# Online messages to reduce users' engagement with child sexual abuse material

A REVIEW OF RELEVANT LITERATURE FOR THE RETHINK CHATBOT

Authored for The Lucy Faithfull Foundation by Jeremy Prichard, Joel Scanlan, Paul Watters, Richard Wortley, Charlotte Hunn and Eamon Garrett

Published by the University of Tasmania

ISBN: 978-1-922708-21-2

**Declaration**

This literature review was funded by The Lucy Faithfull Foundation. The authors declare that they have no interests, including financial interests, which might influence the approach they have taken to preparing for and writing this document

THE LUCY FAITHFULL FOUNDATION

Working to protect children

# Table of Contents

# Executive summary

## Introduction

- The aim of this literature review is to provide information that will assist The Lucy Faithfull Foundation and other stakeholders in the development of an interactive chatbot, *reThink*, designed to increase engagement with individuals attempting child sexual abuse material (CSAM) searches on PornHub UK and direct then to support services.

## The problem of CSAM

- The Internet has dramatically increased the problem of CSAM by providing easy opportunities to view CSAM
- The scale of the problem is such that it overwhelms efforts to police the Internet, and increased efforts need to be made in prevention.
- Individuals who access CSAM vary considerably in their involvement and level of technical expertise. The most striking characteristic offenders is their ordinariness not their deviance.
- One pathway into CSAM is thought to be via legal pornography. This suggests that prevention efforts might be particularly directed to those at the very start of a potential CSAM-offending career.
- Situational crime prevention provides a framework for developing prevention strategies aimed at deterring and disrupting searches for CSAM in legal pornography sites.

## Online warning messages

- There is a long history of using warning messages in the offline world to alert individuals to imminent hazards. Research has identified the key features of effective warning messages. They need to attract attention, be easy to understand, be believable, come from a credible source, and impart explicit information about how to avoid harm.
- Warning messages are increasingly being used online, where they can reach billions of people simultaneously, often at the moment that they are about to engage in the behaviour that is the focus of the message.
- Warning messages can be triggered when a user enters a specified keyword or attempts to access a specified URL and can be implemented by a range of individuals and actors within the technology, government, non-government and private sectors.
- Messages can appear on users' screens as pop-up messages or banner messages. The most effective messages are 'active', requiring the user to take some action to continue to use the website or application.
- Messages aim to change behaviour by alerting the user to the consequences of non-compliance, providing information to allow the user to come to a reasoned decision, and/or seeking to engage the user's emotions.

- Recent research by members of the current research team has demonstrated empirically that warning messages can deter individuals from attempting to access websites that purport to contain potentially sexually deviant material.

**Chatbots**

- Chatbots owe their origins to mid-20th century research which theorised how computers could 'think' and engage in conversation.

- Chatbots have many names and are used in all sorts of ways, such as in health, IT support, hospitality and tourism, education and finance.

- They have a common architecture involving: comprehending what a user has written or said; a decision-making process; and communicating a response to the user.

- The most sophisticated chatbots in use today have been developed by big tech companies. Two voice-operated examples are Alexa (Amazon) and Siri (Apple).

- Smaller off-the-shelf chatbots are becoming increasingly affordable.

- Internet users are becoming au fait with chatbots – and in some cases have a preference for chatbots over interacting with a person.

- Chatbots are currently being used to prevent crime and raise awareness.

- With regards to CSAM, it appears that chatbots are currently used to detect online grooming as well as Internet users who are searching for illegal images.

**Concluding observations for reThink chatbot**

- The *reThink* project may play an important role in addressing the *de facto* normalisation of CSAM in some otherwise legal parts of the Internet.

- Future research on CSAM-warnings needs to examine how: best to interact with different types of CSAM offenders; and compliance may vary over time.

- The broader literature on static warnings and chatbots consistently points to the importance of: specificity; trust, believability and credibility; careful use of language.

- Changing the appearance of warnings and chatbots can counter habituation.

# 1. Introduction

## 1.1 STATEMENT OF THE PROBLEM

The late Steve Jobs was one of the most influential figures in the rise of big-tech. A co-founder and CEO of Apple Inc., he became famous in his lifetime for accurately predicting how technological developments such as the Internet would influence human behaviour – including entertainment (Jobs, 1983). His 1985 interview with Playboy Magazine is still cited (e.g. Tweedie, 2014). That interview did not specifically discuss the likely growth of the pornography industry. Nonetheless, it is safe to assume that growth was anticipated by others because the pornography market had previously expanded with tech advancements, such as the VCR boom of the 1970s and 1980s (Coopersmith, 2006). Of course, the growth and normalisation of pornography since the 1980s has been astounding. Pornography is now a dominant form of entertainment. For example, Pornhub reported 30.3 billion visits in 2018 – an average of 92 million visits per day (Silver, 2018).

In his celebrated speech, 'The Future Isn't What It Used to Be', Jobs (1983) correctly predicted that through the Internet individuals who had eclectic interests would be less isolated because they could connect with likeminded people globally. However, neither Jobs nor any other public figure of the 1980s understood the implications of this connectivity for what the public may refer to as 'child pornography' (now, more appropriately called child sexual abuse material; CSAM). Arguably, of all possible illegal activities that might have escalated with connectivity, CSAM was the hardest to foresee.

In part this was due to the 20th century's erroneous belief that only psychologically 'unhinged' paedophiles, or individuals with serious mental illnesses, would seek CSAM for sexual entertainment. Historically this view seemed to be borne out by the small amount of CSAM in circulation. For instance, in 1980, CSAM was difficult to obtain when mail was the primary means of trafficking in the U.S. (Motivans & Kyckelhahn, 2007). The largest selling 'child pornography' magazine in the U.S. was estimated to sell around 800 copies in 1980 (Wortley & Smallbone, 2012). However, with the advent of the Internet and cheap digital cameras it became obvious that the potential market for CSAM was significantly larger than once thought. This is exemplified in the fact that by 2000, just twenty years on from the trickle of CSAM exchanged in 1980, a single Internet CSAM company was found to have more than 250,000 registered customers (Wortley & Smallbone, 2012). And more recent INTERPOL data indicated that a single website hosting CSAM received 6.5 million views in one month (WeProtect, 2019, p. 11).

CSAM has always been easy to find on the Internet and may be even accidentally encountered on mainstream websites and peer-to-peer networks (P2P) (e.g. Prichard et al., 2013). However, the existence of CSAM on some legal pornography websites (Morgan & Lambie, 2019) has only recently attracted public consternation (and resulted in civil litigation by some of the survivors of the child sexual abuse portrayed in the material).

This convergence of the 'black' market in CSAM with the 'white' market in legal pornography is crucial context for this literature review because it:

- Underscores that the tech-drivers of the boom in legal sex entertainment are interrelated with the continued expansion of the CSAM market

- Demonstrates the global significance of the *reThink* CSAM-prevention strategy that is being pursued by The Lucy Faithful Foundation (LFF) and the Internet Watch Foundation (IWF) in collaboration with Pornhub.

## 1.2 *RETHINK* CONTEXT AND BACKGROUND

MindGeek operates Pornhub, one of the world's largest legal pornography websites. Among other strategies to safeguard the legality of Pornhub,[1] in February 2021 MindGeek implemented a pop-up warning message on all its adult content sites across the globe, including for Pornhub's UK audience. The message was designed by LFF and appeared on users' screens if they entered a CSAM-related search term. The message informs users that:

- Any content associated with the search would probably constitute material that is illegal and where children have been harmed

- Confidential and anonymous help is available from Stop It Now! UK and Ireland for people who are concerned about their behaviours or thoughts towards children, including through their use of CSAM.

The message further advises people not to "cross the line" between legal adult pornography and illegal images of children. The core purposes of the warning are to reduce the prevalence of CSAM searches on PornHub UK and increase client engagement with LFF services.

In 2021-22 LFF will trial a new delivery system; the static online messages will be enhanced with an interactive chatbot, *reThink*. The objectives of the chatbot are to increase engagement with users, further reduce attempts to find CSAM, and increase numbers of referrals to confidential support services.

## 1.3 AIM AND SCOPE OF REVIEW

To assist LFF and other stakeholders, this document provides a scholarly review of literature that:

1. Explains how online messages can be viewed as a crime prevention strategy through the lens of crime science

2. Summarises evidence of the capacity of warnings to influence human decision-making, with a particular focus on online messages and chatbots

3. Describes the characteristics of messages and chatbots that appear to increase or decrease the likelihood of compliance, and

4. Discusses what the above might mean for the use of messages and chatbots in different Internet contexts (e.g. general search engines, gaming websites, adult websites).

[1] https://help.pornhub.com/hc/en-us/categories/360002934613

## 1.4 OUTLINE OF REVIEW

In addition to this introduction (**Section 1**), this review contains four further sections that address the four aims listed above (1.3).

**Section 2** discusses the nature of CSAM and types of offenders. Several points are made that serve as crucial foundation for the rest of the review. Chief among them is that not all CSAM offenders are strongly committed to viewing the material and, indeed, onset – the first deliberate viewing – can occur without a prior sexual interest in children. We draw on situational crime prevention (SCP) theory to explain the need for early intervention strategies to target those at risk of onset and those who have just commenced CSAM offending.

**Section 3** presents an overview of how messages are applied on the Internet and how they are deployed as preventative strategies. We begin by summarising research on the application of messages in the physical world, highlighting the key features of messages that are known to increase effectiveness as measured by their influence on human behaviour. We then examine how messages are applied on the Internet and how they are deployed as preventative strategies. The empirical studies carried out by the current research team are summarised. These demonstrate that pop-up messages can dissuade Internet users from viewing CSAM and sharing potentially illegal sexual images. Importantly, messages which offer users help for problems with their pornography use are just as effective as warnings about criminal law enforcement.

**Section 4** presents chatbots as an interactive form of online messaging. We examine what they are, how they draw on artificial intelligence (AI), what they can do when designed well, and how they have been deployed online in different Internet contexts.

***Section 5*** highlights the critical points raised in this review and offers some concluding observations on the design and operation of the reThink chatbot.

## 1.5 RESEARCH DETAILS

This literature review analysed international literature from multiple disciplines that have examined human responses to messages offline and online, including information technology and cybersecurity, human-computer interaction, occupational safety, psychology, crime science and criminology, health, and business and economics. Literature was sourced through academic databases, including: Web of Science, Ovid, ProQuest, Gartner, Informit, PsychINFO, AGIS, APAIS-Health, Emerald Insight, CINCH and CINCH-Health and Google Scholar. A short annotated bibliography is appended to this report providing summaries of key research studies cited.

# 2. The problem of child sexual abuse material (CSAM)

In this section we examine the growth of the CSAM problem in the Internet era. We argue that the Internet has provided unparalleled opportunities for individuals with even a passing interest in CSAM to satisfy their curiosities immediately, conveniently, cheaply, and at low risk. The situational prevention approach provides a framework for devising interventions to deter and disrupt the accessing of CSAM, particularly by those in the early stages of what may develop into a persistent pattern of CSAM use.

## 2.1 NATURE AND PREVALENCE OF CSAM

Criminal laws that define CSAM vary between countries (Broadhurst, 2019; Westlake 2020, p. 1127). The material ranges in severity from sexual posing and semi-nudity, through to egregious acts of sexual cruelty against minors, including infants (ECPAT, 2018; Internet Watch Foundation, 2017; Westlake, 2020 p. 1231). Like many other types of cybercrime, the scale of the CSAM market is difficult to quantify precisely (Westlake 2020, p. 1230). Nonetheless all available data indicate that the market's growth continues unabated (ECPAT, 2018; EUROPOL, 2020, P. 36; Internet Watch Foundation, 2017, p. 50) and actually appears to have spiked upwards during the COVID-19 pandemic (INTERPOL, 2020, p. 9).

The sheer scale of the problem now undermines any suggestion that policy makers can rely on the criminal justice system to provide a solution (Quayle & Koukopoulos, 2019). In addition to overwhelming policing resources (Holt et al 2020; see also, Broadhurst, 2019; Carr, 2017; Westlake, Bouchard and Girodat, 2017, p. 290), the prosecution of CSAM offenders constitutes a continuous drain on the court and prison systems (Home Affairs, 2018; Motivans & Kyckelhahn, 2007).

## 2.2 WHERE CSAM CAN BE FOUND ON THE INTERNET

It is well known that the Dark Web (e.g. as accessed via a TOR browser) provides efficient mechanisms for accessing and sharing CSAM (Europol, 2021). On the Surface Web open to the general public CSAM links and dedicated CSAM websites can be found using ordinary search engines (Carr 2004; Price, 2020; Steel, 2015; Westlake 2020; Westlake & Bouchard 2015; Westlake, Bouchard & Girodat, 2017, p. 289). (We note that Google has implemented a sophisticated system to prevent CSAM being indexed through their platform and to reduce the chances that it is shown to the user (Google, 2020).)

Search engines aside, CSAM can be accidentally encountered or deliberately sought in all sorts of contexts – legal pornography websites (Fortin, Paquette, & Dupont, 2018, p. 35; Morgan & Lambie, 2019; Ray, Kimonis, & Seto, 2014), P2P networks (Prichard et al., 2011, 2013; Wolak, Liberatore & Levine, 2014), email spam (Krone, 2004), website notice boards (Rushkoff, 2009) and so on.

## 2.3 TYPES OF CSAM OFFENDERS

Deviant pornography and CSAM is a risk factor for sexual assault by individuals already predisposed to sexual aggression (Malamuth, 2018; see also Quayle, 2020). So while it is feasible that viewing CSAM may increase the likelihood that some individuals will go on to sexually abuse children, no causal link has been found to exist (Babchishin, Hanson, & VanZuylen, 2015; Malamuth, 2018).

Studies of offenders who use CSAM (but do not sexually assault children) indicate that, although the pathways leading into CSAM offending vary considerably (Merdian, Perkins, Dustagheer & Glorney, 2018), three main categories of offenders are discernible (Seto & Ahmed, 2014).

1. The first includes individuals whose offending is consistent with a diagnosis of paedophilia.

2. In the second category are those for whom CSAM viewing forms part of a broader range of behaviours that meet criteria for a hypersexual disorder.

3. The third and simplest category includes individuals who start viewing CSAM as the result of impulsive risk-taking behaviour.

The idea that an otherwise 'normal' person might deliberately view CSAM is difficult for many members of the public to accept or understand (Hunn et al., 2021, 2020). However, it is borne out consistently by other studies (e.g. Beech et al., 2008, p. 225; see also Lanning, 2010), including those which have found that onset may commence in the absence of a paedophilic disorder or a self-reported sexual interest in children (e.g. Ly et al., 2018). In fact, CSAM offenders are remarkably heterogeneous in terms of education, employment, and family background (Wolak, Finkelhor, & Mitchell, 2011). Their striking characteristic is their "ordinariness, not [their] deviance" (Wortley, 2012: 193).

## 2.4 THE PATHWAYS MODEL

What causes an individual to possess an interest in CSAM is a matter of academic debate but also of real-world importance. Many researchers have identified a common pattern whereby individuals may progress from legal pornography – and particularly the so-called 'barely legal' genre – to CSAM (Merdian et al., 2018; Morgan & Lambie, 2019; Perkins & Wefers, 2019). Developing a specific interest in CSAM may be gradual and eventually involve crossing a "significant psychological threshold" (Wortley & Smallbone, 2012, p.121).

Showing that legal pornography use predates accessing CSAM does not establish a causal link. However, there is evidence that viewing barely legal pornography can have a corrosive effect and increase the tendency to view children as sexual objects (Paul & Linz, 2008). Research suggests that CSAM use can become progressively more problematic over time for an individual through repeated viewing, habituation and escalation (e.g., Fortin & Proulx, 2018; Paul & Linz, 2008; Quayle & Taylor, 2004). This pathways model of offending highlights the importance of deterring individuals in the early stages of CSAM offending.

## 2.5 SITUATIONAL CRIME PREVENTION AS A FRAMEWORK

The default approach to understanding and preventing crime has long been to identify and attempt to correct the presumed personal deficits of those who commit crime. Situational crime prevention (SCP), on the other hand, suggests that all criminal decision-making occurs as the result of an interaction between **personal** and **situational** factors. An alternative approach to prevention involves changing immediate environments in ways that make crime less likely to occur. SCP may be characterised as a shift from an *offender* focus – understanding why some individuals develop criminal propensities – to an *offending* focus – understanding why an individual commits a crime at a given time and place. SCP does not deny the importance of personal factors, but suggests that given the right circumstances even normally law-abiding individuals may offend (Clarke, 2017; Mayhew et al., 1975).

From the perspective of SCP, the extraordinary expansion of the CSAM market cannot be plausibly explained by an increase in personal factors – such as paedophilia – at the population level. The simplest and most convincing explanation is the Internet era has ushered in a global change in situational factors (Smallbone & Wortley, 2017; Wortley, 2012; Wortley & Smallbone, 2006, 2012). The changes to which SCP refers are readily apparent. Today CSAM can be immediately accessed cheaply, with apparent anonymity and with low risk of detection, from the comfort of home (Merdian et al., 2009; Quayle, 2012; Wortley & Smallbone, 2006, 2012). As Quayle (2012: 110) observes, there are few other crimes that an individual can commit with such "extraordinary ease".

This is not to deny the agency of individuals. Technology has not stripped Internet users of the capacity to decide *not* to view CSAM. We would suggest that the majority of society has no interest in CSAM regardless of the circumstances. It should also be noted that a multitude of other factors may combine to increase the risk of onset in any one individual case. For example, it seems likely that an individual might be at greater risk on an evening where they are alone (e.g. Seto, 2019), already in a sexually aroused state (Taylor & Quayle, 2008) and intoxicated.

SCP is not only useful because of its compelling explanation for why CSAM offending has increased. It is also widely recognised as providing a valuable framework for mapping out prevention and early intervention strategies (Quayle & Koukopoulos, 2019). In short, SCP prevention strategies involve making the Internet a less conducive environment for CSAM offending, for example, by increasing the perceived risk of offending (Smallbone & Wortley, 2017; Wortley, 2012; Wortley & Smallbone, 2006b, 2012). Warning messages are one example of a situational strategy because they attempt to influence the decision-making of potential offenders at the very time they are seeking out CSAM. Messages can increase users' perception of the legal risks of CSAM offending. But it is also feasible messages can counter cognitive distortions or excuses that users may be entertaining at a point in time (Prichard et al., 2021).

# 3. Online warning messages

Warning messages of one kind or another are a ubiquitous feature of modern life – almost to the point where we are not aware of their influence on our behaviour. In the course of a single morning we might follow cautions printed on the labels of pharmaceutical products or foodstuffs, obey dozens of traffic signs while driving to our workplace, adhere to various posters reminding us of COVID-19 regulations, participate in a fire drill prompted by a siren, and, more recently in the online world, accept a computer-generated invitation to upgrade some software on our laptop.

We begin this section by examining the bank of empirical knowledge that has been developed over the past 50 years about the conditions in which humans are likely to comply with warning signs in the real offline world. We then move to the online environment, looking at the different ways that online messages can be delivered, before the reviewing research on the effectiveness and underlying psychology of warnings. We conclude by looking at the application of warning messages to the problem of online CSAM.

## 3.1 WARNING MESSAGES IN THE OFFLINE WORLD

In the real offline world, warning signs – in the form of symbols, words, flashing lights, sirens and so on – have been used for decades to communicate the existence of hazards relevant to public health (e.g. Hall et al., 2021; Mollen et al., 2017; Rosenblatt et al., 2018), occupational safety (Taylor & Wogalter, 2019), road safety, consumer protection, ergonomics and so on (e.g. see Wogalter & Mayhorn, 2006; Wogalter, 2019). A smaller body of work has examined the efficacy of warnings as an offline crime prevention strategy. But several studies have empirically demonstrated that messages can work to varying degrees to prevent crime. For instance, experimental research has demonstrated that warnings can reduce home burglaries (Chainey, 2021; Kyvsgaard & Sorensen, 2021) and the theft of black goods on university campuses (Chernoff, 2021). Similarly postal letters have been shown to be potentially useful to reduce insurance fraud (Blais & Bacher, 2007) and tax evasion (Coleman, 2007) and to protect victims of online fraud (Cross, 2016).

Generally, warning signs are considered to be legitimate when used as a last resort to mitigate risks. In other words warnings are an appropriate strategy only where hazards cannot be completely (a) removed or (b) neutralized with guarding mechanisms (Cacha, 2021; Wogalter, 2020).

**Individual factors** heavily influence whether a person will comply with a warning message. These include: cognitive processes involved in memory and comprehension; beliefs and attitudes, such as personal perceptions of risk; and motivation, which can be influenced by modeling (e.g. witnessing the behaviour of others), whether an individual is pressed for time, and the amount of effort required to comply (see Wogalter, 2019).

These factors are consistent with behaviour we have all observed in everyday life. For example, a confident 20-year-old may not be bothered to read warning labels on a lawnmower on the basis that he can generally remember how to operate it and perceives it as a low-risk machine. But the same labels may be routinely observed by his 50-year-old father. Interestingly, perceptions of risk appear to be affected more by the severity of the hazard – the harms that might result – than the likelihood of injury (Lenorovitz, Karnes & Leonard, 2014). So, the same hypothetical 20-year-old may be more inclined to heed the warning labels on a chainsaw because he understands that an accident with that

machine could be far graver than an accident with a lawnmower, irrespective of the relative probabilities of injury.

However, **design factors** also matter. Scientific studies have examined the features of messages that increase the likelihood of compliance. Most of the findings strongly appeal to common sense. For example, not surprisingly, signs that are not easily noticed, are too wordy or are vague are unlikely to be effective. To express this positively, for warnings to influence individual's decision-making they need to attract attention and efficiently impart explicit information about a specific hazard and the behavior necessary to avoid harm (Lenorovitz, Leonard & Karnes, 2012).

Empirical evidence also shows that compliance is more likely when messages are: believable (Riley et. al., 2006); from a credible source (Selejan et al., 2016; Wathen & Burkell, 2002; Wogalter & Mayhorn, 2008); clear and concise (Laughery & Paige-Smith, 2006); and when they match the degree of danger to specific colors, alert symbols (Ng & Chan, 2009; Zielinska, Mayhorn, & Wogalter, 2017) and signal words, such as "caution" or "warning" (ANSI, 2016; see also Kim & Wogalter, 2015).

| **Messages are more likely to influence decision-making when they:** |
|---|
| Attract attention |
| Are clear and concise |
| Impart explicit information about specific hazards, potential harms, and what to do to avoid harm |
| Are believable |
| Come from a credible source |
| Match the degree of danger to specific colours |
| Match the degree of danger to alert symbols like "!" |
| Match the degree of danger to signal words like "caution" or "warning" |

**Table 1**: Features of message-design that influence human decision-making

However, lists like the one contained in Table 1 cannot be applied as a generic 'checklist'. On the contrary, it is vital to design messages for the context of different hazards. By way of example, to minimise the chances of injury, very different warnings are needed for recreational and domestic vehicles (e.g. jet skis, ride-on mowers, and utility terrain vehicles) (Lenorovitz, Karnes & Leonard, 2014).

Finally, providing a warning in multiple modalities or formats tends to be better than relying on a single form (see Wogalter, 2019: 36). For instance, to evacuate a building because of the threat of fire it seems that people will respond best to a combination of physical signs (in the form of pictures and text) and auditory warnings (e.g. sirens) (Taylor & Wogalter, 2019).

## 3.2 REGULATORY SYSTEMS FOR MESSAGE DESIGN

In some countries regulatory agencies control the standards of warning signs and labels for use in the *offline world*. Examples include:

- The UK Health and Safety Executive[2]

[2] https://www.hse.gov.uk/workplacetransport/safetysigns/index.htm

- Product Safety Australia, Australian Competition and Consumer Commission,[3] and
- The American National Standard for Product Safety Signs and Labels (ANZI, 2016).[4]

Among other things, these agencies ensure that warning signs comply with the scientific evidence listed in Table 1. However, an analogous level of regulation for *online warnings* is difficult to find. In fact, we have encountered examples of online warnings that are used by government agencies that fail to comply with best-practise in offline message design (see Box 1).

**Box 1: Official online warnings that may not comply with message-design literature**

The U.S. National Institute of Standards and Technology (NIST) implemented an online warning to deter computer hacking (see Testa et al., 2017: 700). The warning stated:

> *The actual or attempted unauthorized access, use, or modification of this system is strictly prohibited. Unauthorized users are subject to institutional disciplinary proceedings and/or criminal and civil penalties under state, federal, or other applicable domestic and foreign laws. The use of this system is monitored and recorded for administrative and security reasons. Anyone accessing this system expressly consents to such monitoring and is advised that if monitoring reveals possible evidence of criminal activity, the Institution may provide the evidence of such activity to law enforcement officials.*

With reference to the features of messages recommended in Table 1 (above), in our view the main deficiency of this warning was its lack of clarity or conciseness.

Excessive length – over 170 words – was also an obvious departure from the message design literature in a CSAM warning that was trialled in Norway by the Child Sexual Abuse Anti-Distribution Filter (see further Wortley & Smallbone, 2012: 119).

## 3.3 MOVING TO THE ONLINE ENVIRONMENT

Warning messages are now a common part of the online experience. A key feature of online messaging is the ability to transmit a message to literally billions of people simultaneously, often at the moment that they are about to engage in the behaviour that is the focus of the message.

There are four broad ways that an online message may be delivered:

1) Passively, as simple advisory text (Figure 1);
2) As a modal dialog box requiring interaction through acknowledgment (Figure 2);
3) Requiring a decision to be made, based on the advice given (Figure 3), or;
4) Simply as a list of facts implicitly requiring interpretation and action (Figure 4).

[3] https://www.productsafety.gov.au/product-safety-laws/safety-standards-bans/mandatory-standards
[4] https://www.nema.org/standards/view/american-national-standard-for-product-safety-signs-and-labels

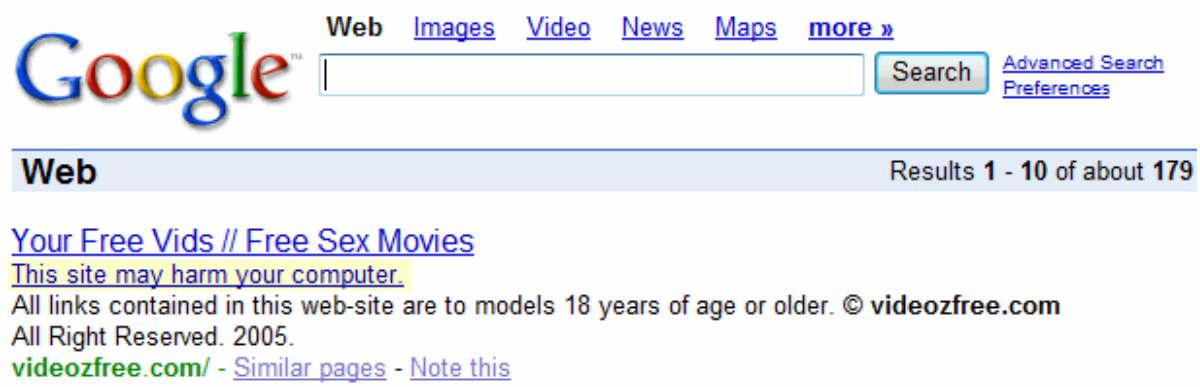



**Figure 1** – Passive Advisory Warning Message

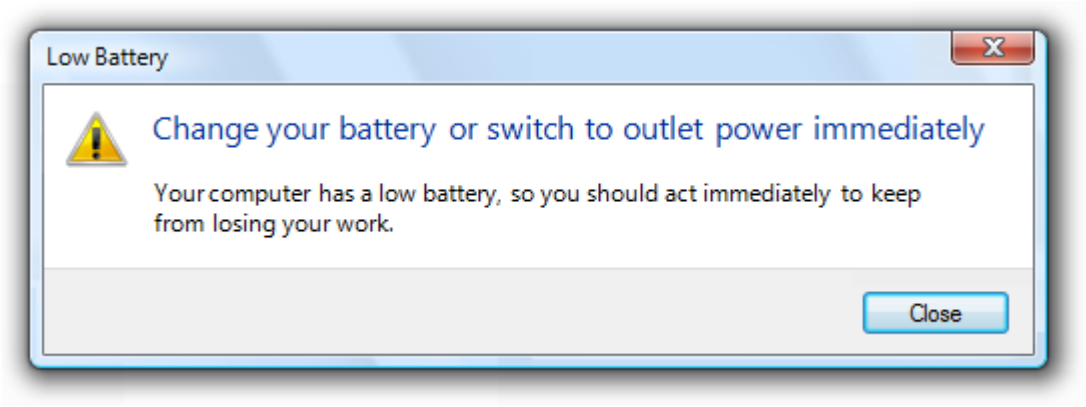



**Figure 2** – Modal Dialog Box Warning Message

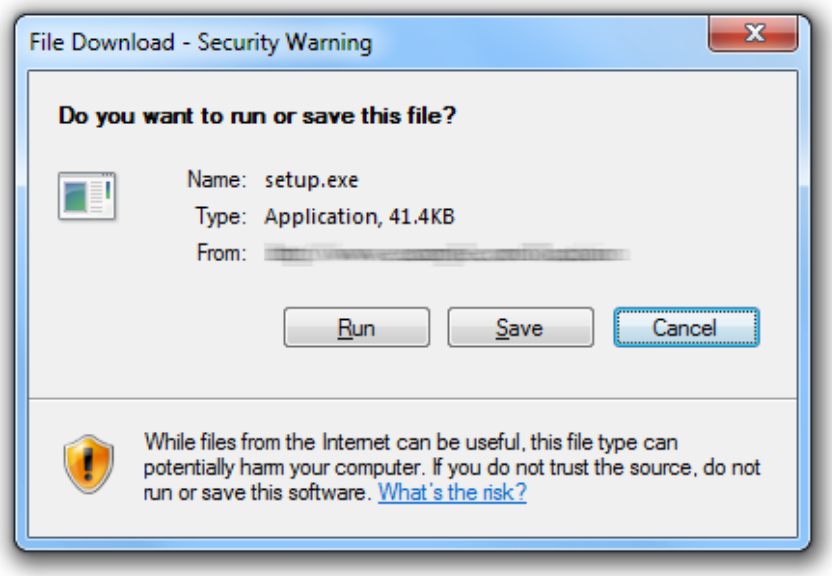



**Figure 3** – Warning Message Requiring A Decision


**Figure 4** – Factual Warning

Online messages can be delivered using a range of aesthetic approaches, with explanatory text and visual and auditory cues provided to draw users' attention and the opportunity to use animation. Messages can appear on users' screens in a number of ways, including pop-up messages and banner messages. They can either be passive or active, even to the extent that users must interact with the message in some way to be able to continue to use the website or application that generated the message.

## 3.4 HOW ARE ONLINE MESSAGES DEPLOYED?

Warning messages can be triggered when an Internet user 1) enters a specified keyword as a search query into a search engine, or 2) attempts to access a specified URL. Warnings could be presented in browsers (Reeder et al., 2018), email programs (Petelka et al., 2019), third-party apps (Reuter et al., 2017), or virtual reality (Tomczyk et al., 2021). A recent review by Hunn et al. (under review) outlines the range of individuals and actors within the technology, government, non-government and private sectors who can take action to implement warning messages. They are:

- The account holder: This can include a private individual or an institutional, corporate or other business account-holder that allows others to use that account.
- Operating system (OS) developers: The three most common operating system are Microsoft Windows, macOS, and Linux.
- Browser developers and vendors: Browser developers and vendors create the software, or interface, between a user's computer and the internet.
- Internet search engine (ISE) companies: Internet search engines, including, Google, Bing, and Baidu, to name but a few, are software systems through which a user can systematically search the internet using a text-based search query (i.e., a keyword).
- Internet service providers (ISPs): Commercial internet service providers, such as BT and Virgin Media (among others), sell internet connections and services to private individuals and organisations, including institutions and corporations.

- Virtual private network vendors VPE): VPN services enable a user to send and receive data within, typically, an encrypted private network using the shared or public internet network.

- Domain name Service (DNS) providers: DNS providers translate a human readable domain address into an IP address, that is, into a numerical identifier.

- Tor software: The TOR network provides its own browser, although multiple browsers support access to the TOR system.

- Third parties. Third parties include government departments, statutory bodies (e.g., eSafety Commissioner), agencies (e.g., LEAs), non-government organisations (NGOs) and for-profit companies who operate in the child protection sector.

## 3.5 RESEARCHING ONLINE WARNING MESSAGES

Methods used to measure the effectiveness of online messages include self-report surveys (Ullman, 2017; Zaikina-Montgomery, 2011), uncontrolled field experiments (Silic & Back, 2017), controlled laboratory experiments (Anderson et al, 2016), controlled field experiments (Junger et al, 2017), meta-analyses (Hancock et al, 2020), crowdsourced recruitment (Clayton et al, 2020), surveys (Reeder et al, 2018; Chen & Li, 2017), and randomised controlled experiments using 'honeypots' and naively recruited participants (Maimon et al., 2014; Prichard et al, 2021).

Most studies of online warnings that were reviewed for this report found evidence that behaviour could be modified by the use of a warning, across a range of desk reviews, observational and experimental studies, using both qualitative and quantitative data. Research indicates that Internet users are prepared to heed warnings about hazardous online behaviours related to: perpetrating cyber-attacks (Testa et al., 2017); malicious software use (Silic & Black, 2017; Anderson et al., 2016); exposure to malware (Haddad et al., 2020); online piracy (Ullman & Silver, 2018); fake news (Clayton et al, 2020); phishing (Neupane et al., 2016); weather and civil defence or emergencies (Casteel, 2016); online gambling (Caillon et al., 2021; Gainsbury et al., 2015; Ginley et al., 2017); pro-anorexia websites (Martijn et al., 2009); and disclosing personal information (Carpenter et al., 2014).

## 3.6 PSYCHOLOGICAL MECHANISMS BEHIND WARNING MESSAGES

Warning messages may operate via three different psychological domains, namely behavioural, cognitive and affective. Through a ***behavioural*** lens, an online message might encourage a certain type of behaviour by providing a reward, or reduce behaviour by delivering a punishment. From a ***cognitive*** perspective, an online message might present a series of facts, appealing to a user's intellect and capacity to process information. From the ***affective*** perspective the message may seek to engage the users' emotions, such an anxiety, for example, by scaring a user from engaging in a future behaviour. In practice these three domains overlap and warning messages may rely on elements from each. For example, a warning that a particular action might result in legal action is behaviourally based in that it relies on the threat of punishment to change behaviour. But the prosect of punishment can induce anxiety, while the judgement as to the likelihood that the punishment will actually occur involves rational-choice decision-making.

Across these domains, research indicates that the more salient the message is made to the user the more effective it is. For example, Silic and Black (2017) used warning messages to try and dissuade

users from installing malicious software. They compared the effectiveness of 1) a low impact message in which malicious software was said to be against security policies, 2) a medium impact messages, where malicious software was said to be illegal, and usage was monitored, and 3) a high impact message where malicious software was potentially dangerous and harmful to the user. The messages resulted in malicious software being installed less often, and with shorter duration, and less repeated use, with the levels of desistence 34% for low, 44% for medium, 63% for high, and 10% for the control group (which received no warning).

A major concern with online warning messages is ***habituation*** – that is, that over time weary users ignore the message and it loses its effectiveness. This is a risk for messages both in the online (e.g. Amran, Zaaba & Mahinderjit Singh, 2018) and offline context (Kim & Wogalter, 2009; for a recent example in the crime prevention context see Kyvsgaard & Sorensen, 2021).

However, habituation is not inevitable. Recent findings based on >6,000 surveys with Internet users in situ – that is, immediately after the participants viewed real-life online warnings – suggest that "rather than acting out of habit, participants made decisions based on the specific circumstances and the context of each warning" (Reeder et al., 2018: 9).

Additionally, research indicates that habituation effects can be countered by changing the appearance of warning signs (Amran, Zaaba & Mahinderjit Singh, 2018; Anderson et al., 2016; Kim & Wogalter, 2009). Online warnings that constantly change appearance are called ***polymorphic***. We should note here that, compared with messaging in the offline world, changing the appearance of online warnings will usually be a simple and efficient process. In other words, the capacity to deploy polymorphic messaging is a major advantage of the Internet environment.

Other strategies may assist too. Habituation may be reduced by using ***active rather than passive*** messages (Egelam, Cranor & Hong, 2008). For example, Kaiser et al. (2021) found that interstitial warnings were the most effective type of active warnings; that is, users could not proceed until they were forced to interact with the active message, such as a link being disabled until a warning was dismissed. Other active message approaches include moving warnings closer to user activity on the screen, or presenting the warning when a mouse "hovers over" a link (Petelka et al., 2019).

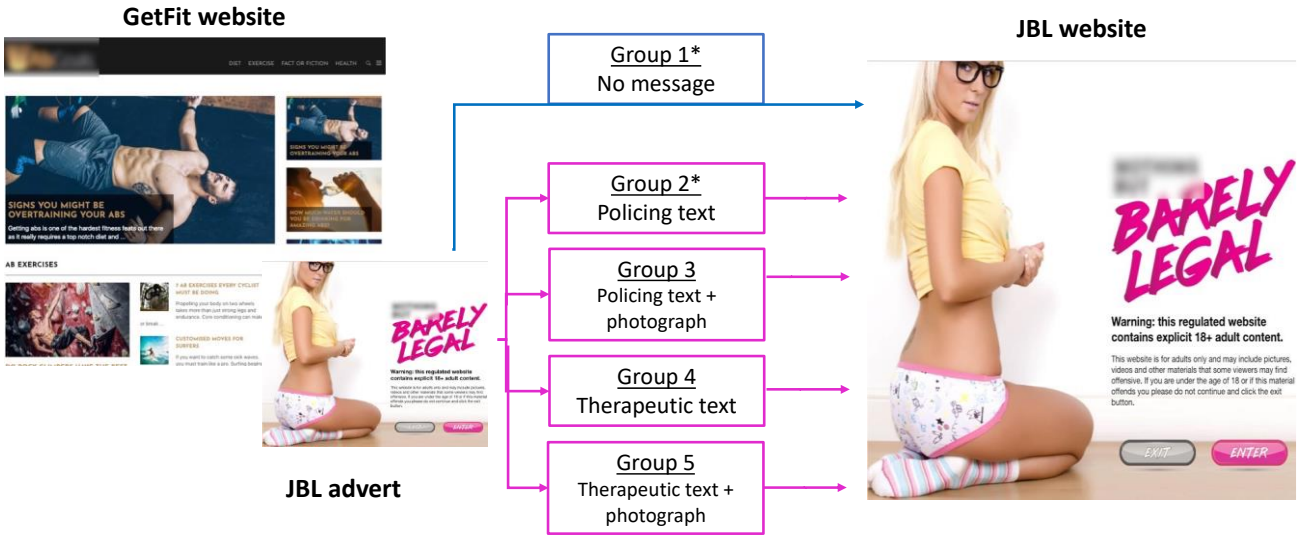
## 3.7 CSAM WARNING MESSAGES

Online messages to prevent CSAM have been trialed by law enforcement agencies (e.g. Global Alliance Against Child Sexual Abuse Online, 2014) and are used by some Internet companies (e.g. Essers, 2013; Google, 2020). However, until recently there has been little research to demonstrate their effectiveness. To redress this gap, the authors have an ongoing research programme designed to establish the effectiveness of online warning messages and to identify the salient features that contribute to effectiveness.

We subcontracted a commercial agency to design a men's fitness website targeting young adult men, which we refer to as 'GetFit'. Like other 'honeypot' studies (e.g. Testa et al., 2017), we developed this website to covertly observe the behaviour of anonymous Internet users. The primary aim here was to achieve high ecological validity: to examine how users would interact with online warnings about pornography in real life. Our rationale was that the standard approach to message research – through interviews and surveys – would be undermined because participants' might be influenced by embarrassment and similar social factors.

We were granted approval by a registered human research ethics committee on the basis that: our research was for the social benefit (beneficence); there was a low risk of causing distress or harm to participants or others (non-maleficence); we implemented adequate measures to protect the anonymity of participants; no illegal behaviour was observed; and no pornography was shown (Prichard et al., 2021). To date, we have conducted two series of double-blind randomised controlled experiments.

**Experiment 1** examined whether warnings could dissuade users from **viewing** the 'barely legal' genre of pornography, which eroticises adult-minor sex (Peters et al., 2014). For legal and ethical reasons we used this genre of pornography as a proxy for CSAM (Prichard et al., 2021). Our commercial partner designed an advertisement for a (non-existent) website called *Just Barely Legal* (JBL). This was displayed on *GetFit* among real advertisements from unassociated companies. See Figure 5, below.




**See Prichard et al. (2021)*

**Figure 5** – Structure of viewing experiments with 'barely legal' pornography genre

Users who clicked on the JBL advertisement were randomly allocated to a ***control group***, which received no message and landed directly at the JBL website. There they could click 'enter' or navigate away. If they clicked 'enter' they received an error message about routine maintenance.

The ***experimental groups*** received one of seven different types of warning messages. For simplicity, in this document we are presenting four messages which are most relevant to the current review.[5] The messages were designed in accordance with the literature presented in Table 1 (above) and pretended to come from the administrators of the GetFit website. Group 2 stated that "Police may get IP addresses to track users" (Policing text). Group 3 combined the same message with an image of an arrest (Policing text+image). Group 4 asked users "Concerned about your porn use? Visit mensline.org.au for support" (Referral text). Group 5 combined the message with an image of contemplation (Referral text+image) (see Figure 6, below).

[5] The results for Group 1 and three other conditions not described here can be found in Prichard et al (2021). The results for groups 3, 4 and 5 await publication. This document is not intended to replace the peer-review process.

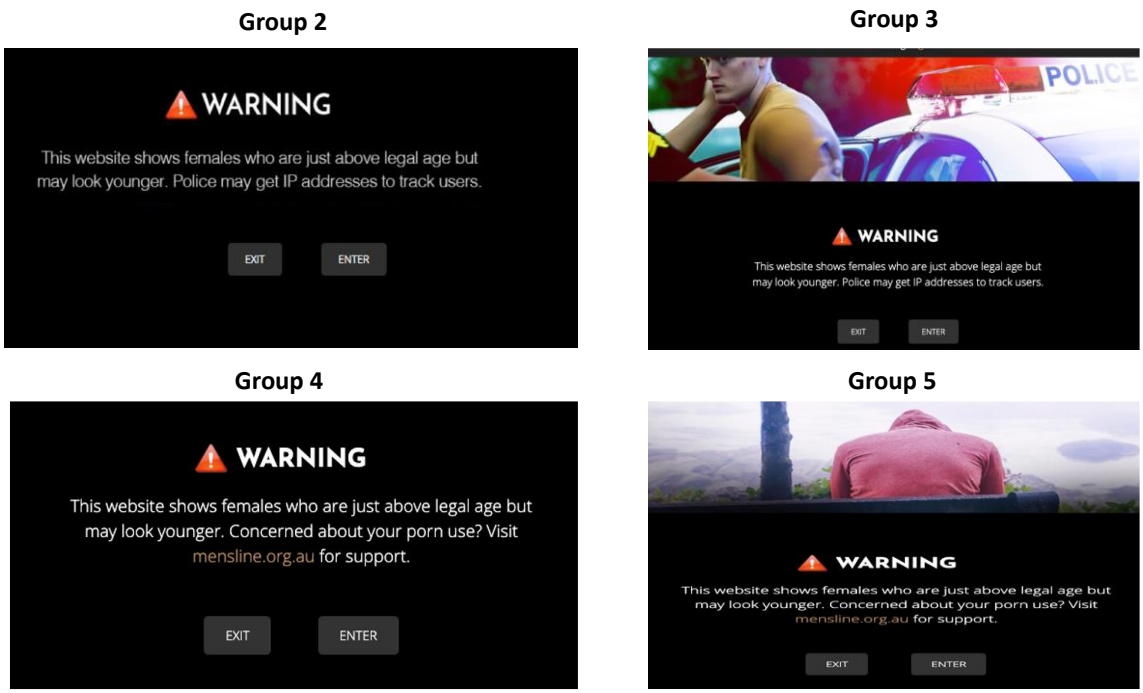



**Figure 6** – Messages used in 'viewing' experiments

We recorded whether participants click 'enter' on the JBL website ('persistence'). The per centage of participants in each group that attempted to enter were as follows:

- Group 1 No message (N=100) – 73%
- Group 2 Policing text (N=81) – 51%
- Group 3 Policing text+image (N=117) – 35%
- Group 4 Referral text (N=120) – 40%
- Group 5 Referral text+image (N=137) – 47%.

Chi square tests were conducted between Group 1 and each of the experimental groups. These tests revealed that the differences were statistically significant and meaningful. All messages dissuaded users from attempting to enter the JBL website. Consistent with SCP, we predicted that the deterrent-focussed policing messages (Group 2 and 3) would reduce persistence (see Prichard et al., 2021).

We were surprised by the effectiveness of the referral messages and we have not yet concluded how best to interpret these findings. What is clear, however, is that the effect of including an image with the message is dependent upon the message in question; it appears to have value-added for the police message but not for the referral message. This finding underscores the point made in 3.1 that context matters. Overall, these unpublished results appear highly salient to the reThink project, given that it offers users routes to therapeutic interventions.

**Experiment 2** (Prichard et al., in press), examined whether warning messages would dissuade adult men (N=528) from ***sharing*** a sexual image of their girlfriend. We used this context as a proxy for the sharing of CSAM. The message warned about the illegality of uploading images of underaged girls

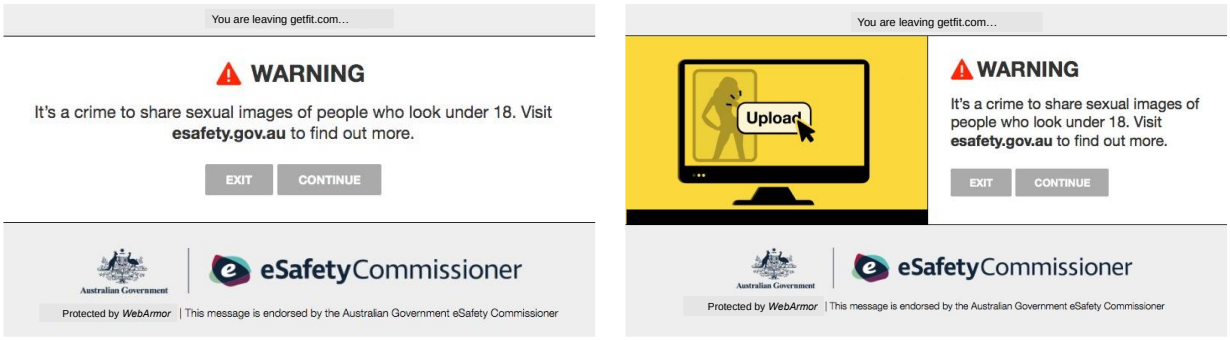



**Figure 7** – Messages used in 'sharing' experiments

and came in two formats – text only and text plus animation (see Figure 7). A key feature if this experiment was the use of an authoritative source – the eSafety Commissioner – to give credibility to the message.

Both messages were effective in significantly dissuading individuals from proceeding to the site, with no significant difference between the two formats (control=60%, text only=43% and text animation=38%).

The results indicate that warning messages can dissuade users from viewing or sharing sexual imagery. We have argued that the effects we observed would be stronger in the CSAM context. In other words, if users' complied with warnings about potentially deviant but ultimately legal pornography, it seems likely that compliance would be greater if messages were received when users were search for (or preparing to share) CSAM.

The experiments were point-in-time; they recorded users' first encounter with a warning message. We have never held the view that CSAM warning messages would have a permanent effect on the behaviour of all users. However, they could permanently influence some users – especially given the wide variety of genres of legal pornography to choose from.

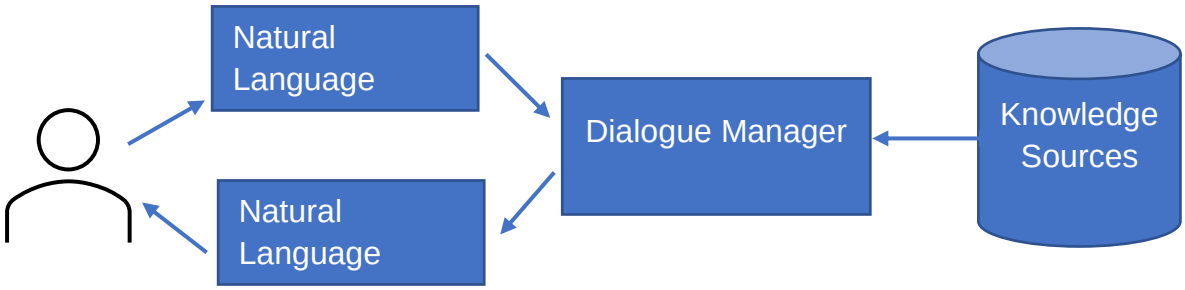
# 4. Chatbots

## 4.1 HISTORICAL CONTEXT OF CHATBOTS

The concept of a computer based conversational agent is nearly as old as computers themselves. The Turing Test, proposed by Alan Turing (1950), is the ultimate test of whether a computer can think for itself, and whether that 'thought' can then be identified as being not of human origin (as the 'imitation game'). The more convincing a conversational agent is, the more human it appears – encouraging users to respond in a natural and true way to those that control the agent (Rapp, Curti & Bodi, 2021). As such, a modern chatbot that approaches passing the Turing Test is a tool that can be used to educate, inform, and even alter the behaviour of target users. Modern chatbots present an opportunity to create new awareness and education-based campaigns to prevent harm in online contexts.

## 4.2 BASIC ARCHITECTURE OF CHATBOTS

Conversational agents, or chatbots, have been used in a broad range of contexts, and are known by many names with over 160 synonyms having been recorded.[6] However, despite their broad usage, and indeed variation between implementations, their basic architecture can be described as having three main components: Natural Language Understanding, Dialogue Manager and Natural Language Generation (McTear, 2020).




The ***Natural Language Understanding*** component where input is received from the user, via whatever interface is being used. The interface is commonly a form of typed content, however spoken content that is then converted to text via automatic speech recognition is also common. The primary purpose of this content is to understand the 'what' (entity extraction) and 'why' (intent extraction) is being communicated by the user (Kulkarni et al, 2019). These objectives are undertaken by using syntactic analysis to analyse the input against grammar rules and semantic analysis to understand the meaning. A representation of the users input, containing confidence values about the entities and intents extracted, is passed to the Dialogue Manager.

The ***Dialogue Manager*** is the primary component of the system, and it is responsible for deciding what the user meant, and what the next course of action will be. This is undertaken by examining what is currently known about the user, combined with the input that was just received. What is currently known about the user is stored within a State Tracking or Context Model subcomponent - which is implemented in a variety of ways across different conversation systems. This context,

[6] https://www.chatbots.org/

combined with the input, is examined by another subcomponent called the Decision Model or the Dialogue Policy module – which is again implemented in a variety of ways depending on the conversation system. If the Dialogue Manager can understand the input, a response is then formulated in accordance with its intended purpose, however, if it is unable to understand, it then moves to establish which information may be missing and will then formulate a response to attain the required information. This could be through responding to the user, or interacting with ***Knowledge Sources*** for more information before then re-evaluating the situation and then responding to the user.

The final component, ***Natural Language Generation***, converts the structured data produced by the Dialogue manager into a human readable form. This process involves multiple stages including what content is in the reply (possibly aggregating from multiple sources), how it is structured, what phrasing is used, and additional features to maximise how life like the response is (Kulkarni et al, 2019). These stages are designed to produce as realistic and natural a response as possible.

## 4.3 EVOLUTION INTO A TOOL FOR MULTIPLE ROLES

Since the first chatbot was created in 1966 (Weizenbaum, 1966) they have advanced dramatically. Early approaches were hand crafted rule-based systems. Such systems commonly implemented with knowledge being handwritten into rules. Artificial Intelligence Markup Language (AIML) is a common format for encoding knowledge that is still used (McTear, 2020). The knowledge encoded into the rules is the core component informing decisions within the Dialogue Manager. Such systems are still common today, used in many task-based systems such as knowledge retrieval, eCommerce, IT support or other domain specific knowledge context.

An alternative approach to Rule Based systems are Statistical Data Driven systems – where large amounts of conversational data is used to train a system without direct human input to formulate the responses. There is a broad range of systems in this category, with the original approaches being statistical model driven approaches, comprised of multiple subcomponents, through to active current research where an end-to-end neural network-based approaches (McTear, 2020).

The most advanced chatbot systems are being produced by the 'big tech' players such as Google, Microsoft, Amazon, Apple and others. They have all invested heavily in the artificial intelligence systems, particularly deep learning approaches that are used in the systems, and they also have access to massive amounts of communication data between their users to train them. Examples of these can be seen in chatbots like by Google's Meena (Adiwardana et al., 2020) and Facebook's BlenderBot (Roller et al., 2020), in addition to commercial products such as Google Assistant, Amazon Alexa, Microsoft Cortana and Apple's Siri.

In addition to the big tech players there are a substantial number of smaller companies producing conversational agents for many different contexts. Some of these are highly generalised, aiming at undertaking basic conversations, which can then be paired with additional knowledge stored within AIML to provide specialised information to an end user. The rapid improvement to these systems, due to increased access of data and refinement of deep learning approaches, has resulted in conversational agents becoming commoditised – an 'off-the-shelf solution' – with the barrier to entry for creating a new chatbot to deploy being quite low, with even low or no cost options available.

## 4.4 REVIEW OF CHATBOT USAGES IN DIFFERENT CONTEXTS

The growth in the number of chatbots being deployed in a wide range of use cases has been dramatic in recent years. Customers of popular chatbot platform Pandorabots have created over 325,000 chatbots, which is only one of the over 1,500 estimated conversational agent platforms (Revang et al., 2019). These platforms enable a broad range of modes of interacting with users, from a simple chat interface in a web page, through to being built into platforms like messaging platforms like WhatsApp, Messenger and WeChat, or more fully features work environments like Teams and Slack. The popularity of chatbots appears to be growing. For instance, Business Insider (2021) estimate that almost 40% of Internet users prefer to interact with a chatbot than a virtual agent. As a result, many domains have implemented chatbots in a variety of forms in recent years.

**Support and service provision for businesses** has been a popular context for chatbot deployment. Chatbots have been equipped with specific industry knowledge and vocabulary, and existing support infrastructure (such as call centres or email conversation histories) are leveraged to provide data to train chatbots. These chatbots can respond to common problems or provide required information, or to efficiently triage a user requesting assistance ahead of forwarding onto trained personnel. Such systems have been popular in IT support (Fiore, Baldauf & Thiel, 2019), finance (Quah & Chua, 2019), hospitality and tourism (Pilla & Sivathanu, 2020), retail (Chung et al., 2020) and education (Dibitonto et al., 2018).

From the broad field of industrial applications of chatbots three studies are highlighted in Box 2 because of their potential relevance to the *reThink* chatbot.

**Box 2: Industry chatbot studies**

**Building rapport**

A recent systematic review of published research identified 29 studies that empirically examined the features of chatbots that seem to enhance rapport and, in particular, trust with users (Rheu et al., 2021). In our view, the notion of 'trust' here seems closely interrelated to the topics of 'credibility' and 'believability' discussed in the ergonomics literature (see above, 3.1). We return to these issues in Section 5.

Several of the themes explored by Rheu and colleagues related to high-end conversational agents. For instance, those with voices or life-like digital avatars. However, other themes are informative for chatbots of all levels of sophistication. These invite consideration of:

- The **communication style**. Chatbots are more likely to be perceived as trustworthy when they use language that is culturally familiar to the user. So, a UK chatbot may need to use slightly different terminology from an Australian one.

- **Openness with users**. Users appear to respond positively to chatbots' open admissions of their short-comings or limitations. Not surprisingly, trust diminishes if users determine that a chatbot is attempting to deceive them in some way.

- The **quality of the content**. The pertinency and accuracy of the information provided by a chatbot improves how it is perceived by users.

As these points infer, to be fit-for-purpose chatbots, like offline warnings (see 3.1), must be tailored to the context at hand. 'There is no one-size-fits-all design', concludes Rheu et al. (2021: 12-13), and effectiveness relies on 'careful consideration of the context and the goal of the tasks'.

**Static messages paired with chatbots**

Silic's (2020) experiment indicated that 38% of users adhered to a static warning about suspicious ULR links that may contain malware. But adherence increased to 61% when the static warning was combined with a chatbot that provided security advice.

**Customer service chatbots**

The results of a controlled experiment on customer service chatbots were published this year (Adam, Wessel & Benlian, 2021). The authors investigated how to improve the ability of a chatbot to secure customer feedback by

- Making it more human through **anthropomorphic** design features. These included giving the chatbot a name ('Alex'), having it refer to itself in the first person, designing it to engage in small-talk, and programming empathetic reactions to some emotional expressions from users. Note that the anthropomorphic design did not entail pretending to be an actual person; and

- Implementing a sales technique called 'foot-in-the-door' (FITD). This entailed initially only asking customers for a 'small favour' (i.e. one survey question). But if it was answered, the chatbot then asked customers for a 'large favour' (to undertake a longer survey, which took up to seven minutes to complete).

The participants in the control group interacted with a chatbot without the anthropomorphic or FITD features. Two thirds (63%) of the group complied with the request for customer feedback. The compliance rates for the experimental groups were:

- 77% with the addition of the FITD feature only;

- 84% with the anthropomorphic feature only; and

- 95% with both the anthropomorphic and FITD features.

The results demonstrated that "subtle changes" in dialog design can increase compliance and that "when employing [conversational agents], and chatbots in particular, providers should design dialogs as carefully as they design the user interface" (Adam, Wessel & Benlian, 2021:438).

The ***health sector*** is another area which has adopted chatbot technology in multiple roles. While technologically similar to systems used in other industries, it is a context which is more personal in nature with a greater level of trust and assurance to clear for users to accept to enable them to be

successful. A 2019 survey of research into health chatbots (also called 'conversational agents') located 4,145 articles across the preceding 10 years (Montenegrom, da Costa, & da Rosa Righi, 2019). Their research found chatbots used in eight broad areas including general practice, neurology, therapy and endocrinology; with a range of age groups being targeted from students to the elderly; and with a broad set of outcomes aimed for including diagnosis, prevention, and education.

Compared with the substantial body of literature from the health sector, relatively little has been published about the use of chatbots in ***crime prevention.*** There appears to be a broad initiative in the city of Pune, India, which adopted an SMS-based chatbot to, among other things: enable users to notify authorities about criminal behaviour; provide some crime prevention tips to help users avoid victimisation; answer users' queries about a variety of criminal laws (including punishments); and to help users locate a hospital near to them (Surana, Chekkala & Bihani, 2021).

Chatbots have been developed to respond to the CSAM problem. Although some of these use sophisticated graphics and dialogue, their purpose to date has largely focussed on ***detecting CSAM offences.*** The Sweetie chatbot (Henseler & de Wolf, 2019), which attracted international media attention almost a decade ago7, is probably the most well-known example. Its purpose was to pretend to be a 10-year-old girl in an online chatroom. Information about Internet users who attempted to procure sex acts from Sweetie was collated and provided to law enforcement agencies (see also Callejas-Rodríguez et al., 2016; Zambrano et al. , 2017). Rodríguez's (2020) C3-Sex chatbot was similar inasmuch as it also set up an online 'sting' by emulating an individual interested in acquiring CSAM. Other forms of chatbots have been designed with classification models to detect grooming behaviour (Anderson et al., 2019; Ebrahimi, 2016; Gunawan, Ashianti & Sekishita, 2018; Meyer, 2015; Michalopoulos, Mavridis, & Jankovic, 2014; see further Sunde & Sunde, 2021).

[7]See e.g. https://www.bbc.com/news/av/world-europe-24819538

# 5. Concluding observations for *reThink* chatbot

The colossal increase in the popularity and availability of CSAM in the space of a few decades has caught policy makers, child welfare organisations, private companies and big tech off-guard. An increasing number of government and private agencies are responding to the CSAM phenomenon from multiple angles over and beyond criminal justice responses. The appearance of some CSAM on legal adult entertainment websites, such as Pornhub, is a worrying social development. It appears to be a contemporary example of the *de facto* normalisation of CSAM on mainstream parts of the Internet which are otherwise law abiding (see e.g. Dines, 2009: 124; Prichard et al., 2013: 997; Warner, 2010: 395).

In this context it is difficult to overstate the potential importance of the *reThink* project and the collaboration of LFF, IWF and MindGeek. At the micro level the project might help individuals avoid CSAM onset, or even to desist from entrenched CSAM offending. The wider implications are hard to predict. But it is feasible that MindGeek's stance constitutes an important cultural reminder from a major pornography company about the boundary between (a) socially acceptable online sexual entertainment and (b) material that is illegal and commonly gravely exploitative.

Situational crime prevention (SCP theory; 2.5) predicts that warnings will dissuade some Internet users from viewing CSAM because at the point of criminal decision-making they (a) act as a deterrent or (b) prick the conscience of the user and remove excuses for the behaviour (Wortley & Smallbone, 2012). Our experiments did not study this exact scenario. Instead, for legal and ethical reasons we examined how naïve users would react to warnings when they attempted to: view adult-minor themed legal pornography; and share sexual images of women (Prichard et al. 2021; forthcoming). The results showed that deterrent policing messages influenced behaviour, as well as simple messages about problematic pornography use (3.7).

Although we contend that our experiments have provided useful evidence about online CSAM messages, crucial facts are unknown about several issues. Of these, two stand out for the *reThink* project, namely:

- The **types of CSAM offenders** most likely to pay attention to and comply with warnings. There are compelling reasons to assume that users who are contemplating onset, or who have just commenced viewing CSAM, will be most responsive. However, it also seems feasible that some users with more complicated offending profiles – e.g. consistent with paedophilia, or hypersexual disorder (Seto and Ahmed, 2014; 2.3) – might become motivated to desist; and

- **Users' reaction to messages over time**. Habituation is clearly one risk. But perhaps the opposite is true too. For instance, it may be some users only comply through repeated exposure to messages. After all, this is exactly the approach taken in anti-smoking campaigns. Reference, for example, the "never quit quitting" campaign in Great Manchester.[8]

[8] https://www.pat.nhs.uk/home-news/Never-Quit-Quitting-campaign-encourages-Greater-Manchester-smokers-to-quit-smoking.htm

While the empirical study of CSAM warning messages has just begun, a staggering amount of research from unrelated fields shows that static warnings and chatbots can and do influence human behaviour. The arc of this literature has uncovered some widely applicable principles that are informative for the management of *reThink*:

**Be specific.** 'Messaging', by which we mean the design literature on static warnings and chatbots, works more effectively when it is clear and tailored to the specific context at hand. Based on the information we have been provided, the *reThink* project achieves this quite well. Rather than a broad 'scatter gun approach' it attempts to communicate with UK users who have just sought CSAM on Pornhub.

**Habituation is not inevitable.** Context matters for compliance (Reeder et al., 2018) and the agency of users to *choose* compliance cannot be discounted (Wogalter, 2019; 3.1, 3.6). In other words, a user may decide to engage with *reThink* even after repeated (and perhaps irritating) exposure. Importantly, *reThink* is offering information and help about (a) disengaging with sexual behaviours that attract extreme levels of social condemnation, and (b) avoiding imprisonment. It is not attempting to sell something, or advise how to safely operate a familiar machine, or to install software that will nominally improve a computer's performance. This context seems a significant strength of *reThink.* Notwithstanding, a one strategy to counter habituation will be to periodically change the appearance of *reThink* (see polymorphism; 3.6). The planned introduction of the chatbot (in combination with the static warning) is a case in point. Other options, including audio and images, can be contemplated in future.

**Trust, believability and credibility are important.** The material on how individuals perceive messages and the agencies that deliver them deserves consideration. Among other things it indicates that transparency with users will be appreciated with regards to their anonymity. Evidently users can likewise appreciate frank admissions by chatbots of their own shortcomings (see Box 2).

**Simple changes can encourage compliance.** Digital sophistication is a significant advantage because it opens up so many opportunities for engaging with users, including voice activation, avatars, and complex dialogue architecture (see 4.2). However, improvements in compliance might be achieved through comparatively simple changes to the language employed by the *reThink* chatbot, such as the use of colloquialisms and commencing interaction with a question (see Box 2).

# References


Adiwardana, D., Luong, M. T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., ... & Le, Q. V. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Anderson, B. B., Vance, A., Kirwan, C. B., Jenkins, J. L., & Eargle, D. (2016). From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. Journal of Management Information Systems, 33(3), 713-743.

Anderson, P., Zuo, Z., Yang, L., & Qu, Y. (2019, June). An Intelligent Online Grooming Detection System Using AI Technologies. In 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-6). IEEE.

ANSI (American National Standards Institute). 2016. *American National Standard Design Principles for Environmental/Facility Safety Signs and Product Labels* (ANSI Z535.X-2016). National Electrical Manufacturers Association.

Babchishin, K. M., Hanson, R. K., & VanZuylen, H. (2015). Online child pornography offenders are different: A meta-analysis of the characteristics of online and offline sex offenders against children. Archives of Sexual Behavior, 44(1), 45-66.

Beech, A. R., Elliott, I. A., Birgden, A., & Findlater, D. (2008). The Internet and child sexual offending: A criminological review. *Aggression and Violent Behavior, 13*(3), 216-228. https://doi:10.1016/j.avb.2008.03.007

Broadhurst, R. (2019). Child sex abuse images and exploitation materials. Child Sex Abuse Images and Exploitation Materials, in Roger Leukfeldt & Thomas Holt, Eds. Cybercrime: the human factor, Routledge.

Business Insider, 2021, Chatbot market in 2021: Stats, trends, and companies in the growing AI chatbot industry, https://www.businessinsider.com/chatbot-market-stats-trends?r=AU&IR=T Accessed: October 30th, 2021.

Cacha, C. A. (2021). Printed Warning Signs, Tags, and Labels: Their Choice, Design, and Expectation of Success. In Handbook of Standards and Guidelines in Human Factors and Ergonomics (pp. 677-692). CRC Press.

Caillon, J., Grall-Bronnec, M., Saillard, A., Leboucher, J., Péré, M., & Challet-Bouju, G. (2021). Impact of Warning Pop-Up Messages on the Gambling Behaviour, Craving, and Cognitions of Online Gamblers: A Randomized Controlled Trial. *Frontiers in Psychiatry*, 12, 711431–711431.

Callejas-Rodríguez, Á., Villatoro-Tello, E., Meza, I., & Ramírez-de-la-Rosa, G. (2016, September). From dialogue corpora to dialogue systems: Generating a chatbot with teenager personality for preventing cyber-pedophilia. In International Conference on Text, Speech, and Dialogue (pp. 531-539). Springer, Cham.

Carr, J. (2004). Child Abuse, Child Pornography and the Internet. London: NCH.

Carr, J. (2017). A brief history of child safety online: child abuse images on the internet. Online Risk to Children: Impact, Protection and Prevention, 5-21.Carpenter, S., Zhu, F., & Kolimi, S. (2014). Reducing online identity disclosure using warnings. *Applied Ergonomics, 45*(5), 1337-1342. https://doi.org/10.1016/j.apergo.2013.10.005

Casteel, M. A. (2016). Communicating increased risk: An empirical investigation of the National Weather Service's impact-based warnings. Weather, Climate, and Society, 8(3), 219-232.

Chainey, S. (2021). A quasi-experimental evaluation of the impact of forensic property marking in decreasing burglaries. *Security Journal*, 1-20.

Chen, H., & Li, W. (2017). Mobile device users' privacy security assurance behavior: A technology threat avoidance perspective. Information & Computer Security

Chernoff, W. A. (2021). The new normal of web camera theft on campus during COVID-19 and the impact of anti-theft signage. *Crime Science*, *10*(1), 1-10.

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. Journal of Business Research, 117, 587-595.

Clarke, R. V. (2017). Situational crime prevention. In R. Wortley & M. Townsley (Eds.), *Environmental criminology and crime analysis* (2nd ed., pp. 286–303). Routledge.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., ... & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Political Behavior, 42(4), 1073-1095.

Coopersmith, J. (2006). Does your mother know what you really do? The changing nature and image of computer-based pornography. History and Technology, 22(1), 1-25.

Dines, G. (2009). Childified women: How the mainstream porn industry sells child pornography to men. In S. Olfman (Ed.), The sexualization of childhood, 121–142. Praeger.

Dibitonto, M., Leszczynska, K., Tazzi, F., & Medaglia, C. M. (2018). Chatbot in a campus environment: design of LiSA, a virtual assistant to help students in their university life. In International Conference on Human-Computer Interaction (pp. 103-116). Springer, Cham.

Ebrahimi, M. (2016). Automatic identification of online predators in chat logs by anomaly detection and deep learning (Doctoral dissertation, Concordia University, Montreal, QC, Canada).

ECPAT International. (2018). Towards a global indicator: on unidentified victims in child sexual exploitation material – summary report. Bangkok.

Egelman, S., Cranor, L. F., & Hong, J. (2008, April). You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1065-1074). ACM.

Essers, L. (2013). Google to warn users of 13,000 search terms associated with child pornography. PCWorld. https://www.pcworld.com/article/2064520/google-to-warn-users-of-13000-search-terms-associated-with-child-pornography.html;

EUROPOL. (2020). *Internet Organised Crime Threat Assessment (IOCTA) 2020* [Report] . Retrieved from https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020

Fiore, D., Baldauf, M., & Thiel, C. (2019). " Forgot your password again?" acceptance and user experience of a chatbot for in-company IT support. In Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia (pp. 1-11).

Fortin, F., Paquette, S., & Dupont, B. (2018). From online to offline sexual offending: Episodes and obstacles. Aggression and Violent Behavior, 39, 33–41.

Fortin, F., & Proulx, J. (2019). Sexual interests of child sexual exploitation material (CSEM) consumers: four patterns of severity over time. *International journal of offender therapy and comparative criminology*, *63*(1), 55-76. https://doi.org/10.1177/0306624X18794135

Gainsbury, S., Aro, D., Ball, D., Tobar, C., & Russell, A. (2015). Determining optimal placement for pop-up messages: Evaluation of a live trial of dynamic warning messages for electronic gaming machines. *International Gambling Studies, 15*(1), 141-158. https://doi.org/10.1080/14459795.2014.1000358

Ginley, M. K., Whelan, J. P., Pfund, R. A., Peter, S. C., & Meyers, A. W. (2017). Warning messages for electronic gambling machines: evidence for regulatory policies. Addiction Research & Theory, 25(6), 495-504.

Global Alliance Against Child Sexual Abuse Online. (2014). Norway. https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/organized-crime-and-human-trafficking/global-alliance-against-child-abuse/docs/reports-2014/ga_report_2014_-_norway_en.pdf

Google. (2020). Fighting child sexual abuse online. https://protectingchildren.google/intl/en/

Gunawan, F. E., Ashianti, L., & Sekishita, N. (2018). A simple classifier for detecting online child grooming conversation. Telkomnika, 16(3), 1239-1248

Haddad, A., Sauer, J., Prichard, J., Spiranovic, C., & Gelb, K. (2020). Gaming Tasks as a Method for Studying the Impact of Warning Messages on Information Behavior. Library Trends, 68(4), 576-598.

Hall, M. G., Lazard, A. J., Grummon, A. H., Higgins, I. C., Bercholz, M., Richter, A. P. C., & Taillie, L. S. (2021). Designing warnings for sugary drinks: A randomized experiment with Latino parents and non-Latino parents. Preventive Medicine, 148, 106562.

Hancock, P. A., Kaplan, A. D., MacArthur, K. R., & Szalma, J. L. (2020). How effective are warnings? A meta-analysis. Safety Science, 130, 104876.

Henseler, H., & de Wolf, R. (2019). Sweetie 2.0 Technology: Technical Challenges of Making the Sweetie 2.0 Chatbot. In Sweetie 2.0 (pp. 113-134). TMC Asser Press, The Hague.

Holt, T. J., Cale, J., Leclerc, B., & Drew, J. (2020). Assessing the challenges affecting the investigative methods to combat online child exploitation material offenses. *Aggression and violent behavior*, 101464.

Home Affairs Committee. (2018). Policing for the future: Report. House of Commons.

Hunn, C., Prichard, J., & Cockburn, H. (2021). Internet users' beliefs about a novice-user of child sexual abuse material (CSAM): what can they tell us about introducing offender-focused prevention initiatives?. Victims & Offenders, 1-21.

Hunn, C., Watters, P., Prichard, J., Wortley, R., Scanlon. J., Spiranovic, C., & Krone, T. (under review). Implementing online warnings to prevent CSAM use: a technical overview.

Hunn, C., Spiranovic, C., Prichard, J., & Gelb, K. (2020). Why internet users' perceptions of viewing child exploitation material matter for prevention policies. Australian & New Zealand Journal of Criminology, 53(2), 174-193.

INTERPOL. (2020). *Threats and trends child sexual exploitation and abuse: COVID-19 impact*. Retrieved from https://www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-highlights-impact-of-COVID-19-on-child-sexual-abuse

Internet Watch Foundation. (2017). *Annual report 2017*.

Jobs, S. (1983) The Future Isn't What It Used To Be. Conference proceedings of the International Design Conference, Aspen, Colorado. https://www.youtube.com/watch?v=tSoxljtN4Xk accessed 05/11/2021.

Junger, M., Montoya, L., & Overink, F. J. (2017). Priming and warnings are not effective to prevent social engineering attacks. Computers in human behavior, 66, 75-87

Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J. N., & Mayer, J. (2021). Adapting Security Warnings to Counter Online Disinformation. In 30th {USENIX} Security Symposium ({USENIX} Security 21).

Kim, S., & Wogalter, M. S. (2015). Effects of emphasis terminology in warning instructions on compliance intent and understandability. Journal of safety research, 55, 41-51.

Kim, S., & Wogalter, M. S. (2009, October). Habituation, dishabituation, and recovery effects in visual warnings. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 53, No. 20, pp. 1612-1616). Sage CA: Los Angeles, CA: SAGE Publications.

Krone, T. (2004). *A typology of online child pornography offending*. Canberra: Australian Institute of Criminology.

Kulkarni, P., Mahabaleshwarkar, A., Kulkarni, M., Sirsikar, N. and Gadgil, K. (2019) September. Conversational AI: An Overview of Methodologies, Applications & Future Scope. In 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA) (pp. 1-7). IEEE.

Kyvsgaard, B., & Sorensen, D. W. (2021). Do stickers indicating the use of forensic property marking prevent burglary? Results from a randomized controlled trial. *Journal of Experimental Criminology*, *17*(2), 287-303.

Lanning, K. V. (2010). Child molesters: A behavioral analysis for professionals investigating the sexual exploitation of children (5th ed., pp. 1-212): National Center for Missing & Exploited Children.

Laughery, K., & Page-Smith, K. (2006). Explicit information in warnings. In M. S. Wogalter (Ed.), *Handbook of warnings* (pp. 419-428). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Lenorovitz, D. R., Karnes, E. W., & Leonard, S. D. (2014). Mitigating product hazards via user warnings alone: When/why "warnings-only" approaches are likely to fail. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *24*(3), 275-297.

Lenorovitz, D. R., Leonard, S. D., & Karnes, E. W. (2012). Ratings checklist for warnings: a prototype tool to aid experts in the adequacy evaluation of proposed or existing warnings. *Work*, *41*(Supplement 1), 3616-3623.

Ly, T., Dwyer, R. G., & Fedoroff, J. P. (2018). Characteristics and treatment of internet child pornography offenders. *Behavioral Sciences and the Law, 36*(2), 216-234. https://doi:10.1002/bsl.2340

Maimon, D., Alper, M., Sobesto, B., & Cukier, M. (2014). Restrictive deterrent effects of a warning banner in an attacked computer system. *Criminology, 52*(1), 33-59. https://doi:10.1111/1745-9125.12028

Malamuth, N. M. (2018). "Adding fuel to the fire"? Does exposure to non-consenting adult or to child pornography increase risk of sexual aggression?. *Aggression and violent behavior*, 41, 74-89.

Martijn, C., Smeets, E., Jansen, A., Hoeymans, N., & Schoemaker, C. (2009). Don't get the message: The effect of a warning text before visiting a pro-anorexia website. *International Journal of Eating Disorders*, 42(2), 139–145

Mayhew, P., Clarke, R. V. G., Sturman, A., & Hough, J. M. (1975). *Crime as Opportunity.* London: Home Office Research and Planning Unit.

McTear, M. (2020). Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. *Synthesis Lectures on Human Language Technologies*, *13*(3), 1-251.

Merdian, H. L., Perkins, D. E., Dustagheer, E., & Glorney, E. (2018). 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*, 1-19. https://doi:10.1177/0306624X17748067

Meyer, M. Machine Learning to Detect Online Grooming. Master's Thesis, Department of Information Technology, Uppsala University, Uppsala, Sweden, 2015

Michalopoulos, D., Mavridis, I., & Jankovic, M. (2014). GARS: Real-time system for identification, assessment and control of cyber grooming attacks. Computers & Security, 42, 177-190.

Mollen, S., Engelen, S., Kessels, L. T., & van den Putte, B. (2017). Short and sweet: the persuasive effects of message framing and temporal context in antismoking warning labels. *Journal of health communication*, *22*(1), 20-28.

Montenegro, J. L. Z., da Costa, C. A., & da Rosa Righi, R. (2019). Survey of conversational agents in health. Expert Systems with Applications, 129, 56-67.

Morgan, S., & Lambie, I. (2019). Understanding men who access sexualised images of children: exploratory interviews with offenders. *Journal of Sexual Aggression, 25*(1), 60-73. https://doi.org/10.1080/13552600.2018.1551502

Motivans, M., & Kyckelhahn, T. (2007). *Federal prosecution of child sex exploitation offenders, 2006*. Washington, DC: US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Neupane, A., Saxena, N., Maximo, J. O., & Kana, R. (2016). Neural markers of cybersecurity: an fMRI study of phishing and malware warnings. IEEE Transactions on information forensics and security, 11(9), 1970-1983.

Ng, A. W., & Chan, A. H. (2009). *What makes an icon effective?* AIP Conference Proceedings.

Paul, B & Linz, D.G. (2008), The effects of exposure to virtual child pornography on viewer cognitions and attitudes toward deviant sexual behavior, *Communication Research*, 35(1), 3-38.

Perkins, D., & Wefers, S. (2019). Treatment of Internet-related sexual offenders. In J.L. Ireland, C.A. Ireland & P Birch (eds) *Violent and Sexual Offenders: Assessment, Treatment and Management (2nd ed).* Abingdon, Oxon: Routledge.

Petelka, J., Zou, Y., & Schaub, F. (2019, May). Put your warning where your link is: Improving and evaluating email phishing warnings. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-15).

Peters, E. M., Morrison, T., McDermott, D. T., Bishop, C. J., & Kiss, M. (2014). Age is in the eye of the beholder: Examining the cues employed to construct the illusion of youth in teen pornography. *Sexuality & Culture*, *18*(3), 527-546.

Pillai, R., & Sivathanu, B. (2020). Adoption of AI-based chatbots for hospitality and tourism. International Journal of Contemporary Hospitality Management.

Price, C. (2020). *17 Great Search Engines You Can Use Instead of Google*. Retrieved from https://www.searchenginejournal.com/alternative-search-engines/271409/#close

Prichard, J., Scanlan, J., Krone, T., Spiranovic, C., Watters, P., & Wortley, R. (in press). Warning messages to prevent illegal sharing of sexualised images: Results of a randomised controlled experiment. Trends and Issues in Criminal Justice, Australian Institute of Criminology.

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.

Prichard, J., Watters, P. A., & Spiranovic, C. (2011). Internet subcultures and pathways to the use of child pornography. Computer Law & Security Review, 27(6), 585-600.

Prichard, J., Wortley, R., Watters, P. A., Spiranovic, C., Hunn, C., & Krone, T. (2021). Effects of Automated Messages on Internet Users Attempting to Access "Barely Legal" Pornography. Sexual Abuse.

Quah, J. T., & Chua, Y. W. (2019). Chatbot assisted marketing in financial service industry. In International Conference on Services Computing (pp. 107-114). Springer, Cham.

Quayle, E. (2020). Online sexual deviance, pornography and child sexual exploitation material. *Forensische Psychiatrie, Psychologie, Kriminologie*, 1-8.

Quayle, E. (2012). Organisational issues and new technologies. In M. Erooga (Ed.), *Creating safer organisations: practical steps to prevent the abuse of children by those working with them*. Chichester: Wiley-Blackwell.

Quayle, E., & Koukopoulos, N. (2019). Deterrence of online child sexual abuse and exploitation. *Policing*, *13*(3), 345–362.

Quayle, E., & Taylor, M. (2004). *Child pornography: An internet crime*. Abington, Oxon: Routledge

Ray, J. V., Kimonis, E. R., & Seto, M. C. (2014). Correlates and moderators of child pornography

consumption in a community sample. *Sexual Abuse*, 26, 523–545.

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. International Journal of Human-Computer Studies, 102630.

Reeder, R. W., Felt, A. P., Consolvo, S., Malkin, N., Thompson, C., & Egelman, S. (2018, April). An Experience Sampling Study of User Reactions to Browser Warnings in the Field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 512). ACM.

Reuter, C., Kaufhold, M. A., Leopold, I., & Knipp, H. (2017, April). Informing the Population: Mobile Warning Apps. In Risk and Crisis Communication for Disaster Prevention and Management-Workshop Proceedings. Wilhelmshaven: Jade Hochschule (pp. 31-41).

Revang, M., Baker, V., Manusama, B., & Mullen, A. (2019). Market guide for conversational platforms. Technical report, Report ID ID:G00367775, Gartner.

Riley, D., Ingegneri, L., Passi, L., & Siqueira, J. (2006) *Modelling exposure outcomes to improve warning assessment and design for chemical consumer products*. Paper presented at the Proceedings of the 16th Congress of the International Ergonomics Association, Maastricht, The Netherlands.

Rodríguez, J. I., Durán, S. R., Díaz-López, D., Pastor-Galindo, J., & Mármol, F. G. (2020). C3-sex: A conversational agent to detect online sex offenders. Electronics, 9(11), 1779.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... & Weston, J. (2020). Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Rosenblatt, D. H., Bode, S., Dixon, H., Murawski, C., Summerell, P., Ng, A., & Wakefield, M. (2018). Health warnings promote healthier dietary decision making: Effects of positive versus negative message framing and graphic versus text-based warnings. Appetite, 127, 280-288.

Rushkoff, D. (2009). The Web's dirtiest site. Retrieved 1 July 2019 from www.thedailybeast.com/blogs-and-stories/2009-08-11/the-webs-dirtiest-site

Selejan, O., Muresanu, D., Popa, L., Muresanu-Oloeriu, I., Iudean, D., Buzoianu, A., & Suciu, S. (2016). Credibility judgments in web page design–a brief review. *Journal of Medicine and Life, 9*(2), 115.

Seto, M. C. (2019). The motivation-facilitation model of sexual offending. *Sexual Abuse*, *31*(1), 3-24.

Seto, M. C., & Ahmed, A. G. (2014). Treatment and management of child pornography use. *Psychiatric Clinics of North America*, *37*(2), 207-214.

Silic, M. (2020). Improving warning messages adherence: can Maya Security Bot advisor help?. Security Journal, 33(2), 293-310.

Silic, M., & Back, A. (2017). Deterrent Effects of Warnings on User's Behavior in Preventing Malicious Software Use. In Silic, M., & Back, A.(2017, January). Deterrent Effects of Warnings on User's Behavior in Preventing Malicious Software Use. In Proceedings of the 50th Hawaii International Conference on System Sciences.

Smallbone, S., & Wortley, R. (2017). Preventing child sexual abuse online. Online risk to children: Impact, protection and prevention, 143.

Steel, C. M. S. (2015). Web-based child pornography: The global impact of deterrence efforts and its consumption on mobile platforms. *Child Abuse and Neglect, 44*, 150-158. https://doi:10.1016/j.chiabu.2014.12.009

Surana, S., Chekkala, J., & Bihani, P. (2021). Chatbot based Crime Registration and Crime Awareness System using a custom Named Entity Recognition Model for Extracting Information from Complaints. International Research Journal of Engineering and Technology (IRJET), Volume 8, issue 4, April 2021.

Taylor, M., & Quayle, E. (2008). Criminogenic qualities of the Internet in the collection and distribution of abuse images of children. *Irish Journal of Psychology, 29*(1-2), 119-130. https://doi:10.1080/03033910.2008.10446278

Taylor, J. R., & Wogalter, M. S. (2019). Specific egress directives enhance print and speech fire warnings. Applied ergonomics, 80, 57-66.

Testa, A., Maimon, D., Sobesto, B., & Cukier, M. (2017). Illegal roaming and file manipulation on target computers: Assessing the effect of sanction threats on system trespassers' online behaviors. *Criminology and Public Policy*, 16(3), 689–726.

Tomczyk, S., Rahn, M., Markwart, H., & Schmidt, S. (2021). A walk in the park? Examining the impact of app-based weather warnings on affective reactions and the search for information in a virtual city. International journal of environmental research and public health, 18(16), 8353.

Turing, A. (2006). Computing machinery and intelligence. Theories of Mind: An Introductory Reader, 51. https://doi.org/10.1093/mind/LIX.236.433

Tweedie, S. (2014) https://www.businessinsider.com.au/best-of-steve-jobs-playboy-interview-2014-9

Ullman, J. R. (2017). The development and testing of potential music piracy warnings (79. University of Nevada, Las Vegas.

Ullman, J. R., & Silver, N. C. (2018). Perceived effectiveness of potential music piracy warnings. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), 1353–1357.

Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology, 53*(2), 134-144. https://doi.org/10.1002/asi.10016

Watters, P. A. (2009). Why do users trust the wrong messages? A behavioural model of phishing. In 2009 eCrime Researchers Summit (pp. 1-7). IEEE.

Watters, P. A., Lueg, C., Spiranovic, C., & Prichard, J. (2013). Patterns of ownership of child model sites: Profiling the profiteers and consumers of child exploitation material. First Monday.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45

We Protect (2019). *Global threat assessment 2019: Working together to end the sexual exploitation of children online*. London: Open Government Licence. https://static1.squarespace.com/static/5630f48de4b00a75476ecf0a/t/5deecb0fc4c5ef23016423cf/1575930642519/FINAL+-+Global+Threat+Assessment.pdf

Westlake, B. G. (2020). The past, present, and future of online child sexual exploitation: Summarizing the evolution of production, distribution, and detection. *The Palgrave handbook of international cybercrime and cyberdeviance*, 1225-1253

Westlake, B., & Bouchard, M. (2016). Criminal careers in cyberspace: Examining website failure within child exploitation networks. *Justice Quarterly*, *33*(7), 1154-1181.

Westlake, B., Bouchard, M., & Frank, R. (2017). Assessing the validity of automated webcrawlers as data collection tools to investigate online child sexual exploitation. *Sexual Abuse*, *29*(7), 685-708.

Wogalter, M. S. (2020). Forensic human factors and ergonomics analysis of a trip and fall event in a parking lot. Theoretical Issues in Ergonomics Science, 21(3), 347-368.

Wogalter, M. S., Laughery, K. R., & Mayhorn, C. B. (2012). Warnings and hazard communications. Handbook of human factors and ergonomics, 4th ed., 868-94.

Wogalter, M. S., & Mayhorn, C. B. (2008). Trusting the internet: Cues affecting perceived credibility. *International Journal of Technology and Human Interaction (IJTHI), 4*(1), 75-93. https://doi.org/10.4018/jthi.2008010105

Wolak, J., Finkelhor, D., & Mitchell, K. J. (2011). Child pornography possessors: Trends in offender and case characteristics. Sexual Abuse: A Journal of Research and Treatment, 23(1), 22-42.

Wolak, J., Liberatore, M., & Levine, B. N. (2014). Measuring a year of child pornography trafficking by US computers on a peer-to-peer network. *Child Abuse & Neglect*, *38*(2), 347-356

Wortley, R. (2012). 'Situational prevention of child abuse in the new technologies', in E Quayle & K Ribisl (eds), Understanding and preventing online sexual exploitation of children, Routledge, London, 188–204.

Wortley, R., & Smallbone, S. (2006). Child pornography on the internet (pp. 5-2006). Washington, DC: US Department of Justice, Office of Community Oriented Policing Services.

Wortley, R., & Smallbone, S. (2012). Internet child pornography: Causes, investigation, and prevention. ABC-CLIO.

Zaikina-Montgomery, H. (2011). The dilemma of minors' access to adult content on the internet: A proposed warnings solution [Doctoral thesis]. University of Nevada.

Zambrano, P., Sanchez, M., Torres, J., & Fuertes, W. (2017). BotHook: An option against Cyberpedophilia. In 2017 1st Cyber Security in Networking Conference (CSNet) (pp. 1-3). IEEE.

Zielinska, O. A., Mayhorn, C. B., & Wogalter, M. S. (2017). Connoted hazard and perceived importance of fluorescent, neon, and standard safety colors. Applied ergonomics, 65, 326-334.

# Annotated bibliography

**Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. Electronic Markets, 31(2), 427-445.**

The authors aimed to gain a greater understanding of what makes chatbots effective, with specific attention given to whether or not the chatbot told the user that it was a chatbot. They ran a randomised control study (providing customer support in a finance setting) exploring the effect of anthropomorphic design cues (i.e small talk, empathy) and foot-in-the-door (to undertake a small task, before the main task) measuring the outcome in relation to the completion of a compliance task. The participants were informed that they were interacting with a chatbot, and the authors found that both methods improved compliance over the control, however the best outcome was using both dropped non-compliance from 37% down to 5%. The study demonstrated the value of anthropomorphism, but not to the goal of having the user believe it is a person (as it stated it was a bot), but to emulate a natural conversation which then receives a social response.

**Babchishin, K. M., Hanson, R. K., & VanZuylen, H. (2015). Online child pornography offenders are different: A meta-analysis of the characteristics of online and offline sex offenders against children. Archives of Sexual Behavior, 44(1), 45-66.**

This paper reports a meta-analysis based on 30 unique samples of online CSAM offenders and contact CSA offenders. Results showed the importance of opportunity in offending: CSA offenders were more likely to have access to children while CSAM offenders had greater access to digital technologies. In terms of psycho-social profiles, CSAM offenders had lower scores on antisocial and psychopathology variables (e.g., fewer prior convictions, fewer problems with supervision, higher education levels, less detached romantic relationships, less childhood difficulties and abuse, less incidence of mental illness), and were more likely to possess 'psychological barriers' to sex offending (e.g., greater victim empathy, greater emotional identification with children, less use of cognitive distortions). CSAM offenders, however, were more likely show a greater sexual interest in children, although the authors suggest that this finding may be an artefact of the respective ways that CSAM and CSA offences are prosecuted. Specifically, they argue that cases against CSAM offenders typically involve possession of images of obviously pre-pubescent children whereas many CSA cases involve victims up to the age of 18 years. In sum, the study highlights the role of opportunity in both CSAM and CSA offending, while also showing that CSAM and CSA offenders differ in important psychological and socio-demographic ways.

**Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. International Journal of Human-Computer Studies, 102630.**

Within this systematic review, the authors found 83 papers which focus on how humans interact with chatbots, mapping out the core themes across the preceding 10 years. The themes included acceptance, expectation, engagement, emotional experience, and humanness (among others). The paper discusses the impact of design goals (such as imitating a human, and including social cues), and highlighted the need for further research. Multiple papers in the review describe positive effects of chatbots being human like including social elements and showing empathy (aligning with Adam, Wessel, & Benlian (2021)), while a group of others did report negative outcomes for chatbots pretending to be human – but highlighting that it can be context dependant.

**Rodríguez, J. I., Durán, S. R., Díaz-López, D., Pastor-Galindo, J., & Mármol, F. G. (2020). C3-sex: A conversational agent to detect online sex offenders. Electronics, 9(11), 1779.**

This paper, like many others (Callejas-Rodríguez et al, 2016; Zambrano et al, 2017; Henseler & de Wolf, 2019), have examined the feasibility of an automated tool, based on a chatbot, for identifying perpetrators online. Instead of using law enforcement officers within chatrooms and other online environments, a chatbot is deployed, serving as a 'honey-pot' (similar to the authors of this literature existing work (refs – or outlined in section X)). Such systems are seen as vastly more scalable than using law enforcement officers, with the C3-sex chatbot used in this study interacting with 7199 users in an eight-week period of the trial.

**Sillic, M. (2020). Improving warning messages adherence: can Maya Security Bot advisor help? Security Journal, 33(2), 293-310.**

This paper bridges the areas of chatbots and awareness messaging, which is highly relevant to the work presented within the literature review. The author compared the efficacy of a static warning, or static warning combined with chatbot within the context of cybersecurity warnings (i.e a user opens a link in an email, and attempts to access a possibly malicious website, and then is show either the chatbot or a warning message). The work demonstrated a significant difference between the static alone, and static with a chatbot. The paper promoted "social facilitation", having the bot be anthropomorphised. This work provides evidence that a chatbot can increase adherence over a static message, providing precedence for the proposed work relating to this literature review.

**Wolak, J., Finkelhor, D., & Mitchell, K. J. (2011). Child pornography possessors: Trends in offender and case characteristics. Sexual Abuse: A Journal of Research and Treatment, 23(1), 22-42.**

This paper is from the prolific research team, led by David Finkelhor at the Crimes Against Children Center, that specialises in conducting large-scale surveys of Internet to collect empirical data on child sexual exploitation, in this case drawing on data from a survey of more than 2,500 US law enforcement agencies. In terms of offender profiles, key findings included: 1) offenders were predominately male (99%), white (89%), older (49%>40 years), employed (61%), and single (68%); 2) Few offenders had backgrounds of mental illness (6%), diagnosis of sexual disorder, including paedophilia (1%), violence (12%), substance abuse 20%), or prior arrests for sexual (9%) or other (27%) offences"; 3) offenders generally showed limited computer skills and took few precautions to hide images and avoid detection; and 4) a minority of offenders arrested initially for possession of CSAM were also arrested for concurrent or previous contact child sexual abuse (16%). In terms of the nature of CSAM possessed: 1) the most common ages of victims was 6-12 years (86%), with 13-17 years next most common (67%); 2) girls were the most common victims (69%); 3) the most common type of CSAM involved graphic sexual images (found in 94% of cases) followed by penetration (82%), non-graphic nudity (82%), sexual contact with an adult (75%), and violence (24%); and 4) In most cases (68%) the offender also possessed pornography involving adults.

Published by the University of Tasmania

ISBN: 978-1-922708-21-2

# Suggested citation

Prichard J, Scanlan J, Watters P, Wortley R, Hunn C, Garrett E, 'Online messages to reduce users' engagement with child sexual abuse material: a review of relevant literature for the rethink Chatbot', University of Tasmania, Hobart. ISBN 978-1-922708-21-2 (2022)

# Author contact details

Jeremy Prichard – jeremy.prichard@utas.edu.au

Joel Scanlan – joel.scanlan@utas.edu.au

Paul Watters – paul.watters@cantab.net

Richard Wortley – r.wortley@ucl.ac.uk