future internet

MDPI

Article

Creating Honeypots to Prevent Online Child Exploitation

Joel Scanlan [1,*], Paul A. Watters [2,*], Jeremy Prichard [3], Charlotte Hunn [3], Caroline Spiranovic [3] and Richard Wortley [4]

[1] Australian Institute of Health Service Management, University of Tasmania, Hobart, TAS 7005, Australia
[2] Cyberstronomy, Melbourne, VIC 3000, Australia
[3] School of Law, University of Tasmania, Hobart, TAS 7005, Australia; jeremy.prichard@utas.edu.au (J.P.); charlotte.hunn@utas.edu.au (C.H.); caroline.spiranovic@utas.edu.au (C.S.)
[4] Department of Security and Crime Science, University College London, London WC1H 9EZ, UK; r.wortley@ucl.ac.uk
* Correspondence: joel.scanlan@utas.edu.au (J.S.); paul.watters@cantab.net (P.A.W.)

Abstract: Honeypots have been a key tool in controlling and understanding digital crime for several decades. The tool has traditionally been deployed against actors who are attempting to hack into systems or as a discovery mechanism for new forms of malware. This paper presents a novel approach to using a honeypot architecture in conjunction with social networks to respond to non-technical digital crimes. The tool is presented within the context of Child Exploitation Material (CEM), and to support the goal of taking an educative approach to Internet users who are developing an interest in this material. The architecture that is presented in the paper includes multiple layers, including recruitment, obfuscation, and education. The approach does not aim to collect data to support punitive action, but to educate users, increasing their knowledge and awareness of the negative impacts of such material.

Keywords: child exploitation prevention; honeypot; social media

Citation: Scanlan, J.; Watters, P.A.; Prichard, J.; Hunn, C.; Spiranovic, C.; Wortley, R. Creating Honeypots to Prevent Online Child Exploitation. *Future Internet* **2022**, *14*, 121. https://doi.org/10.3390/fi14040121

Academic Editor: Tinghuai Ma

Received: 10 March 2022
Accepted: 10 April 2022
Published: 14 April 2022

1. Introduction

The ubiquity of Information and Communications Technology (ICT) to be an integral part of our daily lives has resulted in digital crime becoming commonplace. This wave of technological change has enabled not only new forms of crime to be committed, but for existing crime to move to the digital world. These offences present a dramatically different context to traditional place-based crimes, requiring new approaches in combating the risks being faced.

Cybercrime, as it can be termed, includes both crimes directed at ICT systems (such as computers and networks), and where the ICT system itself is a key part in the commission of the crime [1]. The cost of cybercrime is often calculated based on revenue loss as a result of business disruption, information and productivity loss, and equipment damage as a result of a cyber-attack. These cyber-attacks take multiple forms, including denial of service attacks, hacking and with an increased impact in recent years, from malicious code such as ransomware. These attacks are of a technical nature, and there is a broad set of challenges faced by law enforcement in responding to such attacks which commonly cross jurisdictional boundaries [2]. Other crimes, enabled by cyber systems, do not have a dollar value, as their motivation is psychosocial or geopolitical in nature. Psychosocial cybercrimes include cyber stalking and bullying, and the dissemination of Child Exploitation Material (CEM) [3]. These crimes are people-centric, and cause harms which are psychological in nature, resulting in long-lasting effects for the victims. Geopolitical include cyber based hate speech and vandalism.

As laws increase in volume and complexity to adapt to the changing nature of digital crimes, the need for strategies to educate the public about the parameters of legal rules

becomes more apparent [4]. Relatedly, crime-prevention literature has underscored that the likelihood of criminal decision-making can be reduced by providing potential offenders with information at the point of crime commission [5]. This includes information about the risk of detection by law enforcement agencies and the harms associated with the particular behaviour [5]. By way of example, road signs serve to inform the public of the legal speed for a section of road, enabling users to choose to travel within the limit. The speed limit not only protects the driver, but also other road users and those in the vicinity. When examining digital crimes and their context, these mechanisms are not as easy to apply. We lack a mechanism that is the context equivalent for a speed sign beside the road.

Many traditional approaches within the cyber security field for responding to threats (which could escalate to a cybercrime) are technical, due to the threats themselves being technical in nature [6–9]. Common preventative measures are barriers and system hardening methods, which aim to prevent a future attack from being successful and mitigate the impact of when it does occur. However, these approaches are less effective in addressing people-centric crimes such as the distribution of CEM, in which the initial abuse may occur offline, but the harms caused are perpetuated further online when content featuring the original abuse is circulated. In this context, we are commonly only left with reactive tools such as requesting that content is removed or that a particular user be banned from a platform. Few programs exist [10] which seek to proactively deter potential offenders from engaging in prohibited online behaviour by educating users and raising their awareness about the risk and harm involved in the behaviour.

This paper describes research that has been undertaken to address the current research vacuum in this area. It presents an infrastructure that provides an opportunity to educate Internet users who may be at risk of deciding to commit a cybercrime. The research draws upon existing work within the cyber security area, which focuses on the detection and response to hacking and malware threats but applies it within the area of CEM offending prevention. The infrastructure details a honeypot that was created and used to display warning messages to users who attempted to access a fake pornographic website. The infrastructure has multiple layers and enabled the research to be undertaken in a controlled setting with only limited access to the study to enable careful monitoring of its usage and overall effectiveness.

The paper first introduces the concept of a honeypot, describing their use, contrasted with other cyber security tools. A brief overview of evidence-based literature on CEM prevention is then provided to demonstrate the potential value of honeypot research in this context. In the final section of this paper, we present and describe the infrastructure used within the two experimental trials that we have conducted to date, demonstrating its ability at recruiting naïve participants within an online research study.

2. Background

Cyber security infrastructure has evolved rapidly across the last four decades since its inception in the 1980s. Honeypots fill a role within this infrastructure to enable bad actors to be discovered and monitored [11]. As a preventative tool, they are quite distinct from the other forms of defence that are commonly used within the industry.

When defining cyber security infrastructure, we are typically describing firewalls, detection systems, and cryptographic tools, in addition to other protective measures in our operating systems. These systems are about controlling access to our networks and systems, allowing access to those with permission, and denying access from those who do not.

One of the first tools developed to do this, and the clearest example of denying access, is the firewall. First described and implemented in the 1980s [6], fundamentally, firewalls act as a barrier between our network and others, or even between different sections of our own network. They prevent access and obfuscate the structure of our network and the number of devices they contain, serving to not only keep bad actors out, but to also limit

the knowledge they have of our systems. They are a proactive defensive measure, which prevents attacks from occurring.

A second commonly used network security tool which was first proposed in the 1980s is an Intrusion Detection System [7]. Intrusion detection can take multiple forms, from being installed on a host of interest, or directly upon the network to observe traffic as it travels past the sensor. Like a firewall, intrusion detection systems are an active defensive measure. They can alert the administrator to events that are occurring or respond to the event in real-time itself by denying access to traffic. Antivirus systems operate on a similar premise. These detection and response systems work to thwart an attack in progress, to prevent it from continuing and from causing greater harm. The use of such systems plays a key role in prevention, as they increase the likelihood of discovery, and play an active role in stopping attacks when they are in progress.

Further, cryptography can play a key preventative role in our defences. It is vital in protecting the confidentiality of data; proving the identity of users we are interacting with, or are interacting with our systems; and integrity, by enabling us to monitor unauthorised change. It serves to increase the difficulty of an attack, in multiple contexts, but to also enable users to detect when an attack has occurred.

Other techniques that are commonly discussed and applied relate to our operating systems, including ensuring that software is up to date and that security holes are fixed. The aim here is to not only prevent the possibility of attacks occurring, but also to decrease the attractiveness of the target and increase the difficulty of an attack to, ultimately, reduce the likelihood that it will succeed.

Each of the forms of defence described above aim to harden the system or network, making it more difficult for a bad actor to attack. Honeypots are quite different in their operational philosophy. The next section describes the role that honeypots have, and how they are complementary to these other forms of defences, leading to their application by the authors of this paper to prevent digital crime, namely viewing of CEM online.

2.1. Honeypots

Honeypots represent a lure for bad actors and enable the identification of those intending to engage in prohibited behaviour. Spitzer published the first definition of a honeypot as "a resource who's [sic] value is in being attacked or compromised" [8]. The concept of a honeypot was developed through the late 1980s and early 1990s with the first recorded use (although not using that name) by Clifford Stoll in his widely read book *The Cuckoo's Egg*, which popularised the importance of computer security to a broad audience [12]. Through the 1990s, other authors, and indeed software makers, created tools for producing a honeypot within a network, such as the Deception Toolkit [9].

Simply put, the role of a honeypot is to discover the presence of a bad actor and to distract them away from production systems. Discovery then enables steps to be taken to prevent actions against the production system, increasing their protection. While honeypots do not reduce the number of attacks that occur, they do lower their effectiveness and enable the identity (IP address) and attack methodologies to be discovered. This means that honeypots are not limited to a particular threat but can be deployed in a broad range of contexts [13].

Honeypots present themselves as legitimate objects that are then targeted. They can be created physically, or they can be any part of our digital infrastructure, from servers or routers to entire fake networks [11,13]. They commonly have features that make them of interest to a bad actor, possibly through some clear security flaw or higher availability than typical systems. Ideally, they should be discovered rapidly to enable monitoring systems to discover the attack before they have carried out any actions against a production system.

Schnier describes the purpose of computer security as prevention, detection, and response [14]. A honeypot does not focus on prevention in the first instance, as it is premised on allowing an attack to occur, albeit on a false target. The value of the honeypot in this context lies in detection and enabling an appropriate response to be made. Ideally,

detection and response should occur without the bad actor realising that they were not interacting with a true target. This is quite different to the other forms of defensive cyber security infrastructure described above, which aim to prevent attacks or stop them in process as quickly as possible. This is an infrastructure that can be quite valuable for cyber crime research [15].

Next, this paper describes a honeypot that has been deployed in two studies which aimed to detect and educate naïve internet users who demonstrate an interest in 'deviant' but legal pornography with close parallels to CEM. The next section begins by describing the CEM prevention context, before a description of the honeypot used in this study is provided and key findings are discussed.

2.2. CEM Prevention

The exponential growth of online CEM offending in recent years [16,17] underscores the need for a multifaceted approach to CEM prevention. In this paper, we refer to prevention as a strategy that seeks to dissuade potential or at-risk offenders from choosing to intentionally access CEM online; that is, a form of offender-focused primary and secondary prevention [17]. Until relatively recently, the importance of offender-focused prevention strategies was largely overlooked in favour of toughening the legislative response and increasing the capacity of law enforcement agencies to respond to offending. While the importance of these efforts cannot be overestimated, there is growing recognition that more can, and should, be done to prevent CEM offending.

The success of a number of offender-focused prevention programs operated by charities and other non-government organizations (NGOs) in Europe and the United States underlines this point. Key examples include the *StopItNow!* Program, which is run in a number of European countries and United States [18]; the German program, *Project Dunkelfeld* [19]; and *Thorn*, which is based in the United States [20]. These programs offer a range of services to individuals who may be at risk of offending or are seeking help to stop offending, including anonymous helplines and counselling services. Reports indicate the high demand for these services invariably outstrips the capacity of organizations to respond [21].

While the development of offender-focused prevention programs is largely limited to charities and NGOs, a number of actors within the internet industry and commercial sector have taken action in the broader prevention space. For example, internet search providers, large storage operators, and Internet Service Providers (ISPs) have all played key roles in detecting and removing CEM as well as reporting content and users to law enforcement [2,22]. These actions have resulted in CEM becoming less available on open access areas of the Internet [23].

In addition, attempts have also been made to implement forms of automated prevention, which seek to disrupt the crime-commission process. A good example of this is the trial by Google and Microsoft to use warning messages and a blocking system in an attempt to deter individuals from entering CEM-related search terms into their search engines [22]. This two-pronged strategy appeared to reduce searches for CEM by 67% over a one-year period. However it is not clear whether this effect was due to the warning messages, the blocking system, or a combination of both strategies [22].

The lack of research examining the utility and effectiveness of any form of automated prevention strategy, including the use of warning messages, represents a hurdle to their use [24]. In the following section, we briefly explore the empirical and theoretical basis for the use of such strategies in CEM prevention.

2.3. CEM Education for Deterrence

Despite recognition that the accessibility and availability of CEM on open-access areas of the Internet has declined, due to the combined efforts of law enforcement agencies and the Internet industry, the risk that an individual will be exposed to CEM online remains. Studies indicate that exposure to CEM may occur within a range of online spaces including

a substantial proportion through Peer-to-Peer (P2P) file sharing networks [25] and adult pornography websites.

The prevalence of CEM on P2P networks has been described as 'endemic' [26] (p. 2) with recent estimates suggesting that globally, as many as 3 in 10,000 Internet users may share CEM on P2P networks [27].

Recently, the purported availability of CEM on popular mainstream pornography websites has attracted media attention [28]. The potential for popular genres of ostensibly legal pornography, such as the category of "teen", to represent a stepping stone or gateway to CEM offending has been recognised [29,30]. Moreover, a growing number of qualitative studies with CEM offenders suggest that it is not uncommon for onset—the first deliberate viewing of CEM [31]—to begin following initial inadvertent exposure to CEM while searching for adult pornography [32,33]. Underlining concern, Ray et al. [34] points to the possibility that perhaps as many as 1 in 5 adult pornography users have seen CEM online, although caution is warranted as this is not a prevalence estimate per se.

Explanations of the individual factors that may lead an individual to begin and to continue to view CEM vary. This is unsurprising, given that such offenders tend to have heterogenous demographic profiles [35]. Indeed, it is frequently pointed out that the most striking characteristic of these offenders is their "ordinariness, not [their] deviance" [36] (p. 193). Yet, there is broad agreement within the literature that situational factors, such as opportunity, may well have a greater impact on the likelihood of an individual beginning to viewing CEM than individual factors, including a pre-existing sexual interest in children [37].

A compelling theoretical explanation of this phenomenon comes from the situational crime prevention theory [17,38]. The theory rests on the premise that all criminal decision-making occurs in an interaction between *individual* and *situational* factors. Making this point, Wortley and Smallbone argue that the extraordinary expansion of the CEM market has been primarily fuelled by *situational* factors—namely ready access to digital cameras and the internet—rather than an increase in *individual* factors, like paedophilia. On the internet, users are largely anonymous, have easy access to CEM, and enjoy a comparatively low risk of detection provided they take basic precautions.

Situational crime researchers observe from other data that previously law-abiding individuals can be "drawn into committing specific forms of crime if they regularly encounter easy opportunities for these crimes" [39]. Perceptions of moral ambiguity increase this risk [38]. Studies suggest that cognitive distortions around the harmlessness of viewing CEM may play a role in onset [40,41]. Of broader concern, research examining public perceptions of CEM offending, particularly towards the possession and viewing of CEM, indicate that public awareness of the harms caused by CEM offences may be low [42].

Within a suite of other strategies, automated warning messages have been proposed to tackle some forms of CEM offending, particularly with respect to attempts to search for and view CEM online [43].

3. Methodology

This project aims to establish the extent to which internet warning messages dissuade users from accessing a proxy for CEM; and within this investigation, to establish what kind of messages (whether punitive or harm focused) is most effective. This is a unique study and presents a series of difficulties relating to how to undertake participant recruitment and monitoring in an appropriate manner. This research was approved by the University of Tasmania human research ethics committee (#H0012409). The study was conducted for the public benefit, and it involved a low risk of causing distress to participants, no illegal behaviour was observed, and sizeable lengths were taken to protect the anonymity of the participants [24,44].

Firstly, to study the effectiveness of warning messaging on those who may be about to commit a crime, the data needs to be collected in a realistic situation to ensure natural responses from the participants. In other words, the experiment requires ecological validity.

If the subjects of the study are aware that they are being monitored, then their behaviour will be modified, and the true outcome will not be established. The participants need to believe that the website that the warning messages relate to is a real website, that contains deviant pornographic content as a proxy for CEM (hereafter referred to as CEM-like content), then their responses to the messaging will not be genuine. As such the research study needs to use naïve participants who are unaware of the study being undertaken.

Secondly, a website that is accessible via the internet presents a complex scenario in terms of determining who is visiting the site. When a user visits a web page, they unknowingly present a range of information to the website such as their IP address along with browser and operating system details. Further details relating to demographics of the user are not accessible. Within a study of this nature, it is important to control who is accessing the site within the study, to enable meaningful analysis of the different warning messages to be conducted.

The next section outlines an infrastructure which enables a group of participants to be recruited into a scenario where they are presented with the opportunity to access what appears to be a website containing CEM-like content. If they attempt to interact with the opportunity, they then receive either a warning message or proceed without a warning message, depending on which research condition they have been randomly assigned.

3.1. Honeypot Infrastructure

Perpetrators of digital crimes committed online do so in an environment of increased anonymity compared to crimes committed offline. There is a perceived distance between themselves and those they may be harming. They feel comfortable and safe while they commit the crime. This study, and indeed other warning message initiatives, aim to educate within that situation, to change behaviour, to prevent the crime from occurring. In an online environment, warning messages may be effective in modifying the features of the online environment conducive to crime, such as user perceptions of anonymity. In this study, two honeypot websites were created to then facilitate warning messages to be applied to any internet user who attempted to visit them.

The honeypot websites created within the study did not contain any CEM content. Instead, the two honeypots focused on a closely related category of pornography, called "barely legal" or "teen". This material is legal and widely available [45]. "Barely legal" is a useful proxy for CEM because of its detailed focus on adult–minor sex. For example, techniques used in "barely legal" pornography to enhance the fantasy of an adult actress being a minor include: choosing actresses with small physical statures; clothing (e.g., Catholic school uniforms, pyjamas); child-like behaviour (e.g., giggling, shyness, crying); visual cues (e.g., apparent vaginal bleeding, teddy bears); themes (e.g., storylines involving step-fathers, babysitters, teachers); references to sexual inexperience (e.g., "fresh", "innocent", "virgin"); and the control exerted by male partners [46]. Indeed, Dines [29] refers to "barely legal" as "pseudo-child pornography".

Honeypots derive much of their value from their ability to fool wrongdoers into believing they are authentic. As such, the honeypots used within this study were not just simple websites that were created and placed online for anyone to be able to visit randomly. A platform around them was created to ensure that they were realistic in design and setup to ensure an authentic discovery experience for the participants, and an appropriate experimental environment for measuring warning effectiveness.

The experiment contains three abstracted layers to the infrastructure to ensure appropriate data was gathered from the participants, in as realistic a scenario as possible for a user discovering a website containing CEM material. The three layers are recruitment, obfuscation, and education.

3.2. Recruitment

This study aims to measure the effectiveness of warning messages in deterring participants who are about to commit a crime. The participants are unaware that they were

being monitored. In other words, they were naïve participants within a research setting. This is a truly rare scenario within criminology research. However, the websites that they will visit are on the Internet, and so achieving and maintaining this naivety is the role of the platform.

The first step in this is participant recruitment. The study leveraged social media advertising platforms to gather a targeted cohort of participants. Social media has been used as a segmenting tool for marketing quite successfully since its inception [47]. The true value of social networks is the data that they have on their users and leveraging that knowledge for marketing purposes is their primary business model.

Ads to a benign website were placed on two social networks, targeting Australian men aged between 20 and 30. The benign website is a fitness building website (referred to by the pseudonym 'GetFit' in this paper), tailored to this demographic. This demographic selection is due to men being more likely to be exposed to CEM through their greater use also of P2P networks [31], and men are more likely than females to be pornography users [48], and to have viewed a broader range of pornography [49], like "barely legal". Targeting younger adults will make the results more applicable to CEM-prevention strategies. This is because younger internet users appear more likely than older users to—in situational crime prevention terms—"regularly encounter easy opportunities" to view CEM [5].

Participants who saw the advert on social media and then proceeded to click-through to visit the website fall within the target cohort as judged by the social media platform. Within the research platform, data is collected on their actions on the webpages from that point forward, in addition to their IP address and their visit commenced with a referral from a social media website. No demographics or other information about the participant is collected from the social media platform. They are completely anonymous and unknown to the researchers.

3.3. Obfuscation

Honeypots are only successful if attackers believe that they are authentic. If a honeypot does not appear and respond in an expected manner, it is then obvious that they are false, and interactions with them will cease. Within this study, the honeypot websites are constructed alongside a benign fitness website. The CEM-like websites are presented to users within ad sections within the fitness website, alongside other ads to benign products and services. This tiered approach to the honeypots aims to hide the intentions of the social media ads, and to ensure that the responses to the warning messages when they appear are as authentic as possible to gain a true measure of their effectiveness.

Modern websites are constructed using a sizable infrastructure comprised of services operating on multiple servers around the world. Likewise, the websites we visit are not stored on the same device but are hosted in different locations. While it would be possible to construct the platform used within this study on one server, it would be trivial for a user to notice, and it would then appear strange and obviously false.

The platform within the study involved a fitness website constructed and hosted by a nationally known marketing company. This included content targeted at the demographic of interest to ensure it would be an authentic web experience upon arriving after clicking on the social media ad. Within this website was several ads provided by an advertising service hosted on a server run by the research team. This advertising service delivered a range of ads, including those for the CEM-like honeypot. These were inserted randomly within the fitness website. The warning messages were delivered from same server as the ad-service, but with a different subdomain, referring to a service that provided web filtering services. Finally, the CEM-like honeypots themselves were hosted on a third server, operated by the researchers but at a different location to the ad-service server. The overall architecture of the honeypot is shown in Figure 1.

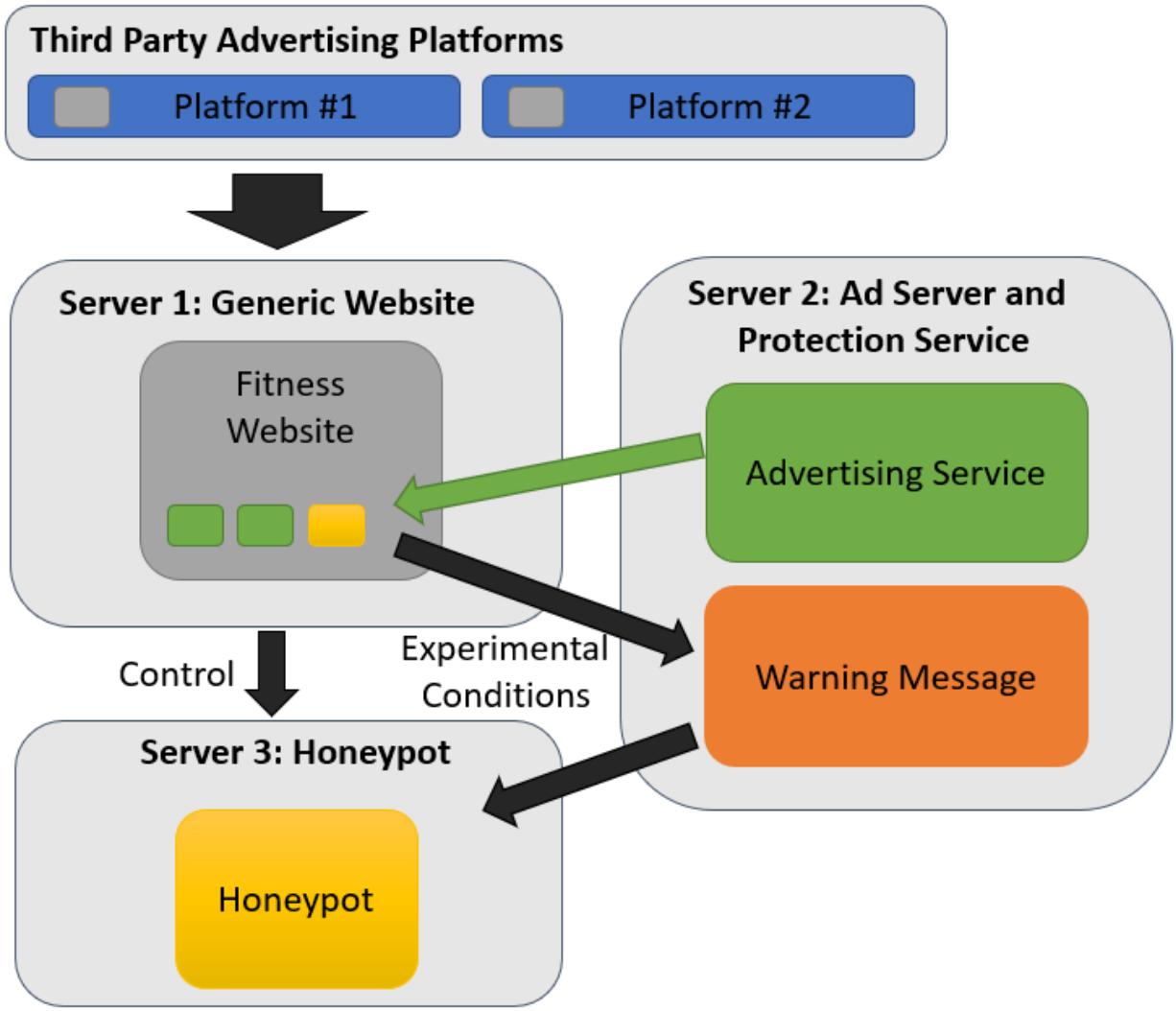

Figure 1. Honey pot server architecture, including third party social media platforms and three servers operated by the research team.

All of the websites, including the CEM-like honeypot, were all protected by Transport Layer Security (TLS), with valid certificates to prove authenticity, even though contact information and business registrations provided to registrars was patently false. TLS is common throughout the web due to pressure from Google to mark websites as insecure without protection [50]. Similarly, malicious websites often use the same protections to appear authentic [51].

3.4. Education

The experiments involved testing the response of participants to warning messages that were displayed to them upon clicking an ad to a CEM-like website. Users who clicked on the link were randomly allocated to a control group (who receive no message), or experimental conditions with different messages about harm and deterrence. Figure 2 displays a screenshot of the website ('GetFit'), and the ad and honeypot website ('JBL') used in one of the two experiments, in addition to a brief description of the experimental conditions.

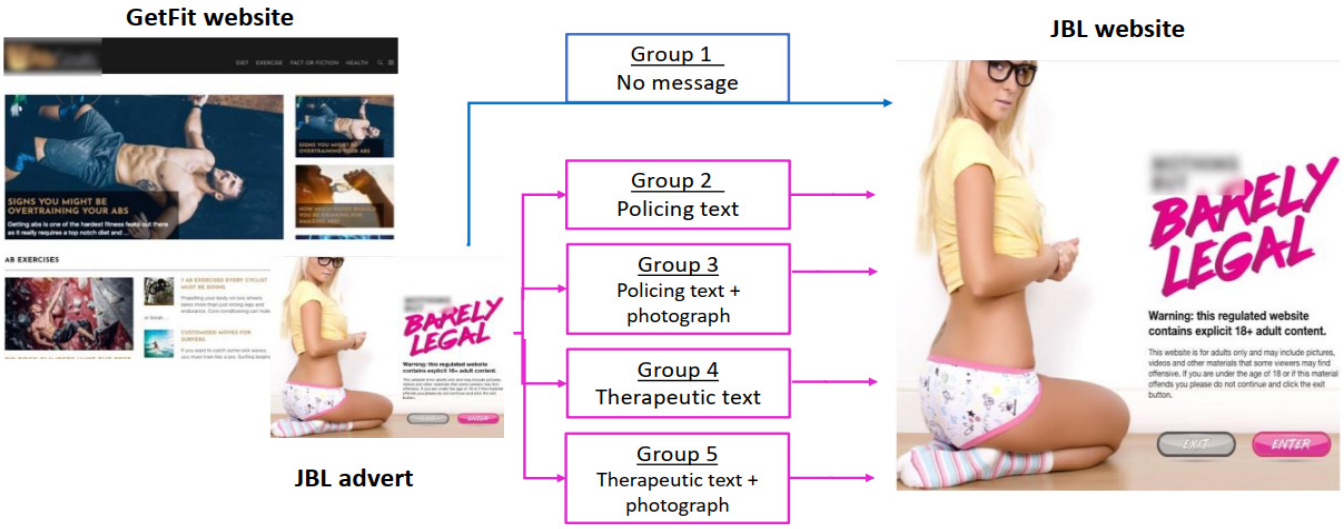

Figure 2. Screenshots of the benign website, GetFit, reached via an advertising campaign on social media and the honeypot ad and website, listing the experiment branches.

When the warning message was displayed, a participant could choose either to click on a 'cancel' button, prompting the message to disappear, or participants would be directed to the entrance to the honeypot. Imagery is present that matches the imagery contained within the ad they have seen. Mimicking other pornography sites, users will have the option of 'exiting' and returning to the previous page or 'entering', which triggers an error message indicating temporary malfunction of the website site. Control participants were directed to the honeypot site without a message and have the 'enter/exit' option. This critical variable records participants' persistence or desistence.

The platform aims to provide an authentic web experience, ensuring that when the participants see the warning message, the educative aspect of the platform, they respond in a natural way. Persisting to view the content, or to opt out, and to leave. The IP addresses are recorded at all three servers: the fitness, warning message, and honeypot. The progression of the participant through the layers of the platform, when compared to the control group without any educative aspect, enables their effectiveness to be measured. The details of the educative elements and the results of the experiment are presented within other publications by the same authors [31,52]. However, examples of the messages are shown in Figure 3 (previous page) of the first experiment, which focused on a 'barely legal' website honeypot. Figure 4 shows two warning messages from the second experiment, whose honeypot was focused on sharing user created material.

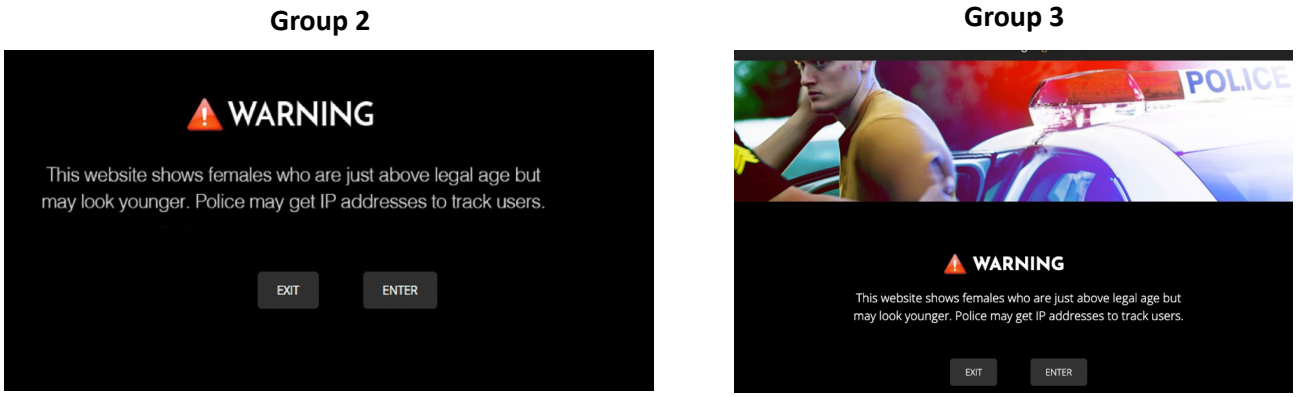

Figure 3. *Cont.*

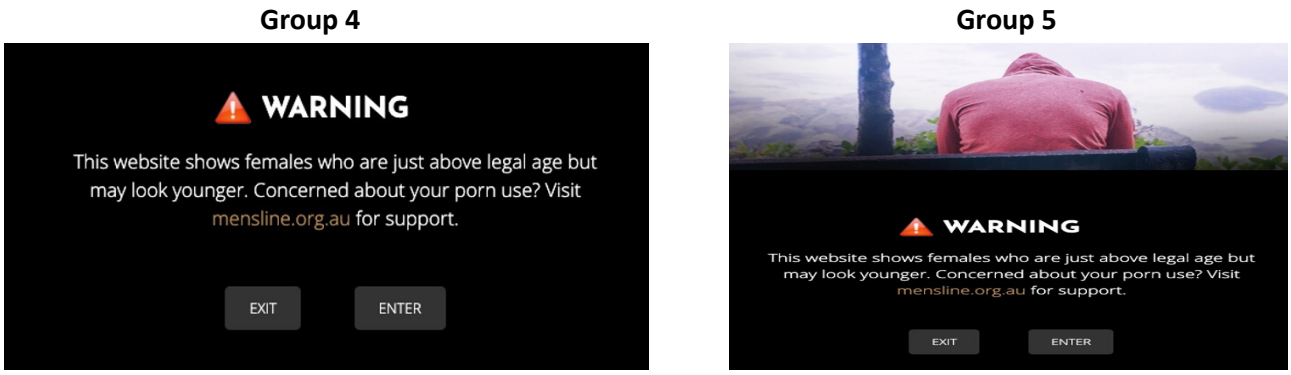

Figure 3. Examples of the messages shown during first experiment with a honeypot styled as the 'barely legal' pornography genre.

Figure 4. Examples of the messages shown during second experiment, which focused on dissuading users from sharing illegal content.

4. Results and Discussion

The research platform produced for the study included paid ads on large social networks and multiple servers operated by the team and are shown in Figure 1. The social networks selected have a global reach, and indeed are 'household names' in their ubiqui-

tousness. These ads buys were targeted at Australian males in the 20–30 age range. The generic website which was contained in these ads was a fitness website targeting the demographic and was built and maintained by a large Australian marketing communications company. This included fake content within the website to ensure it was engaging and believable. It was also discoverable via normal web searches.

Within the fitness website ads were placed by the research team through an ad network operated on a separate server. These ads included health supplements and other fitness products. Users who came from one of the social media networks were also at random shown an ad for the honeypot website. This website, hosted on a third server, contained CEM-like content. When a participant clicked on the ad to the honeypot, they were allocated to either the control or an experimental condition. The control group was forwarded through to the honeypot site directly, while the experimental groups were shown different warning messages. These messages were delivered from a domain hosted on the same server as the ad network. The warning messages shown to users also linked to a website (hosted along with the messages) that described itself as a service that protected users while online. Users could proceed past the warning message to the honeypot. Whether they did so was measured, as was whether they attempted to enter the honeypot website beyond its landing page. These interactions were recorded to measure the effectiveness of the warning message at deterring users from interacting with the CEM-related content. All three servers used within the platform were operated in different locations, with the only connection between them being provided by the ad network.

Using this infrastructure, two experiments were conducted, each with a different final honeypot landing website, but using the same generic fitness website. The relevant ads for the honeypot provided by the advertising service were updated between the two experiments. The advertising campaign run on social media for the fitness website was different between the two experiments, with different imagery and adjusting the campaign to try to get increased engagement from the first experiment. The warning messages were changed as these were the focus of the experiment. The first experiment was operated for four months, while the second operated for seven months.

The first experiment resulted in 1.1 million impressions (views) for the fitness website ads on a large social network by the targeted demographic. These ads resulted in 31,162 clicks from the target demographic through to the fitness website. Of this cohort, 419 saw and then proceeded to click on an ad for the honeypot website, being allocated into the control or experiment conditions.

The second experiment resulted in 3.4 million impressions for the fitness website ads on one social network and 643,000 on a second social network by the targeted demographic. These ads resulted in 27,234 clicks from the target demographic to the fitness website. Of this cohort, 528 then proceeded to click on an ad for the honeypot website, being allocated into the control or experiment conditions.

The ads for the honeypot performed at an order of magnitude or greater than the other ads on the fitness website. These other ads were for health foods and supplements, which while less alluring than the honeypot, are still relevant ads for a fitness focused website.

The performance of the experimental conditions is discussed within other papers by the authors [24,44,52]. However, in summary of those findings, all the warning messages dissuaded users from attempting to enter the honeypot websites. The most effective resulting in 65% of users not proceeding to attempt to enter the website—compared to 27% in the control. The results for each experimental branch in Figure 2 are shown below, detailing how many users still proceeding to enter the honeypot website after seeing a warning message (group 1 is the control):

- Group 1 No message—73%
- Group 2 Policing text—51%
- Group 3 Policing text and image—35%
- Group 4 Therapeutic text—40%
- Group 5 Therapeutic text and image—47%

There was a measurable significant difference between the control group and each condition in experiment one. The most effective message related to the legality of the content and the possible involvement of the police. These styles of messages were then the focus within experiment two (which focused on not sharing illegal content), where similar results were attained. The results indicate that warning messages can dissuade users from viewing or sharing sexual imagery. Full description of this data and its analysis is available in [24,44,52].

However, the performance of the platform as a mechanism for undertaking research, to recruit a cohort of participants displaying behaviours, which could be described as precursor activity to viewing CEM, is clearly demonstrated here. The study was able to recruit and observe 947 participants within two experiments, across a period of 11 months. The platform has enabled the study of what kinds of deterrent messaging could be effective in increasing the awareness and education of the harms of CEM material within a cohort that appears to be at risk of onset (i.e., viewing CEM for the first time). This platform shows promise for future usage of similar tools to be used by NGOs or law enforcement to insert messaging around existing problematic websites.

This research platform has demonstrated the ability to capture, observe, and educate a cohort that would otherwise have been difficult to isolate in a meaningful way. Such techniques could be applied in other contexts where people exhibit behaviours which are harmful to themselves or others. Similar honeypots could be created and applied within the area of drug abuse or gambling to not only educate but also establish the most effective messaging for reaching such troubled cohorts. The approach could be valuable in responding to geopolitical or socioeconomic cybercrimes [3], such as combat hate speech or fraud.

5. Conclusions

This paper described a novel approach to recruiting and interacting with participants in a hard to target cohort involved potentially in digital crime. The platform created included multiple servers containing websites that aimed to give a realistic web browsing experience to users who are displaying an interest in material similar to CEM. The platform was created in such a manner to obfuscate its true purpose, to ensure users were unaware of what was taking place. The platform enabled warning messages to be displayed to the cohort, to educate them to the harms and risks of such content. The intent of the platform was to reach those who were first displaying interest in such content, to educate them, and change their behaviour before it progressed any further. The platform leveraged social network advertising as an intake mechanism to ensure only participants with the target demographics were included in the experiments. The experiments operated over a period of 11 months, recruiting 947 participants, displaying a range of warning messages to educate them to the harm of interacting with CEM content online.

Future work will concentrate on expanding the effectiveness of aesthetic factors in the construction of effective warning messages, as well as identifying new and emerging digital ad markets and social media within which to engage users before they proceed to harmful CEM usage. Cross-site scripting and other technologies, such as chatbots using artificial intelligence, provide a range of future pathways to increase the realism as well as the reflective dialogue with potential offenders.

Author Contributions: Conceptualization, P.A.W., J.P, C.H., C.S. and R.W.; methodology, J.S., P.A.W., J.P., C.H., C.S. and R.W.; software, J.S., P.A.W.; validation, J.S. and P.A.W.; writing—original draft preparation, J.S. and P.A.W.; writing—review and editing, J.P., C.H., C.S. and R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the Australian Research Council (DP160100601) and the Australian Institute of Criminology.

Institutional Review Board Statement: This research was approved by the University of Tasmania human research ethics committee (Reference #H0012409). The full study and its privacy controls

were reviewed by the committee. The study was conducted for the public benefit, and it involved a low risk of causing distress to participants, no illegal behaviour was observed, and significant lengths were taken to protect the anonymity of the participants.

Informed Consent Statement: Informed consent was waived due to the observational nature of the research, and the requirement for the participants to be naïve to the existence of the study taking place, to ensure a natural response to the warning messages. The controls around the anonymization of the participants, in the context of their consent not being received, was central to the ethics application and the controls around data privacy put in place. No data was recorded about their social media account which was active at the time of data collection—only the verification that they were referred from a social media website via the placed advertisement, to ensure that they were within the target group demographically.

Data Availability Statement: The data has not been made available as a part of the privacy controls in place as the participants did not consent to participate, and the dataset includes their IP address.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Australian Federal Police. 2022. Available online: https://www.afp.gov.au/what-we-do/crime-types/cyber-crim (accessed on 9 April 2022).
2. Holt, T.J. Regulating Cybercrime through Law Enforcement and Industry Mechanisms. *Ann. Am. Acad. Political Soc. Sci.* **2018**, *679*, 140–157. [CrossRef]
3. Ibrahim, S. Social and contextual taxonomy of cybercrime: Socioeconomic theory of Nigerian cybercriminals. *Int. J. Law Crime Justice* **2016**, *47*, 44–57. [CrossRef]
4. Ashworth, A.J. *Positive Obligations in Criminal Law*; Bloomsbury Publishing: London, UK, 2013.
5. Wortley, R.; Smallbone, S. *Child Pornography on the Internet: Causes Investigation and Prevention*; Pager: Santa Clara, CA, USA, 2012.
6. Ingham, K.; Forrest, S. A History and Survey of Network Firewalls. Ph.D. Thesis, Department of Computer Science University of New Mexico, Albuquerque, NM, USA, March 2002. Available online: https://www.researchgate.net/publication/228602583 (accessed on 4 April 2022).
7. Denning, D.E. An intrusion-detection model. *IEEE Trans. Softw. Eng.* **1987**, *SE-13*, 222–232. [CrossRef]
8. Spitzner, L. Definitions and Value of Honeypots. 2003. Available online: https://www.eetimes.com/definitions-and-value-of-honeypots/ (accessed on 8 May 2003).
9. Cohen, F. The deception toolkit. *Risks Dig.* **1998**, *19*, 1998.
10. Hunn, T.; Watters, C.; Prichard, P.; Wortley, J.; Scanlan, R.; Spiranovic, J.; Krone, C. Implementing Online Warnings to Prevent CSAM Use: A Technical Overview. 2022; *in press*.
11. Spitzner, L. The Honeynet Project: Trapping the hackers. *IEEE Secur. Priv.* **2003**, *1*, 15–23. [CrossRef]
12. Stoll, C. *The Cuckoo's Egg: Inside the World of Computer Espionage*; Doubleday: New York, NY, USA, 1989.
13. Nawrocki, M.; Wählisch, M.; Schmidt, T.C.; Keil, C.; Schönfelder, J. A survey on honeypot software and data analysis. *arXiv* **2016**, arXiv:1608.06249.
14. Schneier, B. *Secrets and Lies: Digital Security in a Networked World*; Wiley: New York, NY, USA, 2003.
15. Perkins, R.C.; Howell, C.J. *Honeypots for Cybercrime Research BT—Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches*; Lavorgna, A., Holt, T.J., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 233–261.
16. Holt, T.J.; Cale, J.; Leclerc, B.; Drew, J. Assessing the challenges affecting the investigative methods to combat online child exploitation material offenses. *Aggress. Violent Behav.* **2020**, *55*, 101464. [CrossRef]
17. Wortley, R.; Smallbone, S. *Internet Child Pornography: Causes, Investigation, and Prevention*; ABC-CLIO: Santa Clara, CA, USA, 2012.
18. Knack, N.; Winder, B.; Murphy, L.; Fedoroff, J.P. Primary and secondary prevention of child sexual abuse. *Int. Rev. Psychiatry* **2019**, *31*, 181–194. [CrossRef]
19. Beier, K.M.; Grundmann, D.; Kuhle, L.F.; Scherner, G.; Konrad, A.; Amelung, T. The German Dunkelfeld Project: A Pilot Study to Prevent Child Sexual Abuse and the Use of Child Abusive Images. *J. Sex. Med.* **2015**, *12*, 529–542. [CrossRef]
20. Thorn. 2020. Available online: https://www.thorn.org (accessed on 4 April 2022).
21. Assini-Meytin, L.C.; Fix, R.L.; Letourneau, E.J. Child Sexual Abuse: The Need for a Perpetration Prevention Focus. *J. Child Sex. Abus.* **2020**, *29*, 22–40. [CrossRef]
22. Steel, C.M.S. Web-based child pornography: The global impact of deterrence efforts and its consumption on mobile platforms. *Child Abus. Negl.* **2015**, *44*, 150–158. [CrossRef] [PubMed]
23. Westlake, B.G.; Bouchard, M. Liking and hyperlinking: Community detection in online child sexual exploitation networks. *Soc. Sci. Res.* **2016**, *59*, 23–36. [CrossRef] [PubMed]
24. Prichard, J.; Krone, T.; Spiranovic, C.; Watters, P. Transdisciplinary research in virtual space: Can online warning messages reduce engagement with child exploitation material. In *Routledge Handbook of Crime Science*; Routledge: New York, NY, USA, 2018; pp. 309–319.

25. Koontz, L.D. *File Sharing Programs: The Use of Peer-to-Peer Networks to Access Pornography*; US Government Accountability Office: Washington, DC, USA, 2005.
26. Brennan, M.; Hammond, S. A methodology for profiling paraphilic interest in Child Sexual Exploitation Material users on peer-to-peer networks. *J. Sex. Aggress.* **2016**, *23*, 90–103. [CrossRef]
27. Bissias, G.; Levine, B.; Liberatore, M.; Lynn, B.; Moore, J.; Wallach, H.; Wolak, J. Characterization of contact offenders and child exploitation material trafficking on five peer-to-peer networks. *Child Abus. Negl.* **2016**, *52*, 185–199. [CrossRef]
28. Grant, H. World's Biggest Porn Site under Fire over Rape and Abuse Videos. *The Guardian*. 9 March 2020. Available online: https://www.theguardian.com/global-development/2020/mar/09/worlds-biggest-porn-site-under-fire-over-videos-pornhub (accessed on 8 March 2022).
29. Dines, G. Childified women: How the mainstream porn industry sells child pornography to men. In *The Sexualization of Childhood, Westport*; Olfman, S., Ed.; Praeger: Westport, CT, USA, 2009; pp. 121–142.
30. Knack, N.; Holmes, D.; Fedoroff, J.P. Motivational pathways underlying the onset and maintenance of viewing child pornography on the Internet. *Behav. Sci. Law* **2020**, *38*, 100–116. [CrossRef]
31. Prichard, J.; Watters, P.A.; Spiranovic, C. Internet subcultures and pathways to the use of child pornography. *Comput. Law Secur. Rev.* **2011**, *27*, 585–600. [CrossRef]
32. Steely, M.; Bensel, T.T.; Bratton, T.; Lytle, R. All part of the process? A qualitative examination of change in online child pornography behaviors. *Crim. Justice Stud.* **2018**, *31*, 279–296. [CrossRef]
33. Rimer, J.R. Internet sexual offending from an anthropological perspective: Analysing offender perceptions of online spaces. *J. Sex. Aggress.* **2016**, *23*, 33–45. [CrossRef]
34. Ray, J.V.; Kimonis, E.; Seto, M. Correlates and Moderators of Child Pornography Consumption in a Community Sample. *Sex. Abus.* **2013**, *26*, 523–545. [CrossRef]
35. Henshaw, M.; Ogloff, J.; Clough, J. Looking Beyond the Screen: A Critical Review of the Literature on the Online Child Pornography Offender. *Sex. Abus.* **2015**, *29*, 416–445. [CrossRef]
36. Wortley, R. Situational prevention of child abuse in the new technologies. In *Understanding and Preventing Online Sexual Exploitation of Children*; Quayle, E., Ribisl, K., Eds.; Routledge: London, UK, 2013; pp. 188–203.
37. Elliott, I.A.; Beech, A.R.; Mandeville-Norden, R.; Hayes, E. Psychological profiles of Internet sexual offenders: Comparisons with contact sexual offenders. *Sex. Abus.* **2009**, *21*, 76–92. [CrossRef] [PubMed]
38. Cornish, D.; Clarke, R. Opportunities, precipitators and criminal decisions: A reply to Wortley's critique of situational crime prevention. *Crime Prev. Stud.* **2003**, *16*, 41.
39. Clarke, R. Situational crime prevention. In *Environmental Criminology and Crime Analysis*; Wortley, R., Mazerolle, L., Eds.; Willan Publishing: Devon, UK, 2008; pp. 178–195.
40. Merdian, H.; Wilson, N.; Boer, D. Characteristics of Internet Sexual Offenders: A Review. *Sex. Abus. Aust. N. Z.* **2009**, *2*, 34.
41. Quayle, E.; Taylor, M. Paedophiles, Pornography and the Internet: Assessment Issues. *Br. J. Soc. Work* **2002**, *32*, 863–875. [CrossRef]
42. Hunn, C.; Spiranovic, C.; Prichard, J.; Gelb, K. Why internet users' perceptions of viewing child exploitation material matter for prevention policies. *Aust. N. Z. J. Criminol.* **2020**, *53*, 174–193. [CrossRef]
43. Taylor, M.; Quayle, E. Criminogenic qualities of the Internet in the collection and distribution of abuse images of children. *Ir. J. Psychol.* **2008**, *29*, 119–130. [CrossRef]
44. Prichard, R.; Scanlan, J.; Krone, J.; Spiranovic, T.; Watters, C.; Wortley, P. Warning Messages to Prevent Illegal Sharing of Sexualised Images: Results of a Randomised Controlled Experiment. Trends Issues Crime Crim. Justice. 2022. *in press*. Available online: https://discovery.ucl.ac.uk/id/eprint/10144666/ (accessed on 4 April 2022).
45. Jensen, R. A content analysis of youth sexualized language and imagery in adult film packaging, 1995–2007. *J. Child. Media* **2010**, *4*, 371–386. [CrossRef]
46. Peters, E.M.; Morrison, T.; McDermott, D.T.; Bishop, C.J.; Kiss, M. Age is in the Eye of the Beholder: Examining the Cues Employed to Construct the Illusion of Youth in Teen Pornography. *Sex. Cult.* **2013**, *18*, 527–546. [CrossRef]
47. Lorenzo-Romero, C.; Alarcón-del-Amo, M.-C. Segmentation of Users of Social Networking Websites. *Soc. Behav. Personal. Int. J.* **2012**, *40*, 401–414. [CrossRef]
48. Carroll, J.S.; Padilla-Walker, L.M.; Nelson, L.J.; Olson, C.D.; McNamara Barry, C.; Madsen, S.D. Generation xxx pornography acceptance and use among emerging adults. *J. Adolesc. Res.* **2008**, *23*, 6–30. [CrossRef]
49. Svedin, C.G.; Åkerman, I.; Priebe, G. Frequent users of pornography. A population based epidemiological study of Swedish male adolescents. *J. Adolesc.* **2011**, *34*, 779–788. [CrossRef] [PubMed]
50. Google. A Secure Web Is Here to Stay. 2018. Available online: https://security.googleblog.com/2018/02/a-secure-web-is-here-to-stay.html (accessed on 4 April 2022).
51. Volkman, E. *Abuse of HTTPS on Nearly Three-Fourths of all Phishing Sites*. 2020. Available online: https://info.phishlabs.com/blog/abuse-of-https-on-nearly-three-fourths-of-all-phishing-sites (accessed on 8 March 2022).
52. Prichard, J.; Wortley, R.; Watters, P.A.; Spiranovic, C.; Hunn, C.; Krone, T. Effects of Automated Messages on Internet Users Attempting to Access "Barely Legal" Pornography. *Sex. Abus. J. Res. Treat.* **2021**, *34*, 106–124. [CrossRef] [PubMed]