Users' Perceptions of Online Child Abuse Detection Mechanisms

ELMIRA DELDARI, University of Maryland, Baltimore County, United States
PARTH THAKKAR, University of Maryland, Baltimore County, United States
YAXING YAO, Virginia Tech, United States

Child sexual exploitation and abuse (CSEA) online has become a major safety issue for children to access the Internet. To combat CSEA, electronics services providers (ESP) have implemented various mechanisms to detect child sexual abuse materials (CSAM). However, these mechanisms, despite their capability to prevent the mass distribution of CSAM online, may raise significant privacy concerns among general users. In this paper, we conducted a semi-structured interview study with 23 participants to understand their privacy perceptions of two types of online CSAM detection mechanisms. Our results suggested that users were concerned about the transparency of the detection process, inappropriate access to users' data, and unclear boundaries of such mechanisms. Our results also highlight that, even though the majority of participants choose to sacrifice their privacy for societal benefits, they still have privacy concerns that need to be addressed. We discuss the design and policy implications for ESP to improve users' awareness of the data practices of these mechanisms, alleviate users' privacy concerns, and increase societal benefits.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**.

Additional Key Words and Phrases: Online Child Abuse Detection, Child Sexual Abuse Materials, User Perceptions, Privacy, Server-Side, Client-Side

ACM Reference Format:
Elmira Deldari, Parth Thakkar, and Yaxing Yao. 2023. Users' Perceptions of Online Child Abuse Detection Mechanisms. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 147 (April 2023), 26 pages. https://doi.org/10.1145/3637424

1 INTRODUCTION

Child Sexual Abuse and Exploitation (CSEA) is prevalent on today's Internet [25, 39]. For a long time, CSEA has been one of the most significant safety risks for children when they are online [39]. Users with malicious intentions can share different forms of CSEA-related content and materials (e.g., images, text, videos) anywhere on the Internet, such as social media feeds, search engines, photo/video sharing platforms, and streaming platforms [19]. In most cases, offenders share these materials online to execute child exploitation mechanisms; in some other cases, offenders (generally identified as general users) share these materials because they found those materials funny or they felt bad and wanted to spread awareness [79]. While the intention for sharing child sexual abuse materials (CSAM) may not necessarily be malicious, such behaviors are still against the community standards and could even be considered a crime [32]. In this paper, we use "CSEA" to denote the phenomenon of child sexual abuse and exploitation and "CSAM" to denote child sexual abuse materials.

Authors' addresses: Elmira Deldari, University of Maryland, Baltimore County, Baltimore, MD, United States, edeldar1@umbc.edu; Parth Thakkar, University of Maryland, Baltimore County, Baltimore, MD, United States, parthkt1@umbc.edu; Yaxing Yao, Virginia Tech, Blacksburg, VA, United States, yaxing@vt.edu.

ACM 2573-0142/2023/4-ART147
https://doi.org/10.1145/3637424

To combat the sharing and spreading of CSAM, electronic service providers (ESP) have implemented different algorithms or programs (i.e., detection mechanisms) to automatically detect any materials or contents that may be considered inappropriate. Existing CSAM detection mechanisms can be broadly divided into two types (details can be found in Section 2.3), the *server-side mechanism* (i.e., detection is done on ESP's servers) and the *device-side mechanism* (i.e., detection is done on users' devices). For example, Google implemented a server-side mechanism that can scan all the content across all Google service platforms and raise a flag if abusive content is identified [28]. Other companies (e.g., Facebook, Microsoft) also have similar techniques for the same purpose [2, 51]. In 2021, Apple announced its plan to scan for CSAM locally (i.e., a device-side mechanism) through users' devices before they are uploaded to the iCloud [30].

While these detection mechanisms may be designed with good intentions, they have become increasingly controversial because of their ability to scan users' data, causing users significant privacy concerns. For example, when Apple announced its CSAM scanning feature in its operating systems, the public considered it a "privacy disaster" due to its ability to access users' photos, messages, and other documents that were stored locally on users' devices [38]. In fact, due to users' significant privacy concerns, as of January 2022, Apple has removed mentions of their local CSAM scanning from their websites without a clear returning date [74].

In this study, our goal is to explore how users perceive the privacy aspects of CSAM detection mechanisms implemented on devices and servers. We have the following research question: 1) Are users aware of current only CSAM detection mechanisms? 2) What are users' privacy concerns about CSAM detection mechanisms? These are important questions as they shed light on two crucial aspects. First, CSAM detection mechanisms are widely utilized across various online platforms, granting them access to a significant portion of users' data. However, the lack of transparency regarding these mechanisms raises valid privacy concerns among users. Second, since Apple is one of the most popular ESP in the US and the only one implementing device-side CSAM detection mechanisms, these mechanisms are directly applied to users' personal devices where they store personal data like photos and text messages. As a result, users have to make decisions about their actions and privacy without having all the necessary information. Thus, increasing users' awareness of these detection mechanisms and their data practices as well as understanding users' privacy concerns towards these mechanisms may help users make informed decisions and achieve their privacy goals.

To answer our research questions, we conducted 23 semi-structured interviews with US-based participants. To ensure that we include users of both client-side and service mechanisms, all participants were users of both Apple products (e.g., iPhone, MacBook) as representatives of client-side mechanisms and other ESP (e.g., Google and Microsoft) as representatives of server-side mechanisms Our interview focused on how participants perceived the server-side and device-side mechanisms, including the perceived pros and cons of each type of mechanism, their privacy concerns, and their perceived trade-offs between their privacy needs and the social benefits of reducing CSAM spreading online. Our results indicate that most participants were not aware of these detection mechanisms. After going through the introduction of selected mechanisms, they raised many privacy concerns (e.g., accessing an excessive amount of data) and identified some potentially malicious purposes of data collection (e.g., ESP collect users' data for marketing purposes in the name of scanning CSAM). Participants also compared the two types of implementation based on their perceived effectiveness and privacy invasiveness. Finally, while most of our participants are willing to sacrifice their privacy needs to some extent to reduce CSAM distribution online, some participants still believed that social good should not come at the cost of users' privacy.

Our paper contributes to the field in three ways. First, to the best of our knowledge, previous studies primarily focused on the technical aspects of CSAM detection mechanisms. Our study is the

first qualitative research that explores users' awareness of these mechanisms being implemented on their personal devices. Additionally, we investigate users' privacy perceptions of CSAM detection mechanisms. Our findings shed light on various privacy concerns expressed by users, as well as their willingness (or unwillingness) to make trade-offs between privacy and societal benefits. Second, we discussed potential enhancements for future CSAM detection mechanisms. By identifying areas for improvement, we contribute to the ongoing development of more effective and privacy-conscious detection techniques. Lastly, our paper draws design and policy implications based on our findings.

2 BACKGROUND

In this section, we introduce some background knowledge related to CSAM detection mechanisms.

2.1 NCMEC: Central Source of CSAM

In the US, the National Center for Missing and Exploited Children (NCMEC) leads the fight against abduction, abuse, and exploitation. As the nation's centralized reporting system for online child abuse and exploitation, NCMEC's CyberTipline system provides an opportunity for both public and Electronic Service Providers (ESP) to report any incidents or materials related to CSEA [54].

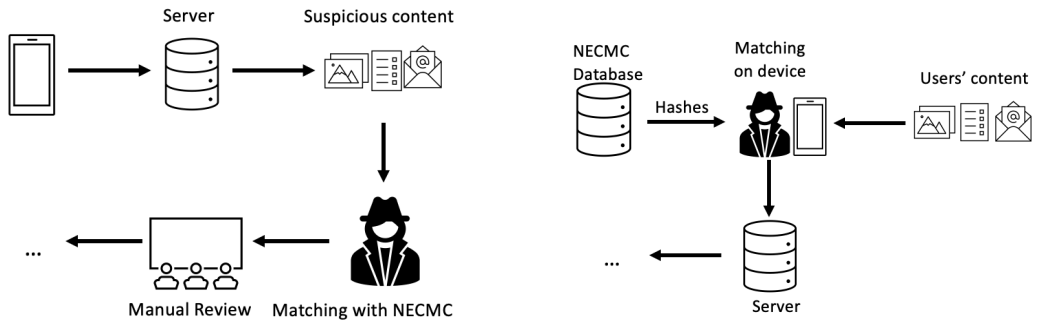
In 2021, more than 29.3 million of these reports were received by CyberTipline, including close to 70 million child sexual abuse images and videos [28]. 29.1 million of these reports were from ESP, reporting instances of apparent child sexual abuse material that they identified on their systems [54]. These reports include online enticement of children for sexual acts, child sexual molestation, child sexual abuse material, child sex tourism, child sex trafficking, unsolicited obscene materials sent to a child, misleading domain names, and misleading words or digital images on the Internet. In particular, ESP are obligated to report all instances of online child exploitation content to the NCMEC upon reasonable suspicion and take action to take down the content (potentially the associated accounts in some cases).

2.2 Hash-Matching Technique

In the US, ESP generally have dedicated mechanisms on their respective platforms to identify CSAM. Most of them use a hash-match technique.In general, the hash-matching database uses the known CSAM image hashes provided by NCMEC and other child-safety organizations. The hashing algorithms analyze an image and convert it to a unique number specific to that image. Only another image that appears nearly identical can produce the same number. For example, images with different sizes or transcoded quality will still have the same hash value [7]. In practice, each ESP may implement its hash-matching algorithms in different ways. We broadly categorize these mechanisms into two types, i.e., the *server-end* mechanism, and the *device-end* mechanism. It should also be noted that both mechanisms only work on non-end-to-end encryption (E2EE) data, as it is nearly impossible for ESP to scan E2EE data [14] Next, we introduce how each mechanism works.

2.3 Two types of CSAM detection mechanisms

Server-Side Mechanisms. In a typical server-side mechanism, the detection algorithm is applied to the data stored on the ESP's servers. Most ESP, such as Google [27], Meta [2], Microsoft [51], have chosen this mechanism for their CSAM detection. Even though each company may have developed its own version detection technologies that best fit its product line, they all follow a similar approach. Figure 1a shows an overview of a typical server-side mechanism. As shown in the figure, after users upload their data to the ESP's server, a pool with suspicious content is identified, typically through users' reporting, crawlers, or other pre-filters (other classifiers). Then, each ESP runs its own detection algorithms to match the suspicious content with a national CSAM



(a) Illustration of the server-side mechanism.

(b) Illustration of the device-side mechanism.

Fig. 1. Overview of two CSAM detection mechanisms.

database provided by NCMEC. When a match is found, the content will then be reviewed manually for further confirmation. If the content is confirmed to be CSAM or related, ESP may take further actions including removing content from the platform, immediately banning the user, and reporting the case to NCMEC.

Device-side Mechanism Unlike server-side mechanisms, the detection algorithms will run on users' devices. As of now, Apple is the only ESP that plans to adopt the device-side mechanism [8] in all of its operating systems. Figure 1b shows an overview of Apple's device-side mechanism. In this process, Apple first converts the dataset from NCMEC into a non-readable format and stores it on users' devices as part of the operating system. Then, the detection algorithm will match the content (e.g., photos, videos, messages) on users' devices with the hashed dataset before the content is uploaded to Apple's server (i.e., iCloud). Any matched results will be uploaded to iCloud in the form of a Safety Voucher. When the count of Safety Vouchers reaches a certain threshold (i.e., 30 images), Apple will further review the instances, inform the users about the instance, remove the relevant content, and file a report to NCMEC [47].

3 RELATED WORK

3.1 Children's Online Safety and Countermeasures

Research studies have shown that CSEA can negatively impact children's mental and physical health [43]. It can cause a child to experience isolation, fear, anxiety, and distrust, which may further lead to lifelong psychological consequences such as academic difficulties, low self-esteem, depression, and difficulty forming and sustaining relationships [13, 18, 36]. Unfortunately, the rapid development and adoption of electronic devices have made CSEA extremely prevalent online [15, 19, 63, 66]. Typically, online CSEA may happen in the form of uploading inappropriate images/videos of child pornography [31, 42], online grooming [29], sex chatting [4, 46], and cyber-bullying [5, 57, 75].

To ensure a safe online environment for children, numerous attempts have been made to combat online CSEA. One approach is to equip children and their parents with safety advice and knowledge, such as computer security advice [55], privacy setting suggestions [9, 19, 75], and self-education program [59]). Parents may also adopt various technology-mediated tools (e.g., Net Nanny[1], Minor Monitor[3]) to monitor their children's online activities to ensure their safety.

In recent years, another approach to combat online CSEA has become more and more prevalent, i.e., to detect and identify predators, offenders, and any content (e.g., images, videos, text) that may be relevant to CSEA using techniques from computer science. For example, Safer, an AI-powered tool, allows technology platforms to identify, remove, and report CSAM at scale [67]; ChildSafe.ai

can actively collect signals of exploitation threats from online ecosystems, then model the child exploitation risk on the web [17]; Griffeye and CEASE.ai uses computer vision tools (e.g., facial recognition, image recognition) to scan images to evaluate their nudity and ages [26]; Spotlight uses predictive analytics to identify potential victims of human and child trafficking and child sexual abuse [64].

Despite the good intentions behind these tools and practices, since all techniques mentioned above rely on access to users' data (e.g., photos, videos, messages, and documents), they may raise significant privacy concerns among users. Yet, how the general public perceives these practices, particularly from the privacy aspect, remains understudied. Our study attempts to fill the gap by investigating how users perceive two different implementations of the "hash-matching" CSAM detection mechanisms, i.e., the server-side mechanism and the device-side mechanism.

3.2 CSAM Detection Mechanism

In the past, there were early approaches that relied on specially designed image descriptors to identify nudity and minimized the need for manual examination. For example, the NuDetective Forensic Tool [20] focuses on enabling forensic examiners to analyze images efficiently at crime scenes by automatically detecting nudity in images. Similarly, the iCOP toolkit [56] is designed to identify individuals who share new or unfamiliar child sexual abuse media on peer-to-peer (P2P) networks.

Modern approaches typically make use of advanced deep-learning techniques. For instance, the work by [48] introduced their own methods for estimating the age of individuals and the detection of pornography. Macedo et al. created a dataset that includes images with child pornography content. This dataset was labeled by computer forensic experts and also contains other types of images such as those without people, images with adults, images with children, and adult pornographic images. They also suggested a method that combines a child face detection module with a pornography detector to identify child pornography. Sae-Bae et al. [60] introduced a method that utilizes skin color filters to identify human skin tones and explicit content. They incorporated a child classification aspect into their work by considering texture and facial landmark distances to differentiate between adult and child nudity.

A more recent study [23] breaks down the problem of automatically detecting child sexual abuse into two more manageable tasks; identifying pornographic content and classifying individuals into age groups. These enhancements are utilized to develop distinct models for age estimation and pornography detection. Gangwar et al. [23] also noticed a shortage of publicly available datasets for age estimation and pornography detection. To address this, they created two datasets: Juvenile-80k, containing 80,000 labeled images of children up to 18 years old, and Pornographic-2M, which consists of 2 million diverse pornographic images collected from the internet.

Yiallourou et al. [80] created a dataset of synthetic images containing potentially suspicious content. They developed a method using Haar feature-based cascade classifiers [44, 70] to detect faces in the images and analyzed them to estimate the age, gender, and lighting intensity. They examined the possibility of determining the level of image inappropriateness by using automatically extracted image features instead of relying on manually annotated features. By combining these steps, they derived five image features: child presence, number of people, age diversity, gender distribution, and lighting. These features were used to train a regression model that classifies the images as appropriate, neutral, or inappropriate. While a significant amount of research in CSAM detection has concentrated on engineering features, there has been relatively less investigation into how these mechanisms are perceived by general users.

3.3 Users' Privacy Perceptions of Technologies

The usable privacy community has long been investigating users' privacy perceptions of different technologies. Prior work has studied how users perceive camera-based technologies [21, 33, 34, 83], the Internet of Things devices [11, 45, 53, 77], social media and online activities [52, 69, 78], smart environments [10, 40, 41, 72, 76, 81, 84], etc. Aside from users' various privacy concerns around these technologies, one important and consistent conclusion indicates a fact, i.e., users constantly make privacy trade-offs. For example, people may have privacy concerns, but they may still choose to join a store loyalty program with their personal information in exchange for discounts [6]; choose to turn on web tracking to receive personalized results [50]; or choose to pay a premium in exchange for their privacy [68].

In the context of CSAM detection, one can easily see the benefits of such practices from a societal perspective, i.e., protecting the online safety of the next generation. Geierhaas et al. conducted a survey study with 1062 participants in Germany and found that people were generally supportive of technological solutions for CSAM, including both client-side and server-side detection mechanisms, to combat online child abuse [24]. However, it is not clear how the broader social good may influence users' privacy perceptions and whether users are ready to make the trade-off between privacy and societal gain. In this study, we situate in the CSAM context and aim to understand how users perceive such trade-offs.

4 METHODOLOGY

We conducted a semi-structured interview study virtually to investigate people's perceptions of CSAM detection mechanisms between November and December 2021. In this section, we describe our methodology in detail. This study is approved by our University IRB.

To investigate users' perceptions of CSAM detection mechanisms, we selected a collection of five ESP that have adopted either the server-side or the device-side implementation in their detection mechanisms. For the server-side implementation, we selected Google, Meta, Microsoft, and YouTube, because they remain the top four platforms in terms of CSAM incidents based on the report from NCMEC [54]. It should be noted that even though YouTube is part of Google, it is ranked separately due to the nature of the reported incidents. That is, the reported incidents are mostly images and documents on Google products, while on YouTube, the incidents are primarily videos. For the device-side mechanism, Apple remained the only ESP that implemented it.

4.1 Screening Survey

We recruited participants for our study using Prolific, an online crowdsourcing platform. Our study focused on understanding users' perceptions and experiences of child online safety mechanisms. Initially, we administered a screening survey on Prolific, which was completed by a total of 109 individuals. After filtering out incomplete or invalid responses, we obtained 78 valid responses. Subsequently, we directly contacted these participants using their Prolific ID and invited them to participate in the next phase of our study. Since we aimed to explore users' perceptions of both implementations and allow for possible comparisons between them, we specifically sought participants who had been exposed to mechanisms from both Apple and one of the following platforms: Google, Meta, Microsoft, or YouTube. Hence, the screening survey was designed to identify participants who had experiences with services from both types of platforms.

We started the screening survey by asking participants about their experiences of using ESP, including their three most frequently used ESP. If a participant's answer included Apple and one of the other four ESP (i.e., Microsoft, Google), they qualified for our study. We reached out to all qualified respondents to schedule a follow-up online interview.

4.2 Interview protocol

The interview protocol consists of the following main parts. The complete interview protocol can be found in the appendix.

Demographic and background questions. We started the interview by asking participants their demographic questions, including their ages, self-identified gender, occupation, and education level. We then confirmed with them regarding their most frequently used ESP, and selected two ESP to use throughout the rest of the interview, one server-side mechanism (i.e., one of Google, Meta, Microsoft, and Youtube) and one device-side mechanism (i.e., Apple). We randomly picked one mechanism and continued the interview as described below. Once we finished one mechanism, we moved forward with the other mechanism using the same procedure.

General awareness of online CSEA and CSAM detection mechanisms. The first set of questions focuses on understanding participants' general awareness and perceptions of online CSEA and CSAM detection mechanisms through the two selected ESP. For example, we asked, "How do you define child exploitation and abuse?""Do you think child abuse exists in the online world?""Are you aware of mechanisms carried out by companies like [ESP] to fight against child abuse on their platforms?"

Perceptions and Understandings of CSAM detection mechanisms. Regardless of participants' prior knowledge of CSEA and these mechanisms, we provided an introduction of the mechanisms to ensure all participants have a similar understanding. We prepared a slide deck with details of how each mechanism works. To maintain the neutrality of our introduction and reduce any potential bias, we purposely adopted the introduction and explanation directly from ESP's websites [2, 8, 27, 51]. Participants first went through the slide deck and read the materials, then answered a set of questions regarding their first impression and whether they understood the materials. Then, we offered a chance for participants to ask any questions they may have about these materials, particularly the technical terms. The goal was to make sure that they had a correct understanding of the terms included in the introduction materials.

Then, we asked participants their perceived concerns and benefits of the mechanism, what factors contribute to their perceptions, and how they would like to address their concerns. Additionally, we also asked participants whether they would be willing to trade their personal information for any perceived societal benefits.

At the end of the interview, we allowed each participant to ask us any questions related to online CSEA, CSAM detection, and other confusions they may have.

4.3 Participant recruitment

We selected participants who had successfully completed the screening survey and provided responses to all the questions. Then, we individually contacted each participant to ascertain their availability for a one-hour interview. Participants need to meet the following criteria to qualify for the follow-up interview: 1) they were at least 18 years old; 2) they lived in the US; 3) they used both Apple devices and at least one service from the four ESPs. All survey participants were paid $0.5 upon finishing the screening survey regardless of their qualifications.

We released our tasks on Prolific in batches of 3 throughout the days of the week. That is, we recruited 3 crowd workers each time to fill in the screening survey. By doing this, we hope to eliminate any sample bias caused by crowd workers' working schedules (e.g., weekday vs. weekend, morning vs. evening) and their locations (e.g., east coast vs. west coast).

After selecting suitable participants, we provided them with a Calendly link to choose a convenient time for the interview. Additionally, we shared a physical copy of the consent form with the participants. Subsequently, we invited all eligible participants to take part in a one-hour follow-up

interview. We stopped our recruitment when reaching saturation. In total, we interviewed 23 participants, and each participant was paid $15.

4.4 Data collection and analysis

Upon participants' consent, we recorded the interviews using Zoom. Participants had the option to turn off their cameras during the interview, although it was not required. As such, we only focused on the audio data. We used the auto-transcribing feature in Zoom to generate the initial transcriptions. Two researchers then manually cleaned and checked all transcriptions.

We conducted thematic analysis [12] to qualitatively analyze the data using the following steps. First, three researchers randomly selected two transcriptions as samples and conducted open coding on all samples independently at the sentence level. Then, they discussed and reconciled their codes to resolve any differences and developed an initial codebook. Using this codebook, they coded another randomly selected sample independently, then compared and discussed their codes to ensure consistency.

Next, two researchers divided the rest of the transcripts. During the process, the two researchers regularly checked each others' coding and discussed as needed to ensure consistency. New codes were added to the codebook as needed based on the agreement between the two researchers. In the meantime, the third research supervised all these activities by checking at least half of the coding for each transcription to ensure a high-level agreement. The final codebook contains over 100 unique codes, such as "excessive data collection", "incorrect report", and "unspecified consequences".

When finished coding, all researchers discussed and grouped all the codes into high-level themes. When we had conflicting opinions in grouping the codes, we checked our dataset to determine an optimal group for these codes. In total, we identified 7 high-level themes.

Given the qualitative nature of the study, we refrained from reporting the exact number of participants when presenting the frequency of participants' responses. Instead, we adopted a consistent terminology (similar to prior work [22, 82]) to present the relative sense of such frequency, as illustrated in Fig. 2.

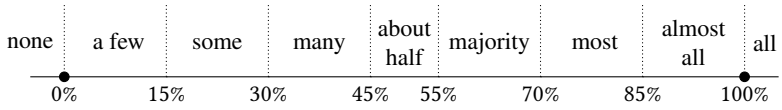


Fig. 2. Terminology used to present relative frequency of themes.

4.5 Ethics

Since our study focused on children's online safety, we prioritized research ethics by implementing rigorous measures throughout the project. First, we provided all participants with a consent form that explained the study's procedures. We requested participants to carefully review the form and return a signed copy via email and prior to the start of each interview, we additionally obtained oral consent from participants. We further emphasized to all participants that they had the right to withdraw from the study at any time and they were under no obligation to answer any questions that made them uncomfortable. In presenting our data, we have made efforts to ensure the anonymity of the participants. The narratives and experiences we share are representative of common themes and are not meant to depict any specific individual's personal circumstances.

4.6 Limitations

Our research has several limitations that should be considered. Firstly, all the participants in this study were from the US. As privacy perceptions may vary in different cultural backgrounds, we can not generalize these results to other countries. Second, due to Apple being the only ESP that implemented device-side mechanisms, our participants were limited to individuals who have used Apple products or services. This could introduce bias into our sample. To mitigate this concern, we carefully recruited participants who actively use various server-side mechanisms such as Google and Microsoft in their daily lives besides Apple products. Additionally, we tried to recruit participants with diverse characteristics, including age, gender, occupation, education, and other relevant factors, further aiding in mitigating potential bias. Finally, while some participants demonstrated a good level of technological knowledge, the majority of our participants were general users of ESPs. This may impact the depth of insight regarding the technical aspect of device-side mechanisms.

5 RESULTS

5.1 Participant

The ages of our 23 participants range from 18 to 71 (Mean = 37, SD = 15.72). Ten participants were male, eleven were female, and two were non-binary. Our participants represent a wide range of locations, occupations, and marital statuses. In terms of participants' technology usage, all participants have used Apple products (e.g., iPhone, iPad, Macbook). For the most frequently used ESP, seven participants selected Google, seven selected Meta, five selected YouTube, and four selected Microsoft. Table 1 summarizes participants' key demographic information.

All participants in our study were aware of online child abuse. Most participants believed that online CSEA exists on social media platforms, text messages, online chats, messaging apps, online gaming platforms, and even emails. They also believed that online CSEA could happen across any devices connected to the Internet, such as computers, tablets, and mobile phones. About half of the participants were able to name four specific types of CSEA, including child pornography, emotional abuse, cyberbullying, and online grooming. About half of the participants have personally encountered online CSEA through Meta, TikTok, and Xbox. Their experiences were mostly about parent monetization, online grooming, and cyberbullying. Some participants have never come across online CSEA, thus believing that online child abuse is moderately severe.

By correctly identifying a few types of online CSEA, almost all participants demonstrated a good level of understanding regarding online CSEA. In addition, the majority of the participants were aware of some ways to report online CSEA. For example, some participants mentioned that they should be able to report suspicious CSEA cases by flagging offensive videos on social media platforms for violating community guidelines. A few participants stated that when seeing any CSAM, they would immediately inform the local Social Service Department or similar agencies for help. Some participants did not know where to report, but they were confident that they could find places to report suspicious cases by searching on Google.

Most participants have not heard of CSAM detection mechanisms. Before we presented the online CSAM detection mechanism, we asked participants whether they were aware of any practices or mechanisms on their ESP to detect potential CSAM content online. Most participants indicated that they have not heard of these mechanisms. Some participants have heard of the detection mechanism from Apple and a few have heard of the one from Google, mostly through technology news. They did not have any experience nor knew the details.

5.2 General Perceptions of CSAM Detection Mechanisms

To ensure all participants have a similar understanding of these mechanisms, we asked participants to read the descriptions provided by their assigned ESP. We then asked participants about their

ID	Age	Gender	Education	Have Kids?	Most Used Platform	Awareness
P1	71	Female	Bachelor	Yes,1	Google, Apple	No
P2	29	Non-binary	Associate	No	Meta, Apple	Yes
P3	45	Female	Bachelor	Yes, 5	Microsoft, Apple	No
P4	20	Female	Master	No	Google, Apple	Yes
P5	19	Female	College	No	Meta, Apple	No
P6	22	Female	Bachelor	No	Google, Apple	No
P7	22	Female	Associate	Yes, 2	Google, Apple	No
P8	28	Non-binary	Bachelor	No	YouTube, Apple	No
P9	27	Male	Master	No	YouTube, Apple	No
P10	45	Male	Bachelor	No	Google, Apple	Yes
P11	22	Male	Bachelor	No	Meta, Apple	Yes
P12	22	Female	Bachelor	No	YouTube, Apple	No
P13	47	Male	Bachelor	No	YouTube, Apple	No
P14	60	Male	High school	Yes,2	Meta, Apple	No
P15	24	Male	Bachelor	No	Microsoft, Apple	No
P16	18	Male	Student	No	YouTube, Apple	Yes
P17	37	Male	Associate	Yes, 2	Meta, Apple	No
P18	28	Female	Bachelor	Yes, 1	Google, Apple	No
P19	48	Female	PhD	Yes, 2	Microsoft, Apple	No
P20	43	Male	Bachelor	Yes, 3	Meta, Apple	No
P21	69	Male	Bachelor	yes, 3	Google, Apple	No
P22	49	Female	College	Yes, 1	Meta, Apple	No
P23	61	Male	Master	Yes, 2	Microsoft, Apple	No

Table 1. Participants' demographic and background information. The "Awareness?" column refers to whether the participant was aware of any child abuse detection mechanism at the time of the study.

first impressions of these mechanisms. Most participants showed positive attitudes towards both types of mechanisms when they initially heard about them. All participants believed that such mechanisms have some clear advantages. In the meantime, however, they also discussed many concerns about these mechanisms. We present the details below.

5.2.1 Benefits of CSAM detection mechanisms were clear. Our participants' overall positive attitude toward the detection mechanisms was rooted in their various perceived benefits. All participants believed that these mechanisms were important steps to help mitigate online CSEA. Specifically, these mechanisms could help stop the dissemination of CSAM content, such as images, videos, audio, and documents. In some cases, these mechanisms could also help track down potential predators by identifying the individuals who uploaded the content. As a result, our participants, especially those who have kids, felt safer when they or their kids were online. Finally, some participants also believed that by having these mechanisms, ESP, such as Google and Meta, may end up with a positive company image and an increased level of trust from the public.

5.2.2 Most participants expressed concerns about CSAM detection mechanisms. In addition to these perceived benefits, our participants also discussed two general concerns related to both server-side and device-side mechanisms, namely, 1) uncertainty about the effectiveness and 2) inappropriate flagging.

Most participants were unsure about the effectiveness. Even though all participants had a positive attitude toward the detection mechanisms, **the majority of them** were unsure about their effectiveness in detecting CSAM. Many participants believed that these mechanisms were moderately effective as there were no appropriate statistics to analyze the performance of these mechanisms. Some participants also believed that, if offenders realized that their routine was blocked, they would potentially seek other alternatives to bypass the detection mechanisms (e.g., offenders may switch their operating systems to bypass the security features on their current computer systems). P16 commented on the effectiveness of Apple's mechanism,

"I'm not sure how effective that would be because you can just switch to Samsung or a different phone, or use your computer for it instead. So, I feel it's just going to catch a small number of people, maybe someone really stupid about it. Like, you can easily get around it if it's just one company." (P16, male, 18)

As P16 stated, if the detection only happened through one company or a few companies, offenders would be able to easily bypass the detection by switching to other devices, making the detection less effective.

Some participants were concerned about inappropriate flagging. Another issue reported by some participants was the possibility of inappropriate flagging. They were concerned that these approaches would inaccurately identify non-abusive content, resulting in undesired legal ramifications. For example, P9 mentioned that if a mother took a picture of her baby while bathing her and sent it to the child's grandmother, it could be flagged as an abusive case. She further explained,

"I'm not sure, I think it was called photoDNA. I would think that maybe the algorithm wouldn't be 100% accurate. So it could detect the false alarm, sort of like false positives." (P9, male, 27)

P9's quote highlighted his concerns over the accuracy of the detection mechanism itself. However, it should be noted that what P9 discussed suggested a clear misunderstanding of how the detection mechanisms work, as the detection mechanisms would match the photos with the flagged ones in the NCMEC database rather than directly checking if the image contained CSAM. Such misunderstanding further indicated that, even after reviewing the introduction on the ESP's website, P9 still lacked an adequate understanding of how the mechanisms work. As P9 later confirmed, her inadequate understanding contributed to her concerns about inappropriate flagging, which further reduced her trust in the mechanisms.

Explanations of CSAM detection mechanisms were difficult to understand for most participants. In fact, P9's example above points to another concern. Most participants either did not understand or only partially understood the explanations provided on ESP's website due to the technical jargon (e.g., "hashes", "PSI") and lack of sufficient details. The high-level explanations without sufficient details also left some participants confused. For example, P21 explained his confusion and suggested some alternative terms,

" There's a lot of technical language, I would have maybe understood better if they could put it in layman's terms, I do have some understanding of how hashing. I'm not a part of an expert on it so if they can put it in when made, they're like fully explained it but also makes it easier to understand the be appreciative." (P16, male, 18)

Overall, the participants voiced positive feedback after learning about detection mechanisms and they believed these mechanisms could help in identifying CSAM content and fighting against CSAE. They also expressed some general concerns regarding the effectiveness and accuracy of these mechanisms, as well as some difficulties in understanding the explanations of the mechanisms.

5.3 Privacy Concerns of CSAM Detection Mechanisms

Aside from the general concerns, discussions around privacy concerns have been salient throughout the study. Participants have expressed various privacy concerns about both server-side and device-side mechanisms.

5.3.1 Possible excessive data access of CSAM detection mechanisms is concerning. One of the most significant concerns our participants have was the amount of data that the detection mechanisms could access. Most participants were concerned that the mechanisms could access not only the necessary information (e.g., photos, videos) but also other types of personal data, such as emails, personal documents, phone numbers, and financial information. It should be noted that the majority of the participants were more concerned about the data access in the device-side mechanism compared to the server-side one. This is because mobile devices generally contain more personal information, thus are more sensitive. For example, P2 spoke on this point when discussing the device-side mechanism. In her opinion, the detection mechanism by Apple could access all data on her phone, including personal documents, personal information, and financial information,

"I think it can access your current and past photos, anything you've shared, and any kind of social media you might have in your contacts. I think it can have my personal number saved. I think it can access your address and all your email accounts. And a lot of people use Apple's service to log into mobile apps for their banks. So even those are potentially at risk in all their passwords and credit cards." (P2, non-binary, 29)

It is worth noting that this was a misconception due to two reasons. First, Apple's implementation of CSAM detection only scans photos before they are uploaded to iCloud. If users opted out of iCloud service, their photos would not be scanned. Additionally, as explained in Section 2.3, all detection mechanisms use the hash-matching technique and match users' photos with the flagged ones in NCMEC's database. Such a misunderstanding indicated the lack of communication and transparency on the data practices of CSAM detection mechanisms, which we will present next.

5.3.2 ESP did not provide transparency on CSAM detection mechanisms and their data practices. Many participants, including some for the device-side mechanism and a few for the server-end mechanism, believed that ESP were not transparent regarding the use of the detection mechanisms in two ways. On the one hand, ESP were not being transparent about the existence of these mechanisms. In fact, many participants were very surprised when they learned about the existence of the mechanisms, even though some of the mechanisms have been implemented and remained active for over a decade (e.g., Google [28]). On the other hand, the data practices of these detection mechanisms remained opaque to the users. During the study, participants reviewed the introduction of their assigned detection mechanisms provided by ESP. However, the majority of the participants reported that they could somewhat understand the explanations, but were not able to parse any details regarding how their data was handled. P10 explained,

"With all sorts of privacy issues that Apple and other tech corporations have, I think they should be more transparent about the detection and I think the government should force them to be more transparent with what they do with our data." (P10, male, 45)

As stated above, P10 believed that ESP needed to be transparent about the practices, particularly given all the privacy issues associated with these ESP in the past few years. He also believed that policy enforcement should also be in place to ensure transparency.

5.3.3 Unclear boundaries of CSAM detection mechanisms. Another key issue brought up by some participants concerned the potential unconstrained capability of CSAM detection mechanisms due to the lack of clear boundaries for these mechanisms. They mentioned that compared to identifying

CSAM, using such detection mechanisms to identify other content could be a more significant threat to their privacy. If the detection mechanisms were capable of identifying CSAM content, they could be used to recognize any other subject in users' data. P6 illustrated this point when asking about the cons of the mechanisms,

"I guess it's kind of two-fold. On one hand, if they have the ability to use this technology, why does this problem still exist online? So, what positive effect it is having on child exploitation issues? And then on the other side, I would wonder, if they could do this, what other kinds of images can they decode? I guess it just makes me think, what other information does Apple know about us through our photos?" (P6, female,22)

In her response, she not only questioned the effectiveness of these mechanisms but also raised further concerns about the power as well as the boundary of such mechanisms. P3 also contributed a similar opinion. When we asked P3 opinions on the differences between the server-side and device-side mechanisms, she commented,

"I think they're pretty much the same ones, same privacy issues and access to information. I know that you said it doesn't take place on the phone, it takes place on Microsoft servers, but for a lot of people, me included, that's an abstract notion. They are still privacy-related. To me, it's unlocking that door. Once you unlock that door, you throw away the key, you can't really lock it. It's just open. So what's going to come through in the future?" (P3, female, 45)

P3's comment indicated her concerns about the potential unlimited possibility for these detection mechanisms to be used without any boundary. Such possibilities introduced significant uncertainties and deepened our participants' concerns about their privacy and data security.

5.3.4 CSAM detection as an excuse for data collection and monetization. Aside from the above privacy concerns, some participants also raised their concerns that ESP may use CSAM detection mechanisms as an excuse to legitimately collect users' data. A few participants mentioned that ESP may collect additional data in the name of ensuring children's online safety, and then use the data for other purposes (e.g., monetization) at their will with fewer restrictions. P10 commented,

"I think it's just a way for these companies to get more information under the guise of doing something for the public good. They're scanning every photo that's being uploaded to iCloud and they're doing it obviously for a good cause, but that doesn't prevent them from doing it for other issues in the same way... like a photo of drinking or something like, they can be sharing that with an insurance company. It's just a breach of privacy. I think they have the unwarranted right." (P10, male, 45)

In his view, regardless of the positive societal benefits, he believed that ESP may collect and share other types of data during the CSAM detection process and cause negative consequences on his life, and he considered the "unwarranted" data collection a "breach of privacy". While it was not clear whether this statement was true or not, we believed that participants' concerns remained valid, which, once again, pointed to the lack of communication and transparency regarding how CSAM detection mechanisms were implemented and how users' data was handled.

5.3.5 CSAM detection mechanisms would not impact ordinary people. While most participants have indicated various types of privacy concerns, a few participants were optimistic that there would not be privacy issues associated with these mechanisms. They believed that these mechanisms would only impact those who had malicious purposes or have contributed to the distribution of CSAM. For ordinary users who have nothing to hide, they should not get concerned. Such perception echoed the notion of optimism bias [62] and has been observed in other technological contexts as well (e.g., social media [71], smart homes [10], etc.). For example, P17 acknowledged that while other users may have privacy concerns, he would not worry about it,

"I think privacy is going to be a big issue for a lot of people about coding pictures that could go wrong, and the system could malfunction. So, it's still an invasion of privacy and I'm sure, a lot of people's minds. I don't mind. Just because I know I did nothing wrong, I wouldn't worry about it." (P17, male, 37)

In summary, our participants discussed various types of privacy concerns. Among them, we highlight the concern related to surveillance, as it was one of the core controversial issues that were raised along with Apple's child safety features [58].

5.4 Comparison Between Server-side and Device-side Mechanisms

In the interview, we asked participants to compare the server-side and device-side mechanisms and whether they, as users, would prefer to see one type of mechanism over the other during their technology usage. Our results suggested that about half of the participants preferred the server-side mechanism while a few participants preferred to see device-side mechanisms. In addition, a few participants did not have a preference while some participants preferred to not have any CSAM detection mechanism at all. We further identified two main factors that influenced their preferences, i.e., perceived effectiveness and the level of perceived invasiveness. We present the details below.

5.4.1 Perceived effectiveness. When comparing the two types of mechanisms, our participants expressed different opinions on which one was more effective. Their opinion further influenced their preferences. About half of the participants believed that, compared to the device-side mechanism, the server-side mechanism would be more effective, because it may be able to provide better coverage for CSAM detection and stronger protection, as it could access all users' data on the server.

Additionally, some participants highlighted that having CSAM detection on users' devices could lead to potential discrepancies and create more issues related to sharing, downloading, and receiving inappropriate content. They argued that focusing on the server side would be more effective in targeting the individuals involved in sharing and accessing such content. For instance P5 said,

" I think the server side is much better. I think that Apple's way of being on people's devices kind of opens the door for a lot of discrepancies. More so, not that maybe we shouldn't be looking into every single person's device. I think it would cut down much better on these issues occurring if we just go from the server side and focus on that first whole bunch of this implementation practice." (P5, female, 19)

This viewpoint highlighted the idea that when the detection process was centralized on the server side, it allowed for comprehensive monitoring and analysis of all uploaded content. By scanning content at the server level, CSAM could be identified more effectively, helping to prevent its spread within the platform.

However, a few participants acknowledged that device-side detection mechanisms also have their merits. By scanning content directly on users' devices, privacy could be enhanced, and potential risks of transmitting sensitive material to servers could be mitigated. These participants expressed concerns that once images or documents left users' phones and were uploaded to servers (such as on YouTube), they became publicly accessible data that could harm children even before being detected and reported. They believed that a device-side mechanism would guarantee that all CSAM could be identified and reported on users' devices without anyone else seeing it, leading to more effective detection. P7 expressed her opinions to illustrate this point,

"What attracted me toward the Apple one was how technical it was, in my opinion. Maybe more advanced because it's on the device, so it's like within the phone. On YouTube, the offender would have already uploaded to YouTube, so the algorithm is only cycling through a smaller group of whatever

data is being uploaded. But on your phone, it's getting all the images videos, texts, and anything that can be detrimental to children." (P7, female, 22)

P7's argument highlighted the need to tackle CSAM at its origin, specifically on users' personal devices. Device-side detection focused on stopping CSAM from being shared or seen by others. It also allowed more localized and immediate action, ensuring that harmful material was identified and dealt with promptly, which ultimately improved the effectiveness of the detection process.

It should be noted that among those participants who did not have a preference for the two types of mechanisms, they primarily considered their overall effectiveness. In their mind, regardless of where these mechanisms were implemented, they would accept the mechanisms as long as they remained effective in identifying potential offenders or CSAM content and protecting the children. For example, P14 stated,

"It sounds better to have it away from me on the server but it probably doesn't make any difference since they are functioning in the same way. So, I don't have a preference, though." (P14, male, 60)

He further explained that both server-side and device-side mechanisms should be able to perform detection on a large number of contents to fight against offenders and stop online CSEA. Given the same purpose of both mechanisms, he would accept both of them.

5.4.2 Level of perceived invasiveness. Participants' preference for server-side CSAM detection mechanisms was significantly influenced by their perception of privacy invasiveness. Our findings revealed that approximately half of the participants believed the server-side mechanism to be less intrusive because it solely examined content voluntarily shared on the platform. They noted that platforms such as Meta and Google have limited access to personal information, primarily encompassing text messages and photos. In contrast, direct device-side detection raised greater privacy concerns as it encompasses all activities conducted on the device. P5 even suggested that utilizing Google or Meta for CSAM detection might entail fewer privacy risks compared to device-side mechanisms. These observations highlighted participants' concerns about privacy and their preference for mechanisms that limit the scope of content inspection to voluntary platform sharing, thus reducing potential privacy violations.

One interesting aspect that P15 touched on is the notion of "privacy trade-off". As he later explained, even though the server-side mechanism of Meta was less invasive of his privacy, there was still a trade-off between his privacy and the societal benefits of preventing the distribution of CSAM. We will further unpack the results related to this point in the next section.

It is also worth noting that some participants who preferred not to have any detection mechanism also discussed their significant privacy concerns. They believed that these mechanisms were a direct violation of their privacy. In the following example, P16 mentioned that since he has been doing legitimate things, the only thing that was sacrificed by having these detection mechanisms was his privacy. In this case, P16 also considered a trade-off between his privacy and any potential benefits,

"Personally I would prefer no detection at all. I know I'm not doing anything that I shouldn't be doing, so for me, it's all just like the aspect of giving up privacy, but I don't really get any of the benefits." (P16, male, 18)

In summary, the effectiveness and privacy invasiveness of the detection mechanisms were two important factors that impacted participants' preferences. Our results showed that more than half of the participants preferred the server-side mechanism while only three of them preferred the device-side one. The reason was that the server-side mechanism, in general, has fewer privacy trade-offs for the users.

5.5 Trade-off between Privacy and Preventing CSAM Distribution

As discussed in prior sections, while all participants understood the benefits of these mechanisms, they also expressed their privacy concerns. Yet, some participants suggested that they would sacrifice their privacy in exchange for a safer online space for children, indicating a potential "trade-off" between users' privacy and the societal benefits of preventing the distribution of online CSAM. In this section, we present participants' perceptions of such trade-offs.

5.5.1 Willing to sacrifice privacy for societal benefits. Majority of our participants suggested that sacrificing some privacy in exchange for preventing CSAM distribution was reasonable, and the need for a safe online environment for children often triumphed over their own privacy. Thus, they were in complete agreement with the trade-off. For example, P20 explicitly highlighted that saving one person's life or preventing one child from being abused was far more important to him than his privacy, thus he would be willing to make a trade-off to ensure the success of these mechanisms. He said,

"I'm okay with my images being put through the database because I would want a program like this to succeed. If that saves or helps any number of exploited children in the grand scheme, if it can help people, I'm okay with my privacy being affected." (P20, male, 43)

5.5.2 Additional information was needed to understand the CSAM detection mechanisms better. Some participants were willing to trade off their privacy for societal benefits but only to some extent. They required further information to gain a thorough understanding of CSAM detection mechanisms. For instance, P2 wanted to see more empirical evidence regarding how companies fighting against child abuse by implementing those detection mechanisms,

"I don't know, again, yeah there's a hard one, I'd want to hear specific information specifically how it was used to combat child abuse, child trafficking, and exploitation I want empirical evidence. And if she could present that, then I would be fine. But I would also want clarity as to who's privacy to these photos and videos." (P2, non-binary, 29)

5.5.3 The lost privacy has positive social values. A few participants were willing to sacrifice their privacy but had a different rationale. For example, P13 explained,

"I think it comes with the territory of using devices and services, they're going to be, for the most part at this point, unregulated by the government and they can sell and trade our information any way they want to. That's already taking place, and I'm supportive and pleased that they're at least doing the right thing, at least protecting children." (P13, male, 47)

In his opinion, given that he already lost control of his personal data through using all the devices and services provided by ESP, it would be comforting to know that he was contributing to protection children online. The positive social value after participants believed that their privacy has gone motivated them to accept the implementation of CSAM detection mechanisms, even though they may still have privacy concerns.

5.5.4 Sacrificing privacy happened too often. A few participants explicitly declined to make a trade-off because similar social benefits have been invoked frequently when their privacy was at risk. As such, they were not willing to sacrifice their privacy for another social benefit cause. For example, P8 specifically did not trust the accuracy and effectiveness of these detection mechanisms. They believed that increasing the accuracy and effectiveness of these detection mechanisms would not be ESP's top priority as ESP was more interested in gaining financial profits using the data. They further commented,

"No trade-off, never. I am pretty concerned about anybody having access to my data. I think that stuff is a real problem. But, it's obviously challenging because it's like everyone can recognize child

abuse is bad. That doesn't mean I want anybody to have access to everything that I own already." (P8, non-binary, 28)

In the following example, P10 held a similar opinion and was not willing to make the trade-off either. He considered privacy as a type of social good and privacy loss as something riskier than online CSEA to individual users. Interestingly, he also believed that law enforcement agencies, instead of private companies, should be the driving force of combating online CSEA. He commented,

"I would say overall, the loss of privacy represents a greater risk to humanity than some people uploading pictures of nude kids. I mean the police should be able to handle that sort of stuff and that should not be left up to a private company. It's where we're heading down a slippery slope here with technology and everything that they're collecting, and that is definitely worse for the collective good. Privacy is the social good." (P10, male, 45)

In summary, even though the majority of our participants were willing to make a trade-off between their privacy needs and social gain, there were still concerns regarding the legitimacy of prioritizing CSAM detection over individual users' privacy, as privacy was still considered to be paramount in their lives.

5.6 Communicating Existing Detection Mechanisms with Users

Although we did not ask questions directly related to enhancing communicating existing detection mechanisms with users, our analysis revealed some promising opportunities. Most participants were interested in learning more about these mechanisms, ideally with greater details compared to what is currently available on the ESP's websites. For example, P11 was one of the participants who were aware of CSAM detection mechanisms before the study. He heard of them from some news articles but did not know too many details. After obtaining more details of the mechanisms through the interview, P11 shared a refreshed viewpoint of these mechanisms and showed a relatively accurate interpretation, indicating the possibility of increasing users' accurate understanding of these mechanisms by providing adequate and detailed information. He noted,

"I think it was more secure and less invasive than I imagined it to be. My previous understanding of this is Apple's actually going to parse through images, and essentially predict or try to detect if there's something that relates to child abuse. But the presentation you just showed me, essentially it's more of a matching than an image analysis from scratch. So, it definitely changed my understanding a little bit about how it works." (P11, male, 22)

Relatedly, as briefly mentioned before, participants also identified several places that they were not able to fully grasp in the mechanism introduction on ESP's websites, mostly technical terminologies (e.g., some participants had a difficult time understanding the concept of hashing). In addition, about half of the participants were also interested in receiving notifications from ESP when CSAM detection existed rather than having to search for the desired information by themselves. They also mentioned the possibility of delivering notifications through various channels, such as emails, text, or even a short animation video.

6 DISCUSSION

The adoption of CSAM detections by various ESP has evolved over time. Early approaches used image descriptions to find nudity. However, with the advancements in technology, modern approaches now employ sophisticated deep-learning techniques and can even scan images directly on users' devices. While previous studies focused on the technical side of CSAM detection, our interview study looked at users' perceptions of these practices. Although the topic of CSAM detection may seem complex and require technical expertise to fully grasp the terminology, we believe it is important to understand the opinions of regular users, especially since device-side scanning

directly affects their privacy on their own devices. In this study, we specifically focused on two types of CSAM detection mechanisms, server-side and device-side mechanisms. Our results showed that a majority of the participants expressed concerns regarding CSAM detection practices. Some of the common apprehensions included issues like inappropriate flagging, lack of transparency, and potential invasion of their data. Participants also noted similarities and disparities among different detection mechanisms and expressed their preferences among two types of detection mechanisms. Furthermore, we found that participants held varying degrees of willingness to sacrifice certain aspects of privacy in exchange for the broader societal benefits of preventing online CSAM distribution.

In this section, we first extend our results by discussing participants' (mis)understandings of CSAM detection mechanisms and how those understandings impacted their privacy concerns. We then discuss the tradeoff between privacy and societal benefits as well as the unclear privacy implications of these detection mechanisms among users. We then draw design implications to enhance the communication between ESP and users.

6.1 Navigating Misunderstandings and Privacy Concerns

Our study revealed that a subset of participants in our research had misunderstandings regarding the mechanisms of CSAM detection practices. These individuals expressed concerns that both device-side and server-side detection mechanisms could potentially invade their privacy and gain access to their personal data. The controversy surrounding Apple's announcement of their device-side detection mechanism exemplifies this issue. Many people believed that Apple would have unfettered access to all content on their personal devices, creating a significant public outcry. In our study, a small number of participants shared similar sentiments, believing that the device-side mechanism could access all of their photos, files, and text messages.

Although Apple provided a detailed document explaining the device-side detection mechanism, it failed to establish complete trust among users. The document did not effectively address users' concerns about privacy implications associated with the detection mechanism. Furthermore, it remained unclear to users whether the underlying algorithms utilized in the detection process served any other purposes. The technical nature of the document also posed difficulties for general users in comprehending the intricacies of the mechanism.

Drawing upon the principles outlined by Saltzer et al. [61] for protection mechanisms in content scanning systems, we suggest the application of the Psychological-acceptability principle by ESP. This principle emphasizes that the policy interface of CSAM detection mechanisms should align with users' mental models of the system. In other words, users must be able to comprehend the mechanics of the protection measures for them to use them correctly. Considering these principles, ESP should design their CSAM detection mechanisms to have restricted scopes, including limitations on the type and format of materials that can be scanned. Additionally, the mechanisms should define which components of memory are accessible to the scanning process and clearly outline who sets the targets and receives alerts. By adhering to these principles, ESP can foster user trust and ensure that users have a clear understanding of how they are protected and how their actions may impact their level of protection.

One prevalent misunderstanding among participants pertains to concerns regarding false positive detection, where participants worried that innocent content, such as a photo of their children in a non-exploitative context, may be flagged as CSAM. It is important to clarify that the hash-matching technique used in CSAM detection matches users' photos against a database of flagged images provided by organizations like the NCMEC. Consequently, not all photos containing nudity would be falsely flagged as CSAM.

To mitigate the risk of false positives, it is essential to design CSAM detection mechanisms that exhibit a low rate of both false negatives (failing to identify traffickers) and false positives (incorrectly flagging users). It is crucial for law enforcement agencies to meaningfully address this concern. Moreover, ESP must ensure that the scanning technology does not exacerbate existing disparities in law enforcement, such as those based on race, ethnicity, class, religion, or gender. It is imperative that errors in the decision-making process do not disproportionately affect minority communities.

Although complete elimination of false positive detection may not be feasible, ESP should communicate with users that efforts are made to keep false positives at a reasonable level. For instance, some ESP like Google [28] incorporate human verification processes after a photo has been flagged, providing an additional layer of scrutiny and reducing the likelihood of false positives.

6.2 Which Is More Important, Privacy or Preventing CSAM Distribution?

In the privacy literature, it is not uncommon that users are willing to make a tradeoff between privacy and other perceived benefits. For example, in the context of smart homes, smart home users are willing to make the tradeoff between privacy and the convenience brought by smart home devices [40, 76, 81]. Some users also believe that online behavior advertising can be smart and useful at times, so they may still choose to make the privacy-utility tradeoff [69]. In our study, it was widely acknowledged among our participants that having such detection mechanisms on ESP's platforms was extremely important to ensure a safe and healthy online environment for the next generation. It helped to protect children from predators who used communication tools to recruit and exploit them and, eventually, have clear social benefits. As a result, the majority of our participants believed that the perceived societal benefit triumphed over their privacy, and thus were willing to sacrifice their privacy in exchange for the large social good.

However, participants' willingness did not suggest that they did not have privacy concerns. Our results suggested that participants had various privacy concerns regarding these detection mechanisms, including their suspicion that ESP may use their data for other purposes in the name of CSAM detection. As such, even though there was a clear societal motivation that drove the detection mechanisms, we believe that ESP should take users' privacy concerns and needs into account and strive to seek a balance between users' privacy and the social good at large.

One potential direction to help achieve this balance is to clearly communicate the privacy implications of CSAM detection mechanisms with users. Our results indicated that the privacy implications of the current CSAM detection mechanisms were not clear to the users, which added another layer of complexity to users' privacy concerns. We further unpack this point below.

6.3 Unclear Privacy Implications of the detection mechanisms

Our participants discussed the pros and cons of each mechanism from two perspectives, i.e., their perceived effectiveness and perceived invasiveness, although the opinions differed across individuals based on their mental models of how the practices work. In reality, both mechanisms utilize the same national database for their match algorithms (i.e., from NCMEC). Additionally, both mechanisms take place before users' data is encrypted (i.e., non-E2EE data). As a result, we believe that both mechanisms should have comparable capabilities and be equally effective in identifying CSAM incidents.

It should also be noted that applying CSAM detection mechanisms to users' data does not provide ESP with greater data access than they already have. Most of our participants, however, seemed to be concerned about the excessive data access by these practices. This phenomenon may be caused by the fact that the process of the detection mechanisms, both the service-end and device-end

mechanisms, remain opaque to the public. It is unclear to the users what data might be used for CSAM detection purposes, how their data might be shared (e.g., with law enforcement agencies) and (mis)used (e.g., for purposes other than CSAM detection, for surveillance purposes), and who has access to their data. As a result, the privacy implication of both mechanisms remains unclear. In fact, Apple announced the suite of child safety features for iOS 15 in August 2021. However, due to the criticisms and controversies raised around the privacy implications of these features (CSAM detection in particular), Apple made changes to their plans, did not release this feature in iOS 15.2, and removed all the mentions related to CSAM detection from their website [74].

6.4 User Expectations and Policy Measures for Privacy Protection

Our study highlighted participants' emphasis on the importance of transparency, simplifying consumer choices, and ensuring privacy-protective measures in the development and implementation of systems. Understanding user expectations regarding privacy and the measures taken by policymakers is crucial in safeguarding privacy. The Federal Trade Commission (FTC) has played a pivotal role in consumer privacy protection since the 1970s, enforcing the Fair Credit Reporting Act (FCRA) [65], one of the earliest federal privacy laws. Over the years, the FTC has pursued consumer privacy protection through law enforcement, policy initiatives, and consumer and business education. For instance, the FTC has employed two primary models in recent years: the "notice-and-choice model" and the "harm-based model." The notice-and-choice model encourages companies to provide privacy notices that inform consumers about their information collection and usage practices. This empowers consumers to make informed choices regarding their personal data. The harm-based model focuses on protecting consumers from specific harms such as physical security risks, economic losses, and unwanted intrusions into their daily lives.

However, the existing privacy model, known as the notice-and-choice model, has limitations. Privacy policies have become lengthier, more complex, and difficult for consumers to understand. Many policies prioritize limiting companies' liability rather than informing consumers about how their information will be used. Additionally, while some companies disclose their data practices, few provide consumers with meaningful controls over those practices. Consequently, consumers face challenges in comprehending privacy policies and exercising control over their personal data. To address these concerns, it is important to enhance transparency and rethink the notice-and-choice model. Privacy policies should be concise, clear, and easily understood by consumers. They should provide individuals with meaningful choices about how their information is used. By striking a balance between legal requirements and user-friendliness, users can be empowered to make informed decisions about their privacy. In the following section, we will delve into these aspects in greater detail.

6.5 Design Implications

Achieving auditability. The design of CSAM detection mechanisms by ESP should prioritize auditability to ensure accountability and build trust. They should enable users to understand which content is scanned, flagged as targeted, and shared with authorities. Additionally, the detection mechanisms should be auditable, allowing for scrutiny of the specific content targeted by the scanning technology. This requires the auditability of target images used to create hash lists and the training data utilized to enhance neural network models, ensuring knowledgeable parties can assess and potentially challenge the system if needed.

Increasing the transparency of the data practice in CSAM detection mechanisms. One major concern raised by our participants is that these detection mechanisms may be used to match dissident content, which is partially caused by the opaque data practices of these practices. Most ESP explain the process of their CSAM detection mechanisms from a high level without zooming

into the details. We suggest ESP increase the transparency of the CSAM detection process with a focus on the data practices, such as what data can the practices access, which parties can access their data, who can participate in the manual review process, and whether their data is used for purposes other than identifying CSAM incidents. More importantly, such information should be presented to the users in an easy-to-understand format. One concrete idea is to adopt the privacy nutrition label [37] and prioritize information that is important to users in the context of CSAM detection mechanisms (e.g., data access, data sharing, etc.).

Providing easy-to-understand documentation for CSAM detection mechanisms. Our results indicated the possibility that knowing how CSAM detection mechanisms work can help alleviate users' privacy concerns. However, most participants reported that the current explanations do not convey the message very clearly. We thus suggest that ESP should consider using alternative media forms to enhance users' understanding. For example, videos or other visual explanatory illustrations have been used to explain difficult terms and are deemed to be effective (e.g., video explanation of the "Right against Solely Automated Decision Making" under GDPR [35], interactive visual explanation of differential privacy [73]), thus can potentially be used in this context. Another concrete idea is to develop an empathy-based approach where ESP allow users to experience the detection process through different personas (similar to [16]).

Consumer education. We believe that there is a need for increased education in this field particularly due to users' limited familiarity with the detection practice mechanisms. To address this need, it would be beneficial to update existing education resources or enhance user education and awareness regarding CSAM detection mechanisms and their privacy implications. For instance, ESP could offer online resources to educate consumers, with a specific focus on children, parents, and teachers, about online safety. These resources would aim to raise awareness about potential online threats such as online predators and phishing. In addition to educating about general online safety, ESP could provide information about CSAM detection mechanisms. This would include details about the types of information being checked, the process involved in detecting CSAM, and the extent of access to user data. It is important to communicate the potential consequences if someone is flagged for engaging in malicious behaviors, as well as the measures in place to address false positives. By doing so, we can alleviate certain concerns and misconceptions that users may have while fostering a more informed and balanced public discussion on this subject.

Communicating the CSAM detection outcomes with users. One key factor that impacts users' privacy preferences and perceptions of both server-side and device-side practices is their perceived effectiveness of these practices. While we believe that both mechanisms should be equally effective, users generally do not have the knowledge to make the judgment. We suggest that all ESP should include a dedicated section on their websites and present some "facts", including how many CSAM cases the practices have identified by the practices and how many are reported to NCMEC, what types of contents these cases come from, how many cases were confirmed by NCMEC to be related to CSAM, etc. This information may convey a positive confirmation of users' dedication and encourage them to continue to help make the Internet safer for children.

6.6 Policy Implications

Refining ESP's privacy policies. The four ESP in our study have all implemented CSAM detection mechanisms, yet none of them have specified the data practices in their privacy policies based on our investigation. Privacy policies have their drawbacks [49], but they remain one of the few channels for users to understand ESP's data practices. We urge that ESP should update their privacy policies and have a dedicated section to explain its company policies around CSAM detection mechanisms. This is not only to inform users of how their data is used for CSAM detection purposes but also to

raise users' awareness of the existence of CSAM detection mechanisms, as our results indicate that most participants were not aware of such practices.

Disseminating knowledge about existing policies. It should be noted that there are several existing laws and policies that have been used to guide ESP on CSAM detection, such as the Eliminating Abusive and Rampant Neglect of Interactive Technologies Act (EARN IT Act) and the Kids Online Safety Act. Yet, our findings suggest that they remain unknown to the general public. ESP are uniquely positioned to disseminate knowledge about existing laws and policies to the general public as a way to raise their awareness of CSEA more generally and encourage everyone to take action.

6.7 Future work

In our study, we have observed some interesting phenomena, yet we cannot draw a definite conclusion from them due to the qualitative and exploratory nature of this study. For example, we notice that depending on whether participants have kids or not or their gender, participants may have different perceptions, particularly towards their willingness to make the trade-off between privacy and social good. Future work can further examine these factors and their impact on users' perceptions.

In addition, future work can look into the users' perceptions and the privacy implications of other types of detection mechanisms, as presented in Section 3.1 of look into how to support users' needs to stay informed of these practices through design. We consider this topic to be of utmost importance and we believe that continuing this study is crucial for a deeper understanding. It would also be valuable to explore the perspectives of other stakeholders, such as service providers, law enforcement agencies, or policymakers.

Finally, as our study focuses on the US population, future research should include participants from other countries. Geierhass et al. studied similar questions in the German context [24]. Their results showed both similarities and differences compared to ours. For example, participants from both countries all supported such technical solutions for CSAM detection. However, German participants were generally supportive of both server-side and client-side mechanisms, yet more than half of our participants from the US preferred server-side mechanisms and only three of them preferred device-side ones. Future research should examine consumers' perceptions of CSAM detection mechanisms in different countries as well as the differences among different societal and cultural backgrounds.

7 CONCLUSION

In this paper, we shared the initial findings from our study with 23 participants who were regular users of electronic services. We aim to understand their privacy perceptions of two types of mechanisms (i.e., a server-side implementation and a device-side implementation) that are used to detect online child sexual abuse materials. Our findings indicate users' various privacy concerns toward these detection mechanisms, such as excessive data access, lack of transparency, and lack of boundary of detection. We also explore the privacy trade-off between users' privacy needs and the societal benefit of preventing the distribution of CSAM. Lastly, we draw design and policy implications for future tool design to combat online child abuse with users' privacy in mind.

8 ACKNOWLEDGEMENT

We thank the reviewers for their thoughtful and constructive feedback. This research is partially supported by the National Science Foundation (Grant 2341187, 2328183), a Meta research award, and a Google research award.

REFERENCES

[1] 2019. Child safety. http://www.netnanny.com/
[2] 2019. Facebook matching. https://about.fb.com/news/2019/08/open-source-photo-video-matching/
[3] 2019. Infoglide, Minormonitor—Facebook Monitoring and Parental Control Software, accessed Jan. http://www.minormonitor.com
[4] 2020. *Guidelines for parents and educators on Child Online Protection.*
[5] 2021. *PARENTING IN THE DIGITAL AGE Parental guidance for the online protection of children from sexual exploitation and sexual abuse.*
[6] Alessandro Acquisti and Jens Grossklags. 2005. Privacy and rationality in individual decision making. *IEEE security & privacy* 3, 1 (2005), 26–33.
[7] Apple. 2021. *CSAM Detection Technical Summary.* https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf
[8] Apple. 2022. Child safety. https://www.apple.com/child-safety/
[9] UNESCO Arnaldo. 2001. *Child abuse on the Internet: Ending the silence.* Berghahn Books.
[10] Natã Miccael Barbosa, Joon S Park, Yaxing Yao, and Yang Wang. 2019. " What if?" Predicting Individual Users' Smart Home Privacy Preferences and Their Changes. *Proc. Priv. Enhancing Technol.* 2019, 4 (2019), 211–231.
[11] Cara Bloom, Joshua Tan, Javed Ramjohn, and Lujo Bauer. 2017. Self-driving cars and data collection: Privacy perceptions of networked autonomous vehicles. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017).* 357–375.
[12] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).
[13] Angela Brown and David Finkelhor. 1986. Impact of Child Sexual Abuse. A Review of the Research. *Psychological Bulletin* (1986).
[14] Andy Burrows. 2021. Why we need to reset the debate on end-to-end encryption to protect children. https://www.computerweekly.com/opinion/Why-we-need-to-reset-the-debate-on-end-to-end-encryption-to-protect-children
[15] John Carr. 2003. *Child abuse, child pornography and the internet.* NCH London.
[16] Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-jun Li. 2023. An Empathy-Based Sandbox Approach to Bridge Attitudes, Goals, Knowledge, and Behaviors in the Privacy Paradox. *arXiv preprint arXiv:2309.14510* (2023).
[17] ChildSafe.ai. 2022. https://childsafe.ai/.
[18] Janet Currie and Cathy Spatz Widom. 2010. Long-term consequences of child abuse and neglect on adult economic well-being. *Child maltreatment* 15, 2 (2010), 111–120.
[19] Julia Davidson and Petter Gottschalk. 2010. Internet child abuse: Current research and policy. (2010).
[20] Mateus de Castro Polastro and Pedro Monteiro da Silva Eleuterio. 2010. Nudetective: A forensic tool to help combat child pornography through automatic nudity detection. In *2010 Workshops on Database and Expert Systems Applications.* IEEE, 349–353.
[21] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. 2014. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2377–2386.
[22] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. 2019. Exploring how privacy and security factor into IoT device purchase behavior. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–12.
[23] Abhishek Gangwar, Víctor González-Castro, Enrique Alegre, and Eduardo Fidalgo. 2021. AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images. *Neurocomputing* 445 (2021), 81–104.
[24] Lisa Geierhaas, Fabian Otto, Maximilian Häring, and Matthew Smith. 2023. Attitudes towards Client-Side Scanning for CSAM, Terrorism, Drug Trafficking, Drug Use and Tax Evasion in Germany. In *2023 IEEE Symposium on Security and Privacy (SP).* IEEE Computer Society, 217–233.
[25] Ateret Gewirtz-Meydan and David Finkelhor. 2020. Sexual abuse and assault in a large national sample of children and adolescents. *Child maltreatment* 25, 2 (2020), 203–214.
[26] Giffeye. 2022. https://www.griffeye.com/.
[27] Google. 2020. https://protectingchildren.google/.
[28] Google. 2021. *safety.google.* https://safety.google/intl/en_nz/stories/hash-matching-to-help-ncmec/
[29] William R Graham Jr. 2000. Uncovering and Eliminating Child Pornography Rings on the Internet: Issues regarding and Avenues Facilitating Law Enforcement's Access to Wonderland. *L. Rev. MSU-DCL* (2000), 457.
[30] Guardian. 2021. Apple's plan to scan for child abuse images 'tears at heart of privacy'. https://www.theguardian.com/world/2021/oct/15/apple-plan-scan-child-abuse-images-tears-heart-of-privacy
[31] Catherine Hamilton-Giachritsis, Elly Hanson, Helen Whittle, Filipa Alves-Costa, Andrea Pintos, Theo Metcalf, and Anthony Beech. 2021. Technology assisted child sexual abuse: Professionals' perceptions of risk and impact on children

and young people. *Child Abuse & Neglect* 119 (2021), 104651.
[32] Christine Heim, Margaret Shugart, W Edward Craighead, and Charles B Nemeroff. 2010. Neurobiological and psychiatric consequences of child abuse and neglect. *Developmental psychobiology* 52, 7 (2010), 671–690.
[33] Jason Hong. 2013. Considering privacy issues in the context of Google glass. , 10–11 pages.
[34] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. 2014. Privacy behaviors of lifeloggers using wearable cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 571–582.
[35] Smirity Kaushik, Yaxing Yao, Pierre Dewitte, and Yang Wang. 2021. " How I Know For Sure": People's Perspectives on Solely Automated Decision-Making(SADM). In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 159–180.
[36] Brian C Kavanaugh, Jennifer A Dupont-Frechette, Beth A Jerskey, and Karen A Holler. 2017. Neurocognitive deficits in children and adolescents following maltreatment: Neurodevelopmental consequences and neuropsychological implications of traumatic stress. *Applied Neuropsychology: Child* 6, 1 (2017), 64–78.
[37] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A" nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. 1–12.
[38] Dan Kennedy. 2021. Apple's attempted crackdown on child sexual abuse leads to a battle over privacy. https://www.wgbh.org/news/commentary/2021/09/08/apples-attempted-crackdown-on-child-sexual\-abuse-leads-to-a-battle-over-privacy
[39] Juliane A Kloess, Anthony R Beech, and Leigh Harkins. 2014. Online child sexual exploitation: Prevalence, process, and offender characteristics. *Trauma, Violence, & Abuse* 15, 2 (2014), 126–139.
[40] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31.
[41] Tu Le, Alan Wang, Yaxing Yao, Yuanyuan Feng, Arsalan Heydarian, Norman Sadeh, and Yuan Tian. 2023. Exploring Smart Commercial Building Occupants' Perceptions and Notification Preferences of Internet of Things Data Collection in the United States. *arXiv preprint arXiv:2303.04955* (2023).
[42] Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. 2020. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation* 34 (2020), 301022.
[43] Rebecca T Leeb, Terri Lewis, and Adam J Zolotor. 2011. A review of physical and mental health consequences of child abuse and neglect and implications for practice. *American Journal of Lifestyle Medicine* 5, 5 (2011), 454–468.
[44] Rainer Lienhart and Jochen Maydt. 2002. An extended set of haar-like features for rapid object detection. In *Proceedings. international conference on image processing*, Vol. 1. IEEE, I–I.
[45] Heather Richter Lipford, Madiha Tabassum, Paritosh Bahirat, Yaxing Yao, and Bart P Knijnenburg. 2022. Privacy and the internet of things. *Modern Socio-Technical Perspectives on Privacy* (2022), 233.
[46] Sonia Livingstone and Peter K Smith. 2014. Annual research review: Harms experienced by child users of online and mobile technologies: The nature, prevalence and management of sexual and aggressive risks in the digital age. *Journal of child psychology and psychiatry* 55, 6 (2014), 635–654.
[47] Ben Lovejoy. 2022. *Apple's CSAM troubles may be back, as EU announces a law requiring detection.* https://9to5mac.com/2022/05/11/apples-csam-troubles-may-be-back-as-eu-plans-a-law-requiring-detection/
[48] Joao Macedo, Filipe Costa, and Jefersson A dos Santos. 2018. A benchmark methodology for child pornography detection. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 455–462.
[49] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp* 4 (2008), 543.
[50] Aleecia M McDonald and Lorrie Faith Cranor. 2010. Americans' attitudes about internet behavioral advertising practices. In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*. 63–72.
[51] Microsoft. 2022. Photodna. https://www.microsoft.com/en-us/PhotoDNA
[52] Anthony D Miyazaki and Ana Fernandez. 2001. Consumer perceptions of privacy and security risks for online shopping. *Journal of Consumer affairs* 35, 1 (2001), 27–44.
[53] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. 2017. Privacy expectations and preferences in an IoT world. In *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*. 399–412.
[54] ncmec. 2020. *electronic service provider.* https://www.missingkids.org/content/dam/missingkids/pdfs/2021-reports-by-esp.pdf
[55] Department of Justice. 2021. https://www.justice.gov/coronavirus/keeping-children-safe-online.
[56] Claudia Peersman, Christian Schulze, Awais Rashid, Margaret Brennan, and Carl Fischer. 2014. icop: Automatically identifying new child abuse media in p2p networks. In *2014 IEEE Security and Privacy Workshops*. IEEE, 124–131.
[57] Ethel Quayle. 2009. Guidelines for Parents, Guardians and Educators on Child Online Protection. (2009).

[58] Adi Robertson. 2021. Apple's controversial New child protection features, explained. https://www.theverge.com/2021/8/10/22613225/apple-csam-scanning-messages-child-\safety-features-privacy-controversy-explained

[59] Elly Robinson. 2013. Parental involvement in preventing and responding to cyberbullying. *Family Matters* 92 (2013), 68–76.

[60] Napa Sae-Bae, Xiaoxi Sun, Husrev T Sencar, and Nasir D Memon. 2014. Towards automatic detection of child pornography. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 5332–5336.

[61] Jerome H Saltzer and Michael D Schroeder. 1975. The protection of information in computer systems. *Proc. IEEE* 63, 9 (1975), 1278–1308.

[62] Tali Sharot. 2011. The optimism bias. *Current biology* 21, 23 (2011), R941–R945.

[63] Rajnesh D Singh. 2018. Mapping online child safety in Asia and the Pacific. *Asia & the Pacific Policy Studies* 5, 3 (2018), 651–664.

[64] Spotlight. 2019. https://www.thorn.org/spotlight/.

[65] FTC Staff. 2011. Protecting consumer privacy in an era of rapid change–a proposed framework for businesses and policymakers. *Journal of Privacy and Confidentiality* 3, 1 (2011).

[66] Janet Stanley. 2002. Child abuse and the Internet [This article is reproduced from Issues in Child Abuse Prevention, no. 15, Summer 2001.]. *Journal of the Home Economics Institute of Australia* 9, 1 (2002), 5–27.

[67] Thorn. 2020. https://www.thorn.org/blog/announcing-safer-built-by-thorn-eliminate-csam/.

[68] Janice Y Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. 2011. The effect of online privacy information on purchasing behavior: An experimental study. *Information systems research* 22, 2 (2011), 254–268.

[69] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *proceedings of the eighth symposium on usable privacy and security*. ACM, 4.

[70] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, Vol. 1. Ieee, I–I.

[71] Yang Wang, Huichuan Xia, and Yun Huang. 2016. Examining American and Chinese Internet Users' Contextual Privacy Preferences of Behavioral Advertising. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 539–552.

[72] Zixin Wang, Danny Yuxing Huang, and Yaxing Yao. 2023. Exploring Tenants' Preferences of Privacy Negotiation in Airbnb. In *32nd USENIX Security Symposium (USENIX Security 23)*. 535–551.

[73] Zikai Alex Wen, Jingyu Jia, Hongyang Yan, Yaxing Yao, Zheli Liu, and Changyu Dong. 2023. The influence of explanation designs on user understanding differential privacy and making data-sharing decision. *Information Sciences* 642 (2023), 118799.

[74] Zack Whittaker. 2021. Apple delays plans to roll out CSAM detection in IOS 15 after privacy backlash. https://techcrunch.com/2021/09/03/apple-csam-detection-delayed/

[75] Sandy K Wurtele and Maureen C Kenny. 2010. Partnering with parents to prevent childhood sexual abuse. *Child Abuse Review: Journal of the British Association for the Study and Prevention of Child Abuse and Neglect* 19, 2 (2010), 130–152.

[76] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. 2019. Defending my castle: A co-design study of privacy mechanisms for smart homes. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[77] Yaxing Yao, Yun Huang, and Yang Wang. 2019. Unpacking People's Understandings of Bluetooth Beacon Systems-A Location-Based IoT Technology. (2019).

[78] Yaxing Yao, Davide Lo Re, and Yang Wang. 2017. Folk Models of Online Behavioral Advertising. *Proceedings of Computer-Supported Cooperative Work* (2017).

[79] Michele L Ybarra and Kimberly J Mitchell. 2007. Prevalence and frequency of Internet harassment instigation: Implications for adolescent health. *Journal of Adolescent Health* 41, 2 (2007), 189–195.

[80] Emilios Yiallourou, Rafaella Demetriou, and Andreas Lanitis. 2017. On the detection of images containing child-pornographic material. In *2017 24th International Conference on Telecommunications (ICT)*. IEEE, 1–5.

[81] Eric Zeng, Shrirang Mare, and Franziska Roesner. 2017. End user security and privacy concerns with smart homes. In *thirteenth symposium on usable privacy and security ({SOUPS} 2017)*. 65–80.

[82] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. 2022. How usable are ios app privacy labels? *Proceedings on Privacy Enhancing Technologies* 4 (2022), 204–228.

[83] Yuhang Zhao, Yaxing Yao, Jiaru Fu, and Nihan Zhou. 2023. {"If} sighted people know, I should be able to {know:"} Privacy Perceptions of Bystanders with Visual Impairments around Camera-based Technology. In *32nd USENIX Security Symposium (USENIX Security 23)*. 4661–4678.

[84] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User perceptions of smart home IoT privacy. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–20.

A APPENDIX

Screening Survey

Q1. Electronic services from which companies do you regularly use?

Q2. How many hours a day do you spend on services provided by the following companies? (Answered on a eleven points Likert scale from 0 to 10)

Q3. Are you willing to participate in a 1-hour online interview about online safety tools?

Interview Protocols

Demographic Questions

Q1. What gender do you identify as? (Female, Male, Non-binary, Other, Prefer not to answer)

Q2. What is your age?

Q3. What is the highest degree or level of school you have completed?

Q4. Do you have kids?

Q5. How much time daily do you spend on the following platforms (insert companies names)?

Q6. Where do you locate?

General Questions about participants' Perceptions of Online Child Abuse

Q7. How do you define child exploitation and abuse? Do you think child abuse exists on the online world?

Q8. In the online world how do you think child abuse is carried out?

Q9. Do you know how to report child exploitation if you accidentally observe it on digital platforms?

Q10. Have you come across any child abuse content online?

Q11. Are you aware of practices carried out by companies like (Apple, Google, Facebook, Microsoft) to fight against child abuse on their platforms?

Technology Usage Questions

Q12. What electronic device do you have?

Q13. Can you please briefly explain what you use them for?

Questions about current detection practices It should be noted that the question was asked separately for two different ESPs. We pick two platforms - Apple, and the other is one platform from our list that they use frequently.

Q14. Have you ever come across the implementation of these practices?

Q15. What is your first impression after knowing about these practices?

Q16. Do you think these companies provide you with enough information to help you understand the implementation of this practice ?

Q17. What additional information can it provide?

Q17. Are you able to understand all the terminologies and technical terms in the statement?

Q19. Will you be comfortable with [Company] using this practice to identify child abuse content?

Q20. Do you think it is necessary to provide you with this information? Would you prefer to receive notification regarding privacy practices?

Q21. What are the pros and cons of this practice?

Q22. How effective or ineffective do you think these practices are in identifying and restricting child exploitation content?

Q23. What types of data or files do you think this practice can access?

Q24. Now you have learned these detection mechanisms, as a user, which one would you prefer to have? Do you prefer to have the detection on the server side, or on your own device?

Q25. Since you mentioned privacy a couple of times, I'm wondering what you think of the trade-off between your privacy and the social benefits you just mentioned.

Received January 2023; revised July 2023; accepted November 2023