending in the first place. [Our 2023 Faithfull Paper on our deterrence campaigns has more information on this approach](#) (5).

In October 2020 the Safe Online Initiative at End Violence Against Children agreed to fund a first-of-its-kind chatbot project proposed by the IWF in partnership with Lucy Faithfull Foundation (LFF). The idea was to develop a chatbot (a piece of software designed to simulate conversation with a human) that could engage and divert people seeking sexual images of children, using IWF's technical expertise and LFF's knowledge of offender motivations and behaviour. Aylo agreed that Pornhub UK, the most popular adult pornography site in the country, could be used for the chatbot test, and the ambitious and pioneering reThink Chatbot experiment launched in March

2022. [Sexual images of children are prohibited on Pornhub](#) (6), however, many users still attempt such searches, some of them multiple times. We had already collaborated with Aylo to present a static warning message on the site since February 2021 aimed at encouraging people to come through to our Stop It Now services. The hope was that the chatbot would boost these numbers even further by being a more engaging intervention, as well as continuing to deter people who were attempting to view illegal images. It may also be the case that the deterrent effect of encountering a warning on Pornhub would give the user enough motivation to curb any other searches or attempts to access illegal material on other websites.

ReThink Chatbot was an unusual collaboration between two charities and a pornography company, but it was to prove fruitful. Over a period of 18 months, data about the chatbot's interactions and the actions of people who viewed the warning message were collected by Aylo, and the data then analysed and evaluated by researchers at the University of Tasmania. What they found provides a crucial insight in a field where there is relatively little published evidence – the effectiveness of warning messages on those seeking to access child sexual abuse images and videos. Before this work, there was no evidence at all about the efficacy of chatbots on stopping people from trying to view sexual images of under-18s.

The chatbot was part of our ongoing commitment to prevent child sexual abuse online – something we have done for many years with our deterrence campaigns and warning messages hosted by tech companies (see below). It couldn't have come at a more crucial time.



It's surprising that there are very few warning messages online to deter people looking for sexual images of under-18s (7). This is in sharp contrast to our everyday life, where we have learnt to accept warning messages as a part of living in a modern society. Digital roadside signs flash our speeds at us if we drive too fast. Sobering TV ads remind us of the dangers of drink-driving. Warnings are so widespread in society that they cover not just illegal activity but harms and risk too (from hazards such as deep water or overhead power lines to sending money via bank transfer to a new contact). They are an integral part of our approach to public health.

But online, even for something as illegal and as serious as searching for sexual images of children, warnings are patchy, often poorly implemented, or not there at all. Websites and tech companies routinely display a generic error message or 'not found'. This means that people searching for illegal content do so largely unchallenged.

Our work is already changing this. We have been pioneering online warning messages with tech companies since 2013. The messages work not only to deter and disrupt immediate illegal behaviour, but to get the right support to change

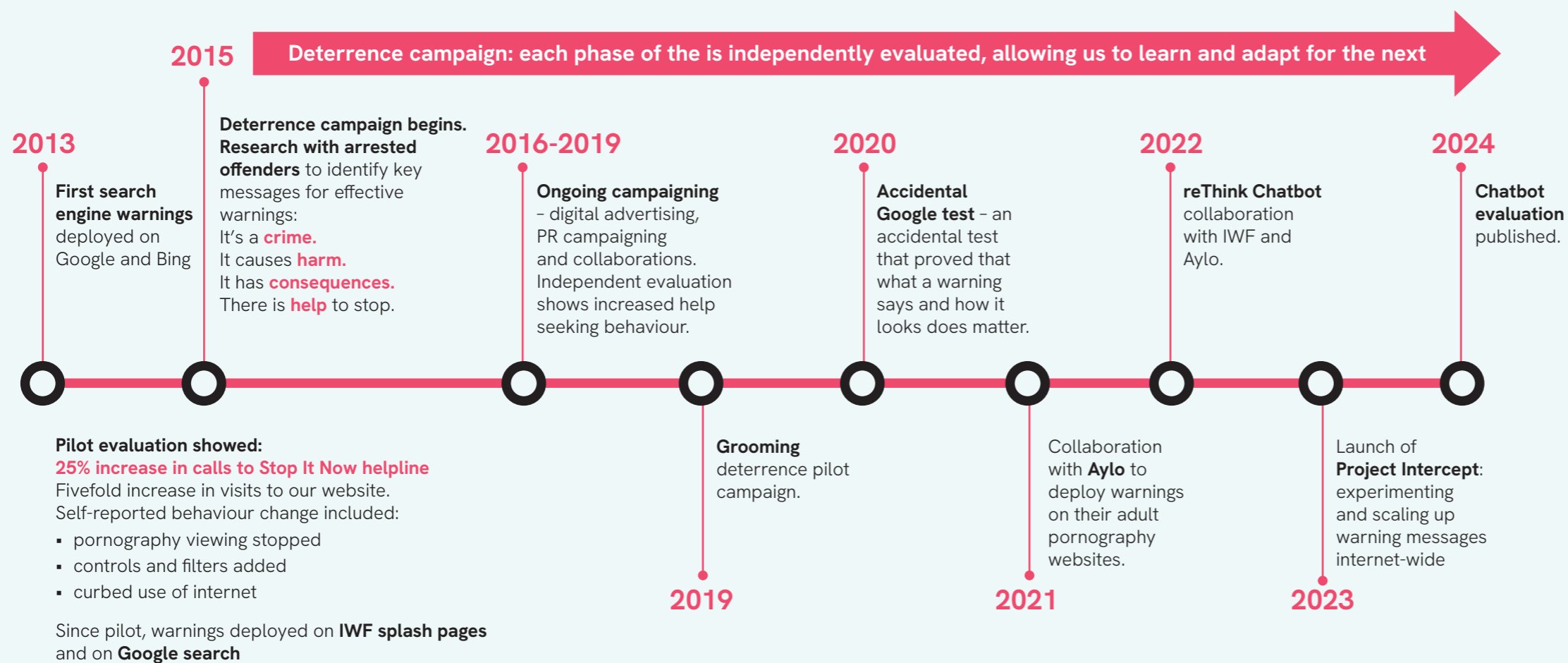
people's long-term behaviour at a point when they need it most. The onus should not be on children to protect themselves. Instead, the responsibility lies with a range of actors – the tech industry, governments, regulators – but importantly also on the people who are behaving harmfully or illegally. Why shouldn't these people be challenged, deterred, reminded, nudged and redirected? And why shouldn't tech companies be challenged to get on board and help make obvious changes that protect children? Warning messages should be the norm, not the exception.

A chatbot, therefore, was a natural progression from our work with warning messages. The question was: could it amplify the deterrence effect of the static warning? Would it stop people searching for illegal sexual images of children?

This report will look at the chatbot experiment, the evaluation of the chatbot and message, but also how the chatbot sits within our wider deterrence work. We will consider whether the chatbot achieved what it set out to do, and what lessons we can learn from this. Finally, we will look to the future, exploring how chatbot can be developed for future applications – and how we hope to get there.

# A history of our deterrence messaging

## Eleven years of deterrence messaging and campaigning



With these messages at its core, our first deterrence campaign ran in 2015 across a range of online and offline spaces, including conventional news media, social media, paid digital adverts, short films, partnerships with law enforcement and other statutory and voluntary organisations. At the end of the pilot campaign [an independent evaluation \(9\)](#) showed that there had been a 25% increase in calls to our Stop It Now helpline, and a fivefold increase in visits to our website. Those who had used our online self-help reported a range of successes, including curbing their use of the internet; stopping their viewing of pornography; and adding controls and filters to their devices to create barriers to offending.

Our deterrence campaign has continued and each iteration is independently evaluated and the insights derived used to inform the next wave of campaigning, for example, we most recently (November 2023-April 2024) focused on deterring people from viewing AI-generated sexual images of children (10).

What we know from our deterrence campaigns is that some people will seek help for their behaviour, if they know that help is available. We also know that effective support can mean people don't offend in the first place. This is part of the wider deterrence effect, which happens when people view news stories or adverts and change their behaviour without reaching out for help (and it is difficult to measure the impact of this). The challenge is to make sure that as many people as possible know about that help, at a time when they are ready to access it. Warnings delivered at the point of attempts to access, share or view illegal images are a key part of our strategy to achieve this.

Our first warnings were deployed in 2013 on Google and Bing search engines, and were triggered by keywords associated with sexual images of children. In 2015, as research showed that instances of online offending were increasing exponentially, we were given funding by the Home Office to design a deterrence campaign to test

whether early intervention could not only disrupt individuals at risk of offending but drive them to our helpline and online self-help resources. We based our deterrence work on our extensive knowledge of sexual offender behaviour, and through research with arrested offenders (8) identified four key messages that would create effective warnings.

These messages were that viewing sexual images of children:

- **is a crime**
- **causes harm to children**
- **has consequences for you and your family**
- **but we provide anonymous help to stop.**

# The reThink chatbot experiment

The reThink chatbot grew out of a discussion with the IWF, whose work it is to remove sexual images of children from the internet, around the need for better, more effective deterrence messaging for people who look for that material.

It is an unfortunate fact that these images and videos can be found relatively easily on the open web (11) including on some legal adult pornography websites.

Aylo, the Canadian multinational pornography company and owners of the website Pornhub, state that they have a zero tolerance approach to any material featuring under-18s and have a range of tools to identify and [remove any sexual videos of children](#) (12) from their sites. However, aware of the fact that searches continue to be made by people attempting to find this material, they agreed to host the chatbot experiment on Pornhub UK to see whether the rate of searching could be slowed.

The project, funded by End Violence Against Children, relied on the collaboration of the three partners: IWF, LFF and Aylo. The chatbot itself was programmed and maintained by the IWF, while LFF was responsible for the content of the chatbot's conversations, which were programmed by our clinical practitioners and psychology experts. All three partners shared data with the independent evaluators from the University of Tasmania (UTAS) to allow a detailed analysis of the actions of users who encountered the warning and chatbot. This agreement by an online tech company to share data so openly is at the heart of what allowed the chatbot experiment to proceed, and we hope that this open sharing of data becomes an industry norm as online tech works towards better protection of children on its platforms.

The chatbot ran uninterrupted from March 2022 until July 2023, when it was turned off for 28 days. At the time of publication of this report (September 2024) it is still active on the Pornhub UK website and engaging users in conversation on a daily basis.

## How the chatbot works

When someone searches for a banned term on the Pornhub website, they are presented with a static warning page telling them that the search term they have used is banned due to its association with sexual images of children. Aylo compiles and moderates its own keywords list, which is constantly being updated, and runs into the tens of thousands. The list is made up from a variety of sources, and comprises data from several NGOs, including the IWF's keywords list, which is available to IWF members. It also includes some data from law enforcement sources. The static warning is live on all versions of Pornhub across the globe, but on the UK site the chatbot also presents itself on the bottom corner of that page (or prominently on a mobile device), aiming to catch the user's attention, and inviting them to a conversation. The chatbot looks like the typical customer service chat option you find on shopping, banking, or utilities websites.

People can click predetermined buttons or can enter free text to talk to the chatbot. Generally, the conversations are most successful when users choose to use the predetermined buttons and the majority of users during the experimental period did so. The chatbot was constructed using Dialog Flow ES, a chatbot platform by Google enabling the easy development and deployment of chatbots for a range of different scenarios. Using text and natural language processing, the chatbot engages with the individual (see figure 1). Based on an automated analysis of the answers received from users of the chatbot, the chatbot provides responses that guide the user and signpost them to LFF services including our Stop It Now helpline,

## The pathway from legal pornography to child sexual image offending

**The collaboration between two charities focused on child sexual abuse prevention and the world's largest pornography platform may seem to be courting controversy. But our belief is that for prevention tactics to work best it is crucial to intervene early and in places where there is an established pathway for offending (our Faithfull Paper on pornography (13) examines this in more detail).**

While of course not all those who view adult pornography also view sexual images of children, the sequence of events from viewing legal adult pornography to viewing sexual images of under-18s has been regularly observed in our work with those who have offended or who are on a path to offending (8). There are several reasons why this move to seeking more extreme material might happen. The Coolidge effect (14) can help to explain a little further. This is when someone experiences a decline in sexual interest when sexual activity is restricted to one stimulus, but then discovers a renewed sexual interest with the introduction

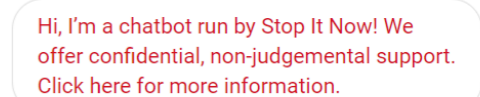
of a new stimulus (such as a new partner or new sexual material). The drive to seek this greater sexual gratification becomes linked to an escalation in viewing habits, including seeking out new and potentially extreme material (15). Almost half of young people under the age of 25 who had sought support to stop viewing illegal images self-reported risky pornography-viewing habits or an escalation of pornography viewing as the main reason for viewing the material (16).

Understanding different offender journeys is important if we are to prevent such illegal behaviour in the future. Some people who have offended state that they had a previous legal pornography problem that "got out of control". Others have a longstanding sexual interest in children. Deterring people who appear to be at risk of moving on to illegal content needs to be a priority, and this is why we worked with Aylo. Catching someone in the moment where offending is about to take place is a hugely important intervention point.

email live chat and online self-help. The user can select which option they prefer. Additionally, the chatbot can also suggest services offered by other agencies which the user might find useful, such as the National Suicide Prevention Hotline and NHS urgent mental health.

The service is anonymous. At the beginning of the conversation, the chatbot asks the user not to enter any personal information as the chat history is being recorded to help with improvements to the service. However, consistent with Stop It Now helpline procedures, if there is any actionable information that a child is in danger or a crime has been committed, information is forwarded to law enforcement authorities. To May 2024, no such information has been provided via the chatbot.

### Figure 1: Chatbot engagement



Hi, I'm a chatbot run by Stop It Now! We offer confidential, non-judgemental support. Click here for more information.

LFF staff review the chatbot logs regularly to see whether any interventions or tweaks to the chatbot's dialogue are needed. The chatbot was created and implemented before the recent explosion of interest in generative AI. All its dialogue is pre-written and it does not create new content in response to inputs from users, which mitigates the risk of inappropriate responses being produced by the chatbot.

### Evaluation

The work of evaluating the chatbot experiment fell to our partners at UTAS. We engaged an independent team led by Professor Jeremy Prichard, including Dr Joel Scanlan, Prof Richard Wortley and Dr Paul Watters. Together they examined the data collected from both the static warning and the chatbot from March 2022 to August 2023 and considered the following questions, proposed by them and agreed by IWF and LFF.

- Did the chatbot reduce searches for sexual images of children on Pornhub UK?
- Was the chatbot effective in engaging 'at-risk' users who have entered search terms indicating they are looking for sexual images of children on Pornhub UK?
- Did the chatbot increase the number of individuals seeking support from LFF?
- Has the experiment enhanced the evidence base concerning online offending around sexual images of children?

There was a huge amount of data generated over the course of the project, with the vast majority being provided by Aylo. Appendix 1 shows the data collected from the three partners.

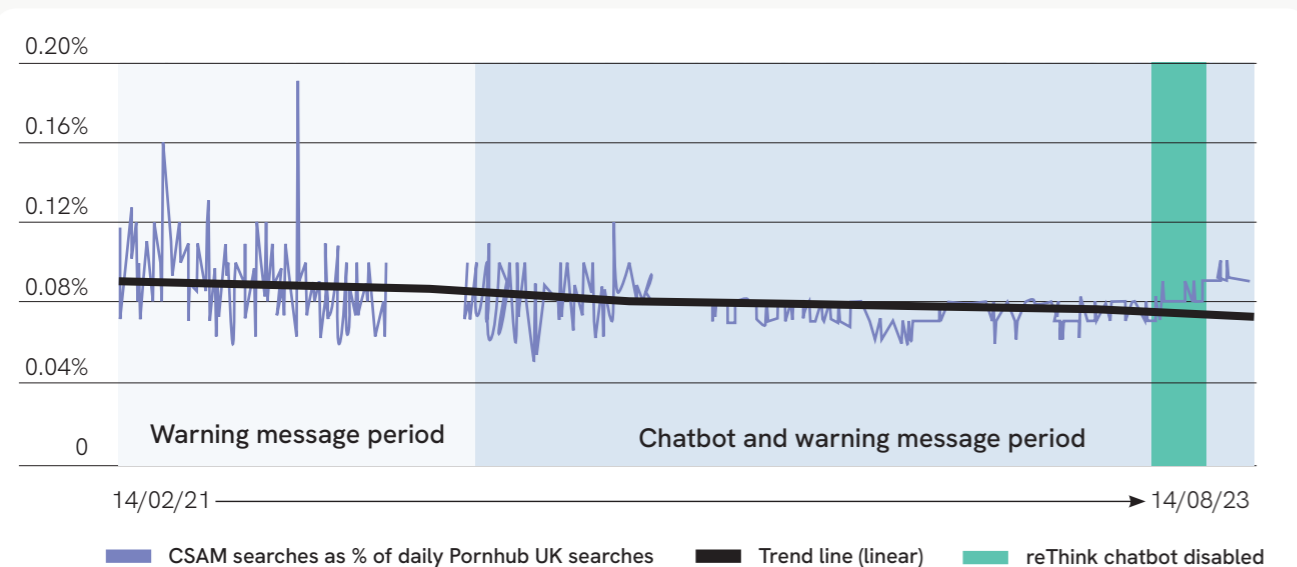
The data collected from Pornhub documented sessions rather than users. We are not able to track people, just sessions. A session is defined as a period of time spent on Pornhub by an individual user. So, this means that a single user might have several different sessions over the 18-month intervention, and it also means that some sessions triggered multiple warnings. Sessions vary in length from a few minutes to weeks at a time - depending on when a user cleared their browser of cookies. This means that there may be an element of double counting when we talk about how many people saw warnings, which should be a consideration when looking at the figures.

### Successes

Between March 2022 and August 2023 there were 2.8 million searches indicative of an attempt to find sexual images of children, or around 0.2% of the total number of searches on the site. During the data collection period more than 99.8% of sessions did not contain a search that triggered a warning, which means that the overwhelming majority of people using Pornhub UK are not searching for illegal material.

The second thing that came clearly out of the evaluation is that the data shows a clear decrease

**Figure 2: Percentage of daily searches triggering a warning**



### The numbers

**99.8%**

During the chatbot data collection period 99.38% - 99.87% of sessions did not contain a search that triggered the warning.

**2.8m**

IWF report 2.77m chatbot sessions. Aylo reported 2.2m sessions were shown warning.

**1,656**

1656 users said they were interested in Stop It Now services.

**490**

428 users from the chatbot to Stop It Now website. 490 sessions on Stop It Now website directly referred from the chatbot.

**68**

46 callers and chatters to the Stop It Now helpline identified as having interacted with the reThink chatbot. Between them these callers and chatters made a minimum of 68 contacts to Stop It Now.

in the number of searches for illegal sexual material over the course of the experiment (figure 2). This deterrence effect is a key aim of the project and answers the first of UTAS' evaluation questions. Although there may be questions as to whether any of this searching behaviour was displaced to other pornography platforms, we can see from the data that many users did stay on Pornhub UK after receiving a warning, watching videos and making other searches. What figure 2 also shows is that when the chatbot was disabled in July 2023, searches for sexual images of children rose again.

Digging down into the numbers, the 2.8 million chatbot triggers resulted in 1,556 requests for more information from Stop It Now services. Following that, people clicked through to our Stop It Now website 490 times, and there were at least 68 calls and chats to our Stop It Now helpline.

Although those numbers may seem small compared to the 2.8 million triggered searches, it is important to understand that they do not show the whole picture. Deterrence is measured not just by clicks through to our website or calls to Stop It Now - this simply reflects the number of people who happened to be ready, at that moment, to seek help. Deterrence is also about the number of people who positively change their behaviour after viewing a warning with no further input from us: in other words, those who stopped what they were doing. This is why we try to ensure that our warnings contain those four important messages mentioned on page 9: viewing sexual images of children is a crime, it causes harm, it has consequences, and we provide help to stop. We know that by including those key elements we increase the possibility that at least one of them will hit home or strike a chord with the user.

Our work with people who offend and those on a path to offending tells us that people seeking sexual images of children are a wide-ranging group with a variety of motivations. Not all are paedophiles (though some will be) (17). Some may use sexual material as a way to cope with emotional and relationship difficulties. Some may have a general atypical sexual interest that does not focus specifically on children. Some may be making a rash choice in the ‘hot moment’ of searching for sexual images while aroused. Some may be pushing at boundaries to see what is available; some may be driven by curiosity or by an enjoyment of engaging in the forbidden. Many of these will be deterred from their course of action simply from the appearance of a warning (and in fact the chatbot data proves this – see below). It is likely that the warning and chatbot drove more people to seek help than these numbers suggest, as some people may have chosen to access the services at a different time or after a period of reflection, and so were not captured in the above data. This hidden deterrence effect represents a methodological challenge which has always been difficult to measure.

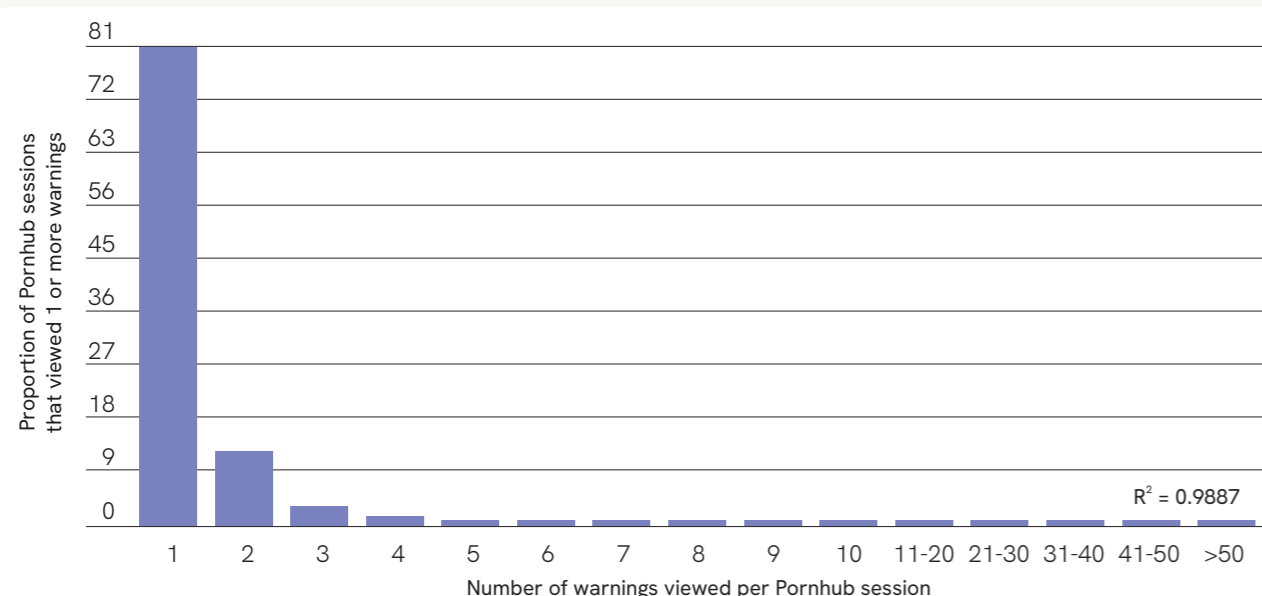
However, it is important to acknowledge the strong attrition rate here, and to see it as evidence that

we need to look further at ways to increase the early take-up of help – because we know that this help can make a huge difference in the lives of those on a path to offending. The section looking at future directions on page 21 has some ideas on how the messaging can be improved to increase engagement. We know from our work with people who have offended that they are more likely to acknowledge and take in a deterrent message when they are not in an aroused state, and we know that decision-making while aroused tends to be more impulsive (18), which is one of the reasons we make sure our messaging and campaigns are seen across a wide variety of media that people will encounter during different points in their everyday lives.

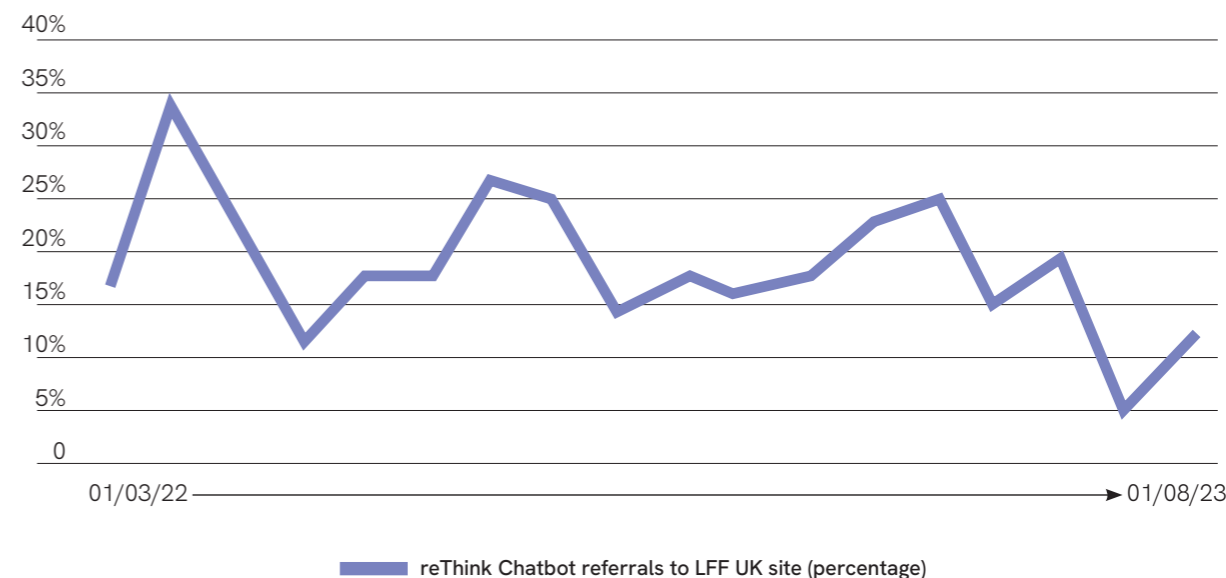
It has always been hard to show this deterrence impact – prevention in action – as it involves trying to prove a negative and show what someone didn’t do. However, because of Aylo’s willingness to share a wide range of data, we can see a hugely important statistic about the behaviour of those who were shown warnings.

One of the most striking things that shines through is that for 82% of sessions that triggered a warning, just one warning was enough to stop searches for illegal material in that session.

**Figure 3: Most sessions desist**



**Figure 4: Monthly user referrals from the reThink Chatbot to Stop It Now as a percentage of all new users to the Stop It Now Aylo UK-specific landing page**



UTAS called this the “desist” group (figure 3). They may have carried on their session on Pornhub by viewing a video, or searching further for a legal term, or they may have closed their browser or laptop, or navigated away from looking at porn altogether. What we know is that they didn’t input any further search terms that triggered the warning during that session. That is a powerful finding that lends weight to the argument for warning messages as a prevention tactic. It should be noted, of course, that a warning is only as good as the keyword list that triggers it – something to bear in mind for any future experiments with the chatbot. We also know that the chatbot may have pushed some people onto other sites to look for illegal images, which is why comprehensive warnings across the web are important.

Digging deeper into the data, UTAS examined when during the sessions warnings were typically triggered. They found that in those sessions where seven or fewer warnings were triggered the rate of triggering decreased as the session progressed. They also noted that a general pattern of behaviour was for people to search less after receiving a

warning (even searching for legal terms), although they stayed on the site and watched videos.

**Driving people to our helpline**

A major research question of the Pornhub experiment was to see if the people who triggered the warnings could be engaged to seek support from our website and helpline. To track this we created specific landing pages for Aylo (one globally and one covering just UK referrals) on our Stop It Now website that showed the traffic coming in from all our Aylo interventions (we have static warning messages across all Aylo platforms including YouPorn and Redtube). What we can see from this landing page data (see figure 4) is that over the course of the experiment there was an average of around 20% of referrals coming directly from the chatbot itself, with the rest coming from the static warnings from the range of Aylo sites. While it’s hard to say whether we would still have engaged those users with the warning message alone, we can certainly say that the chatbot helped to drive users to our website.



## Case study: Alex's story

**In 2022, Alex contacted the Stop It Now helpline on two occasions after receiving a warning\* on Pornhub during his online search for adult content. He said his specific focus was on images related to sexual domination. Alex asserted that his online activities did not involve viewing any illegal material and were solely focused on searching for legal adult pornography. Married with two children and a full-time job, he expressed fear of the potential legal consequences, prompting his call to the helpline.**

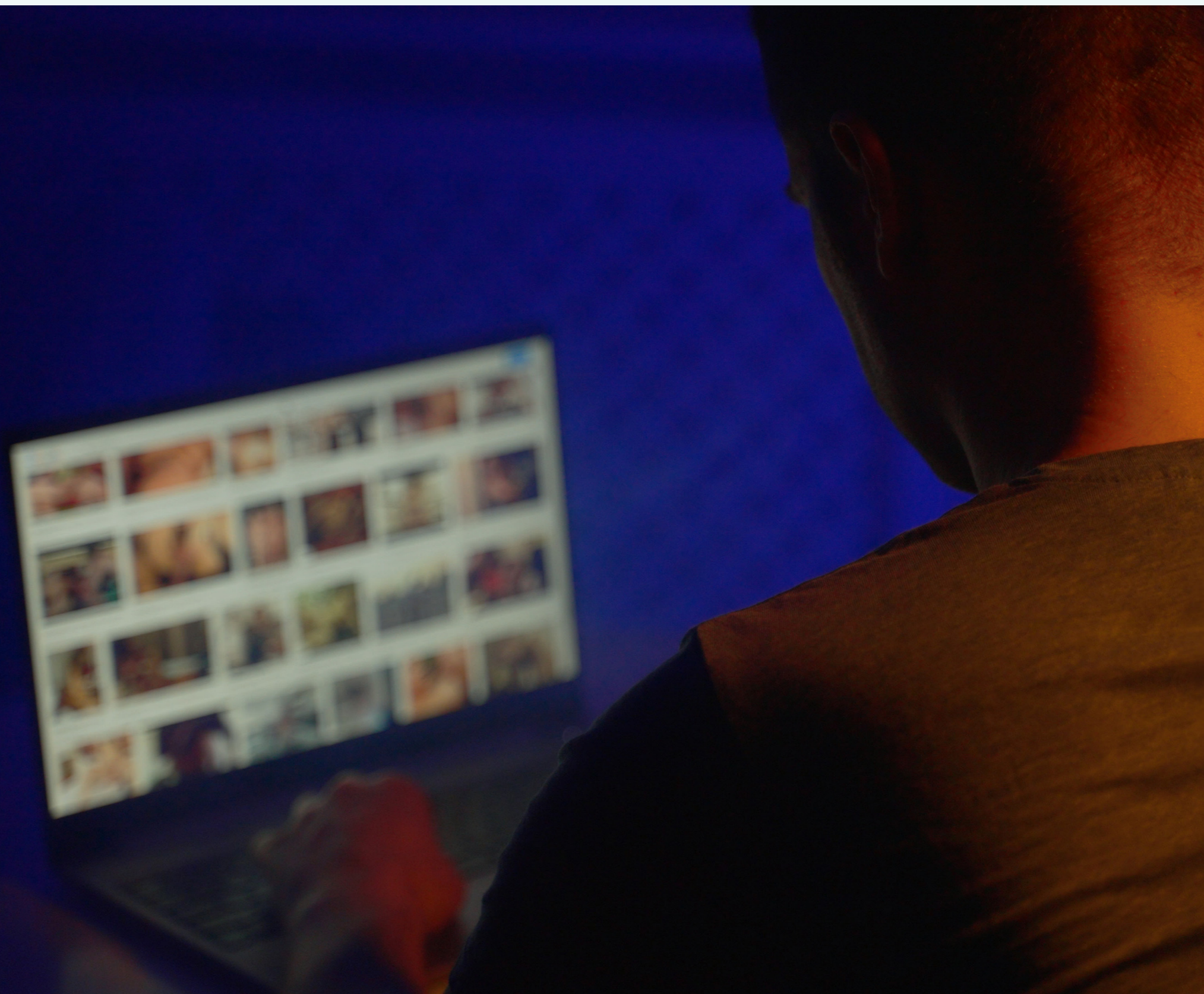
During the first contact, we explored any safeguarding issues, and having established that there were none, we advised Alex to reduce his internet usage and establish parameters to prevent inadvertent access to illegal content. Additionally, he was directed to our Parents Protect website to ensure the online safety for his children and the rest of the family. The helpline advisor also explored Alex's emotional state and encouraged him to engage in self-care activities.

In a subsequent call, Alex reported persistent anxiety related to the warning received a couple of days earlier. He explained that this anxiety was affecting his sleep and eating patterns. Alex was advised to seek support from his GP and was introduced to mindfulness exercises to alleviate anxiety and help with his sleep, which he found beneficial.

Alex shared with the advisor some of the search terms he used to seek adult pornography, one of which included terms like "against her will", which he suspected might have triggered the warning. It was explained to him that even if the individuals in the videos were adults, there might still be legal concerns regarding consent for the depicted sexual activity or being recorded. It was emphasised that search terms implying a lack of consent could trigger the warning, and that the warning's purpose was to attempt to avert any escalation in search behaviour. The advisor also provided information about our collaboration with Pornhub, utilising warning pop-ups as an opportunity for early intervention.

Reassured by the explanation, Alex decided to share his concerns with his wife and continue to implement internet parameters (he had told us on his first call that he had stopped viewing adult content from the moment he received the first warning).

*\*advisors did not distinguish between the static warning and chatbot in their discussions*



Some people who contacted our Stop It Now helpline mentioned seeing or engaging with the chatbot or warning message specifically. In Alex’s case the very appearance of the warning was the driving factor in encouraging him to make contact, as he was anxious that he would get into trouble with the law. Through repeated interaction with the helpline advisors he gained an understanding of the implications of his behaviour and of the role of the chatbot and warning message that left him in a good place to rethink his future actions.

Alex’s case and also the numbers of new users on our website shows there has been an evident benefit from the project in making people aware of our help services and in feeling empowered to click through or call.

Towards the end of the experimental period some changes were made that saw the static warning messaging altered and at one point, the chatbot was turned off all together to see what impact this would have. While the changes were too short-lived for any formal analysis, informally we can see clearly from the graph in figure 2 (page 12) that turning off the chatbot coincided with an increase in the numbers of searches for illegal material.

The different iterations of the warnings are laid out in Appendix 4.

“**Alex’s case and also the numbers of new users on our website shows there has been an evident benefit from the project in making people aware of our help services and in feeling empowered to click through or call.**”

### Limitations and challenges

There were a few issues that UTAS encountered during the evaluation of the experiment which have impacted the results.

#### Data challenges

Firstly, there was no data available to measure a baseline of searches that would have triggered the warnings before the introduction of the static warning alone in February 2021. This means our understanding of how the static message works alone is incomplete.

Secondly, as mentioned above, the messaging for the static warning was changed several times towards the end of the evaluation period in an attempt to measure the effect of alternative wording on people’s behaviours. UTAS felt these periods were too short to produce a reliable analysis. They came to the same conclusion when the chatbot was turned off in July 2023.

Thirdly, the project contained a mix of datasets, each with different timescales, formats and definitions of user sessions, which meant that often comparing like for like was hard.

All of these challenges point to a need to set clear parameters for any future testing scenarios and ensuring baselines are clearly measured before the experiment begins.

#### Complexities

UTAS noted that although the searches for sexual images of children on Pornhub UK decreased over time, so also did the rate of requests for information about LFF services. This suggests that the chatbot and warning messages were less effective over time at engaging users, and perhaps users who are long entrenched in problematic behaviour. There are implications for further experiments. UTAS suggested, for example, that the warning message could be rotated, with different texts/images/videos to stop users becoming habituated. They recommended that we consider integrating videos from our deterrence campaigns, which are high quality and created specifically to target those looking for illegal images of children.

Experimenting with designs that require interaction (for example, to click a button to remove the pop up – something which Meta does (19), or to make an active choice about what to do) or that use warning symbols or bold colours could also help to keep the warnings fresh and maintain engagement, especially if used on a rotational basis.

In addition to this, there were some results in the data that showed not all our deterrence work lands in the way that we would hope. Alongside the desist group mentioned above, there was another group, the “persist” group, who continued to search for illegal images despite multiple warnings. Data showed an increasing rate of warnings being shown in a session after about 10 warnings - some people were served more than 50 warnings in a session. This small group (0.15% of sessions shown the message and chatbot) resisted the warning messages, and for many of those who received more than a handful of warnings, it is clear that there is work to do to obtain the desired deterrent effect. From our work generally with people who offend we know that not everyone is ready to accept help when they are on a path to offending, but that readiness can change depending on various factors. This is one of the reasons that we are so keen for warnings to become the norm rather than the exception – repeated exposure to a message increases the likelihood that it will take root in the viewer’s mind and subsequently affect their behaviour. The Stages-of-Change model (20)

describes five stages that people experience when undergoing behavioural change: precontemplation, contemplation, preparation, action, and maintenance. These persistent users would be at the precontemplation stage; the stage at which there is no intention to change behaviour in the foreseeable future. An interesting future study could take place around this persist group, to see if, for example, there are alternative approaches that would increase engagement with them. This would require some research into understanding the characteristics and context of the group so that targeting of the right intensity and duration could be developed.

#### Chatbot sophistication

The chatbot itself was purposely limited in how it could respond to users to safeguard against inappropriate responses, but this in turn caused frustrations and irritation among some people who used the service. Some of these were users testing out the capacities of the chatbot, and some were users who clearly did not understand that what they had input in the search bar was illegal, leading to aggressive and defensive remarks.

It’s possible that amongst these discussions there were people who may have needed a more sophisticated interaction to encourage them to seek help, and for whom a more persuasive, intelligent chatbot might have pushed them to click through. This is something that will be considered as we look to develop chatbot 2.0.

### A selection of defensive user responses

Why I like young girls

This filtering is shit I didn’t search anything illegal

Can I please watch whatever I want on Pornhub

What did I do stupid chatbot

What the hell are you on about I don’t need help OK

I HAVE STOPPED DUMBASS I DID NOTHING WRONG

## UTAS honeypot experiment

UTAS have previously run their own experiment on the effectiveness of warning messages designed to deter people seeking illegal images of children. The honeypot experiment (21) involved running ads for a 'barely legal' website on a men's fitness site. (This site, GetFit, was designed specifically for the experiment, in order to covertly observe the behaviour of anonymous internet users.)

Clicking on the barely legal ads led the user to be confronted by one of seven different warning messages, each designed with a different emphasis. Some had photographs of an arrested person being led away, some emphasised police might be tracking the user, some offered a more therapeutic approach. A control group were shown no warning message, but were simply taken through to a 'routine maintenance' page).

The results demonstrated that showing any warning message had a deterrent effect, with the numbers of people choosing to click through to the site after receiving a warning message substantially lower than those in the control group who were simply taken to the front page of the site. But the results also showed that the inclusion of photographs, therapeutic text or the threat of police tracking could also make a statistically significant difference.

## Conclusions

Revisiting the questions posed by the evaluators, we have some answers.

- **Did the chatbot reduce searches for sexual images of children on Pornhub UK?**  
Yes, there was a statistically significant decrease over the 18-month project.
- **Was the chatbot effective in engaging 'at-risk' users who have entered search terms indicating they are looking for child sexual abuse images on Pornhub UK?**  
Yes, although the limitations of the chatbot's dialogue left some users frustrated or unwilling to engage
- **Did the warning message and chatbot increase the number of individuals seeking support from LFF?**  
Yes it did: the data shows that hundreds of users were directed to our services via their interactions with the chatbot and warning message.
- **Has the experiment enhanced the evidence base concerning online offending around sexual images of children?**  
Yes: in this study we have seen that just one warning can divert users in up to 82% of sessions from their pathway to offending in an adult pornography setting. Of course this is just one experiment; there is more to be done to understand the effects of warning messages in pornography contexts.

# The future: chatbot 2.0, Project Intercept and more

So, what of the future for the chatbot? The short answer is, we are excited about its potential. At the conclusion of the experiment it was agreed that LFF would take sole ownership of the chatbot as it sits comfortably within our other deterrence work. We are proud to welcome the chatbot into one of our newest programmes, Project Intercept.

## Project Intercept

Project Intercept is a three-year programme funded by Nominet that is attempting to challenge the current patchwork approach to warning messages online by working with tech companies to develop new and improved warning interventions. Through the project, we are also trying to find the gold standard of warning messages by experimenting with existing designs across a range of platforms. Tech giants including Google and Meta are already using warnings, as are gaming platforms (including MovieStarPlanet) and creator platforms (including Only Fans). The project is driven by the need to make warnings a commonplace and accepted intervention in any online space where people are searching for, viewing or sharing sexual images of children – to make a hostile internet for those attempting to cause harm. It also seeks to drive as many as possible to find help. This is an example of [situational crime prevention](#) (22), a theory used across LFF. This theory says that many people who offend will be deterred if we increase the effort that is needed to commit an offence. So by placing warnings widely we are increasing the friction in their journey and attempting to change their decision-making process. It increases perceived risk, increases the effort needed to offend, removes excuses and reduces permissibility (23).

As with all our work at LFF, it is important that everything is rooted in a solid evidence base to guide our actions. As such we are continuing to work with UTAS, who have created an internal

framework for our experiments with tech, as well as 20 hypotheses to test. These hypotheses are divided into different categories such as personalisation and emotion, consequences of detection, and external drivers, and will be chosen for experimentation according to the overall context of a particular tech partner (for example, a tech partner with a strong focus on privacy might be interested to see the impact of a message underlining the consequences of detection).

We also work with IPPPRI (International Policing and Public Protection Institute) to dive deeper into understanding the pathways and motivations of people who offend online. They have created a literature review (24) that synthesises all the current knowledge about pathways to online offending, and at the time of writing a deeper piece of research deriving insights from the dark web and encrypted spaces is about to begin. Both UTAS and IPPPRI are consortium board partners for Project Intercept, along with the IWF and a representative from a specialist law enforcement team. All of them bring years of expertise and guidance from their respective fields of work.

The opportunity to bring the chatbot into Project Intercept was one we jumped at. The chatbot experiment's results are a huge boost to our attempt to build a stronger evidence base for the use of warning messages. Now we can concentrate on improving the chatbot and rolling it out widely, in line with Project Intercept's aims to make the internet hostile to offending.

## Possible future directions

- Rolling out the chatbot to the full range of Aylo's UK adult entertainment websites, which include YouPorn and Redtube, and globally on Aylo sites – first English speaking and then other languages.
- Replicating the success seen on Pornhub UK on other companies' platforms.
- Adapting the chatbot to be used in non-pornography contexts. Searching on legal pornography sites is just one way that people attempt to access illegal images; with a chatbot engaging people at different intervention points across the online world we would increase awareness of the problem and drive more people to seek help.
- Updating the chatbot to be powered by a large language model that is fine-tuned and trained to deliver best practice interventions to make it a vastly more responsive and sophisticated technology, with less day-to-day maintenance from LFF staff needed.
- Experimenting with the design of the chatbot and the warning message to include, for example, LFF deterrence campaign videos to see if they are more engaging for the user, and perhaps experimenting with rotating a series of warnings to ensure there is less chance of people becoming habituated to the messages over time.
- Matching up the key times when chatbot was most active (10pm) with the opening hours of our helpline and live chat.

There is much work to be done to make the chatbot a more flexible and responsive (and therefore helpful) intervention. Throughout the period of the evaluation tweaks were constantly made by LFF staff to improve the dialogue between the chatbot and the user, but in the period since chatbot was launched, the advances made in generative AI have been huge. The possibility of relaunching chatbot 2.0 as a more naturalistic, responsive model that uses generative AI to engage and divert people is exciting. This would, of course, require a period of development and testing for safety.

“  
**The possibility of relaunching chatbot 2.0 as a more naturalistic, responsive model that uses generative AI to engage and divert people is exciting.**  
 ”

## Suggested reading

Scanlan J, Prichard J, Hall LC, Watters P, Wortley R (2024) 'reThink Chatbot Evaluation', University of Tasmania, Hobart. ISBN 978-1-922708-67-0:

[https://www.lucyfaithfull.org.uk/files/reThink\\_Chatbot\\_Evaluation\\_Report.pdf](https://www.lucyfaithfull.org.uk/files/reThink_Chatbot_Evaluation_Report.pdf)

A presentation by Joel Scanlan, part of the UTAS team, about the chatbot evaluation:

[https://joelscanlan.com/reThink\\_Eval\\_1080p.mp4](https://joelscanlan.com/reThink_Eval_1080p.mp4)

# Recommendations for the online tech sector

## Commitment

- The deterrence messages were a success and as such should be adopted in more contexts.
- All online and technology companies should commit to making the internet a hostile place for offending and implement child safety by design. They should also commit to signposting to help for people who are seeking it.
- Prevention is key, but early prevention is even better. Often, tech relies on actions such as identifying, blocking, reporting, removing, which happen after some harm has occurred. Earlier prevention, such as warning messages, wider education, clear expectations of user behaviour and linking to help in resource pages, can help to stop the harm before it happens.
- With the arrival of the Online Safety Act in the UK and the accompanying regulation, tech companies will increasingly be held accountable for the safety of their platforms. Now is the time for them to get ahead of the game and act as leaders in their sector, by working with us to create bespoke warning solutions for their spaces. Ofcom (in its draft illegal harms codes) has already recommended that large search engines should use warnings (with links to support services such as Stop It Now) as best practice. As regulation evolves there will only be more to come.

## Collaboration

- Collaboration is key to stopping online offending. The chatbot project has laid a blueprint for how tech companies can work collaboratively with the third sector to produce meaningful experiments and effective deterrence. Other tech companies should follow this example and share their data and insights to help those working on prevention

and deterrence to broaden our understanding of what works best to protect children online.

- Our research has shown that some people who are at risk of offending are worried about their online behaviour and want to change. They must have somewhere to go to support them in stopping and changing their behaviour – quickly and for good – before they harm a child. By signposting to Stop It Now, tech platforms can help with this goal.
- With a tech company's expertise on its users and platform, and LFF's expertise on the behaviour of people who have offended or are at risk of offending, a powerful intervention can be created that not only keeps bad actors off the platform but drives them to seek help too.
- Those companies who already run warnings and link to our helpline can help further by spreading the word and bringing on board tech who may be hesitant.

## Openness

- Some tech companies have been historically reticent about announcing their initiatives to deter those at risk of offending on their platforms because this is admitting that there is a problem. But offending happens everywhere – no platform is immune. Only by tackling the problem openly and with strong, targeted interventions, will the clear message be sent that this behaviour is not tolerated and will be stopped – but that there is support to stop too.
- In wider society we must make it acceptable and commonplace for people who have offended or are at risk of offending to reach out for help to stop. Tech has a key role to help spread this message and thus help bolster the culture of prevention.

# Appendices

## Appendix 1: Data provided by LFF, IWF and Aylo

### Aylo

#### Warning message/chatbot trigger data

July 2021- August 2023

- Daily total number of searches undertaken on Pornhub UK that triggered the warning message and chatbot
- Daily potential child-sexual-image-related searches as a percentage of all searches on Pornhub in the UK

#### Session data

May 2022-April 2023

- 7.7GB text data
- Date session started and ended (these can range from minutes to months, depending on when a user logs out/deletes the tracking cookie)
- List of text strings covering searches and videos watched

### LFF

#### Helpline data

March 2020-August 2023

- Call and live chat volume
- Notes on whether the contact mentioned chatbot

#### Web traffic data

April 2021-August 2023

- Visits to Stop It Now UK website and LFF global website

### IWF

#### Chatbot conversation data

- A sample that covered 132 discussions randomly sampled across the period
- A range of statistics relating to number of chat sessions and their length

## Appendix 2: Chatbot conversation transcripts

### Example one

User **WELCOME**

Hi, I'm a chatbot run by Stop It Now! We offer confidential, non-judgemental support. Click here for more information.

User **YES**

This page will set out all the services we have for you, they're all confidential and anonymous.

User **CALL**

The helpline is confidential and you don't have to give us your full name or any other details that will identify you. Our experienced advisors speak to and support thousands of people every year.

User **VERY HELPFUL**

Your chat session has now ended. Thank you and goodbye.

### Example two

User **WELCOME**

Just to let you know, we have enabled chat history for improving this service, do not offer any personal details.

User I accidentally searched child how do I get off this?

This page will set out all the services we have for you, they're all confidential and anonymous.

### Example three

User **WELCOME**

Just to let you know, we have enabled chat history for improving this service, do not offer any personal details.

User ho

I'm sorry, I didn't get that. Can you rephrase?

User hello

I'm a chatbot run by Stop It Now! We offer confidential, non-judgemental support. Click here for more information.

User How can I stop watching porn

Viewing sexual images of under-18s is illegal and watching these videos causes harm to children. Click on 'show services' for information on the services Stop It Now! has for people worried about their use of online pornography

User **SHOW SERVICES**

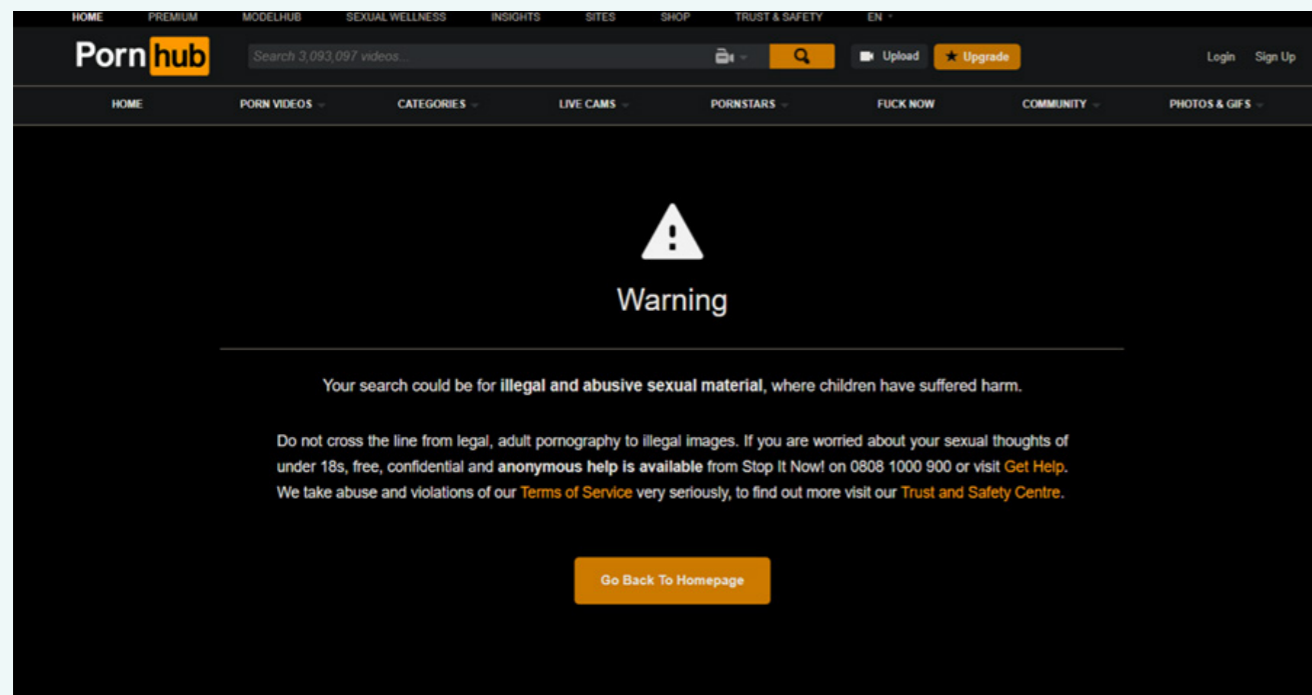
This page will set out all the services we have for you, they're all confidential and anonymous.

User **CHAT**

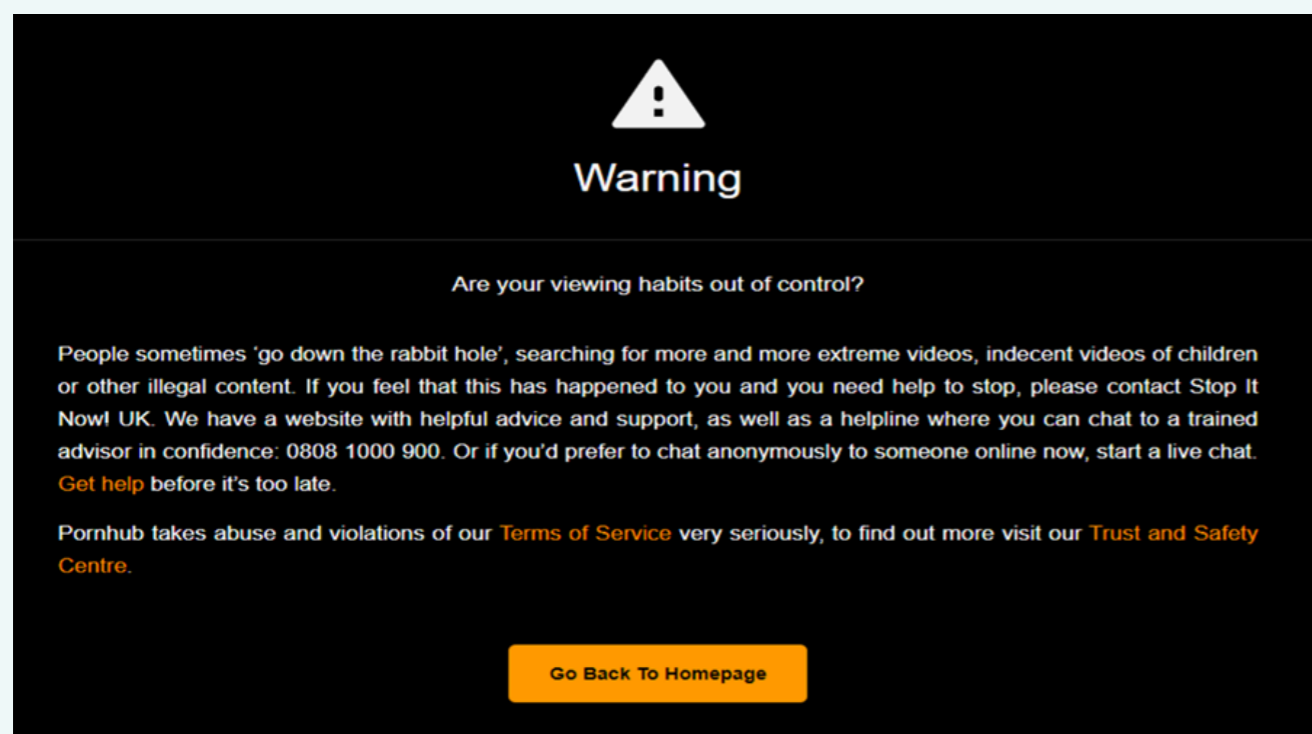
The live chat service is available Mondays: 9am - 12pm, Wednesdays: 6pm - 9pm and Fridays: 2pm - 5pm

## Appendix 3: Text of three different warning messages on Pornhub UK

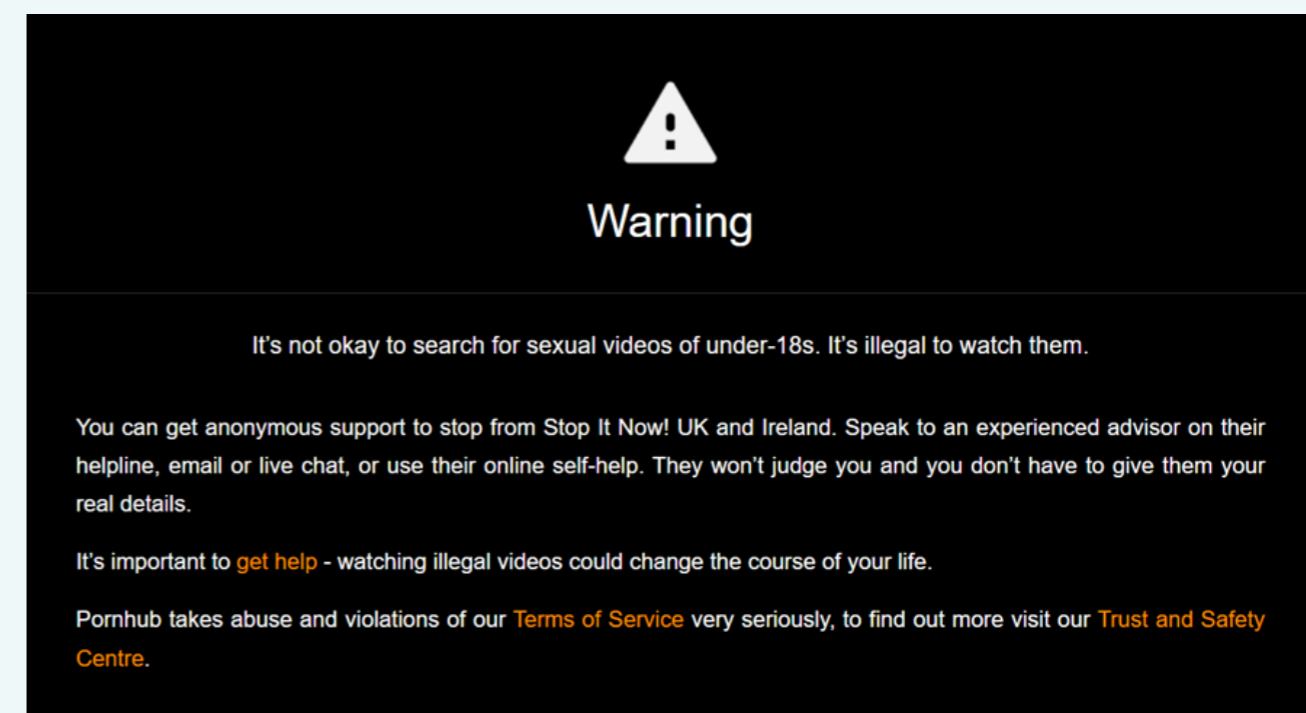
### 1. Original text: from February 2021 to May 2023



### 2. Test 1: from 11 May 2023 to 7 June 2023



### 3. Test 2: from 8 June 2023 to present day



# References

1. Internet Watch Foundation. (2023). IWF Annual Report 2023. <https://www.iwf.org.uk/annual-report-2023/trends-and-data/self-generated-child-sex-abuse/>
2. Davis, P. (17 June 2024). NCMEC on a Mission: President signs law expanding our reach [blog post]. <https://www.missingkids.org/blog/2024/ncmec-on-a-mission-president-signs-law-expanding-our-reach>
3. Internet Watch Foundation. (July 2024). What has changed in the AI CSAM landscape? <https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/>
4. National Crime Agency. (2024). National Strategic Assessment 2024 of Serious and Organised Crime. <https://www.nationalcrimeagency.gov.uk/threats/nsa-child-sexual-abuse-2024>
5. Walsh, M., Denis, D., & Findlater, D. (March 2023). Deterring online child sexual abuse and exploitation: lessons from seven years of campaigning. The Faithfull Papers. [https://www.lucyfaithfull.org.uk/files/LFF\\_Faithfull\\_Paper\\_Lessons\\_From\\_Campaigning\\_FINAL.pdf](https://www.lucyfaithfull.org.uk/files/LFF_Faithfull_Paper_Lessons_From_Campaigning_FINAL.pdf)
6. Pornhub. (June 2024). Community Guidelines. <https://help.pornhub.com/hc/en-us/articles/4419900587155-Community-Guidelines>
7. Prichard, J., Scanlan, J., Watters, P., Wortley, R., Hunn, C., & Garrett, E. (2022). Online messages to reduce users' engagement with child sexual abuse material: a review of relevant literature for the reThink Chatbot. University of Tasmania. [https://figshare.utas.edu.au/articles/report/Online\\_messages\\_to\\_reduce\\_users\\_engagement\\_with\\_child\\_sexual\\_abuse\\_material\\_a\\_review\\_of\\_relevant\\_literature\\_for\\_the\\_reThink\\_chatbot/23171162?file=40871957](https://figshare.utas.edu.au/articles/report/Online_messages_to_reduce_users_engagement_with_child_sexual_abuse_material_a_review_of_relevant_literature_for_the_reThink_chatbot/23171162?file=40871957)
8. Bailey, A., Allen, L., Stevens, E., Dervley, R., Findlater, D., & Wefers, S. (2022). Pathways and prevention for indecent images of children offending: A qualitative study. *Sexual Offending: Theory, Research, and Prevention*, 17, 1-24. <https://doi.org/10.5964/sotrap.6657>
9. Van Horn, J., Eisenberg, M., Nicholls, C. M., Mulder, J., Webster, S., Paskell, C., Brown, A., Stam, J., Kerr, J., & Jago, N. (2015). Stop It Now! A pilot study into the limits and benefits of a free helpline preventing child sexual abuse. *Journal of Child Sexual Abuse*, 24(8), 853-872. <https://doi.org/10.1080/10538712.2015.1088914>
10. Lucy Faithfull Foundation. (16 February 2024). A call to end AI-generated child sexual abuse [blog post]. <https://www.lucyfaithfull.org.uk/featured-news/a-call-to-end-ai-generated-child-sexual-abuse.htm>
11. Prichard, J., Spiranovic, C., Watters, P., & Lueg, C. (2013). Young people, child pornography, and subcultural norms on the Internet. *Journal of the American Society for Information Science and Technology*, 64(5), 992-1000. <https://doi.org/10.1002/asi.22816>
12. Pornhub. (2023). 2023 Transparency report (second half). <https://help.pornhub.com/hc/en-us/articles/27610703413267-2023-Transparency-Report-Second-H>
13. Bailey, A., & Moran, F. (May 2024). What's porn got to do with it? The link between viewing adult pornography and online sexual offending against children. The Faithfull Papers. [https://www.lucyfaithfull.org.uk/files/2024\\_05\\_Faithfull\\_Paper\\_Viewing\\_Pornography\\_Link%20\\_Final\\_Spreads.pdf](https://www.lucyfaithfull.org.uk/files/2024_05_Faithfull_Paper_Viewing_Pornography_Link%20_Final_Spreads.pdf)
14. Wilson, J. R., Kuehn, R. E., & Beach, F. A. (1963). Modification in the sexual behavior of male rats produced by changing the stimulus female. *Journal of Comparative and Physiological Psychology*, 56(3), 636-644. <https://doi.org/10.1037/h0042469>
15. Ovaska, et al. (in press). Study conducted by EU-funded project Concerning Link between Excessive Use of Adult Pornography and Child Sexual Abuse Material Offending. <https://www.uef.fi/en/article/viewing-of-child-sexual-abuse-material-is-often-preceded-by-habitual-consumption-of-adult>
16. Stop it Now & National Rapporteur on Trafficking in Human Beings and Sexual Violence Against Children. (2023, August). 'We have to talk about it': Research into young (potential) offenders of child sexual abuse [English translated summary report]. <https://offlimits.nl/nieuws/insight-into-young-offenders-of-online-child-sexual-abuse>
17. Merdian, H. L., Perkins, D. E., Dustagheer, E., & Glorney, E. (2020). Development of a case formulation model for individuals who have viewed, distributed, and/or shared child sexual exploitation material. *International Journal of Offender Therapy and Comparative Criminology*, 64(10-11), 1055-1073. <https://doi.org/10.1177/0306624X17748067>
18. Skakoon-Sparling, S., Cramer, K. M., & Shuper, P. A. (2016). The impact of sexual arousal on sexual risk-taking and decision-making in men and women. *Archives of Sexual Behavior*, 45(1), 33-42. <https://doi.org/10.1007/s10508-015-0589-y>
19. Davis, A. (23 February 2021). Preventing child exploitation on our apps. <https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/>
20. McConaughy, E., Prochaska, J., & Velicer, W. (1983). Stages of change in psychotherapy: Measurement and sample profiles. *Psychotherapy: Theory, Research & Practice*, 20(3), 368-375. <https://doi.org/10.1037/h0090198>
21. 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*, 155, Article 106955. <https://doi.org/10.1016/j.chiabu.2024.106955>
22. Allardyce, S., Denis, D., & Nelson, S. (January 2024). Creating safer environments for children and young people: situational prevention of child sexual abuse. The Faithfull Papers. [https://www.lucyfaithfull.org.uk/files/Faithfull\\_Paper\\_Situational\\_Prevention\\_Final.pdf](https://www.lucyfaithfull.org.uk/files/Faithfull_Paper_Situational_Prevention_Final.pdf)
23. Wortley, R., & Smallbone, S. W. (2012). Internet child pornography: Causes, investigation, and prevention. Praeger/ABC-CLIO.
24. Kumitz, L. (2024). Rapid review: pathways to CSAM offending [internal Lucy Faithfull Foundation report].

# Acknowledgements

We'd like to thank the Safe Online Initiative at End Violence Against Children for funding the work, the Internet Watch Foundation and Aylo for their collaboration, University of Tasmania for the original evaluation, the peer reviewers of this paper, and Giulia Terragni for writing the case study.



Visit our website to find out more about our work  
or to donate and help us keep children safe  
**lucyfaithfull.org.uk**

Lucy Faithfull Foundation Scotland:

**lucyfaithfull.org.uk/scotland**

Lucy Faithfull Foundation Wales:

**lucyfaithfull.org.uk/wales**

Find out about our Stop It Now helpline  
(0808 1000 900) and campaign:

**stopitnow.org.uk**

Our Shore website provides a safe space for teenagers  
worried about their own or a friend's sexual behaviour:

**shorespace.org.uk**



**Facebook:** [@LucyFaithfullFoundation](https://www.facebook.com/LucyFaithfullFoundation)



**Instagram:** [@lucyfaithfullfoundation](https://www.instagram.com/lucyfaithfullfoundation)



**X (Twitter):** [@Lucy\\_Faithfull](https://twitter.com/Lucy_Faithfull)



**YouTube:** [@Lucy\\_Faithfull\\_Foundation](https://www.youtube.com/Lucy_Faithfull_Foundation)



**LinkedIn:** [The-Lucy-Faithfull-Foundation](https://www.linkedin.com/company/The-Lucy-Faithfull-Foundation)

The Lucy Faithfull Foundation  
2 Birch House, Harris Business Park, Hanbury Road,  
Stoke Prior, Bromsgrove, B60 4DJ

Telephone 01527 591922

**LUCY  
FAITHFULL  
FOUNDATION**

The Lucy Faithfull Foundation is a registered Charity No. 1013025, and is a company limited by guarantee. Registered in England No. 2729957. Registered office: 2 Birch House, Harris Business Park, Hanbury Road, Stoke Prior, Bromsgrove, B60 4DJ.

Important note: All images within this document are for illustrative purposes only. The names of the individuals in our case studies have been changed in order to ensure their anonymity.