Mental Models, Expectations and Implications of Client-Side Scanning: An Interview Study with Experts

Divyanshu Bhardwaj‡
CISPA Helmholtz Center for Information Security, *and* Saarland University
Germany

Carolyn Guthoff‡
CISPA Helmholtz Center for Information Security, *and* Saarland University
Germany

Adrian Dabrowski
CISPA Helmholtz Center for Information Security
Germany

Sascha Fahl
CISPA Helmholtz Center for Information Security
Germany

Katharina Krombholz
CISPA Helmholtz Center for Information Security
Germany

Figure 1: The cell phone and its capabilities have become the center of attention for investigators, privacy experts, and parental advocates. (Collage from multiple DALL-E 2 images, 2023)

Trigger Warning

This paper includes mention of child sexual abuse (CSA) and child sexual abuse material (CSAM).

ABSTRACT

Client-Side Scanning (CSS) is discussed as a potential solution to contain the dissemination of child sexual abuse material (CSAM). A significant challenge associated with this debate is that stakeholders have different interpretations of the capabilities and frontiers of the concept and its varying implementations. In this paper, we explore stakeholders' understandings of the technology and the expectations and potential implications in the context of CSAM by conducting and analyzing 28 semi-structured interviews with a diverse sample of experts. We identified mental models of CSS and the expected challenges. Our results show that CSS is often a preferred solution in the child sexual abuse debate due to the lack of an alternative. Our findings illustrate the importance of further interdisciplinary discussions to define and comprehend the impact of CSS usage on society, particularly vulnerable groups such as children.

CCS CONCEPTS

• **Social and professional topics** → **Privacy policies**; • **Security and privacy** → *Social aspects of security and privacy*; Privacy protections.

KEYWORDS

client-side scanning, child sexual abuse material, surveillance, crime prevention

ACM Reference Format:

Divyanshu Bhardwaj, Carolyn Guthoff , Adrian Dabrowski, Sascha Fahl, and Katharina Krombholz. 2024. Mental Models, Expectations and Implications of Client-Side Scanning: An Interview Study with Experts. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3642310

‡Both authors contributed equally to this research.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

1 INTRODUCTION

The Cyber Tipline of the U.S. National Center for Missing and Exploited Children (NCMEC) received 32,059,029 reports in 2022, 99.5% of which were classified as CSAM (child sexual abuse material) [16]. CSA (child sexual abuse) is an emotionally loaded crime, and Client-Side Scanning (CSS) is often named as a promising technique in the ongoing heated discussion about ways to limit the dissemination of CSAM.

Major service providers practice server-side scanning for CSAM on cloud storage and centralized messaging service [53]. However, whenever end-to-end (transit) encryption (E2EE) is used, server-side scanning is not feasible. Law enforcement and intelligence services argue that E2EE severely hinders their work [7].

Previous attempts at lawful circumvention of E2EE, including the GCHQ's *ghost proposal* [40], a key escrow service, were all met with backlash [57]. CSS is a new approach that promises to find targeted material at the source or destination of a transmission rather than during transit. In the summer of 2021, Apple introduced a version of CSS to fight CSAM [26], which was met with major pushback that led to the project's scrapping in late 2022 [48].

Current legislative proposals or related assessments from the EU [12, 13], the UK [14], and the U.S. [20, 43] still mention CSS as a preventative measure to fight the dissemination of CSAM. However, a major challenge is that CSS is a vaguely defined meta concept that crams vastly different approaches into the naïve definition of locally scanning a client device for targeted material. Some CSS approaches are discussed for known or unknown material, while others talk about different places where the scanning happens. All would lead to vastly different implementations and thus, many discussions are doomed without a standard base definition. Valid arguments about expectations, benefits, implications, and risks associated with CSS may be dismissed by those with a different understanding of what CSS entails. The emotional nature of the issue, combined with the differing interpretations of CSS, complicates a solution-oriented discussion for the original goal of fighting CSA and the distribution of CSAM.

Despite previous research on the topic, the discussion still lacks a comprehensive definition of CSS, as well as an understanding of the expectations regarding different CSS systems and their potential implications. A universal definition is necessary for a solution-oriented discussion that finds new approaches and looks at the problem from different perspectives apart from the currently hardened fronts between proponents and opponents of CSS. Our paper provides a foundation for a universal definition of CSS by structuring the different dimensions of experts' mental models.

To this end, we interviewed 28 stakeholders, from child protection and law enforcement to data protection and academia, about their understanding of CSS, expectations, and potential implications. We decontextualize CSS from the fight against CSA to establish a joint technical foundation of the mental models. Afterward, we recontextualize CSS in the fight against CSAM, discussing the expectations and implications of different CSS methods.

Our results outline the goal of CSS and provide an overview of mental models with the different dimensions related to CSS. Depending on the technical understanding of our participants, the breadth and depth of their mental models were vastly different. Overall, the term CSS combines different types of scanning that depend on the content, how intrusive it would be, and whether a definite output is wished for. In the context of CSAM, the expectations of CSS shed light on goals, adoption and reception, implementation, the expected usage of the technology, and challenges that could be encountered. Most interestingly, many participants pointed out that the lack of a good alternative to CSS makes it a viable solution to the CSA problem. The implications of CSS highlight the consequences that the system would have on a personal and societal level and outline the threats it could pose to certain stakeholder groups. Surprisingly, children, as the primary beneficiaries, can also face significant limitations of their rights, such as privacy and free will, depending on the type of CSS implementation used.

This paper is structured as follows. Section 2 details background knowledge such as E2EE as well as algorithms and protocols that could be used in CSS. Section 3 summarizes related work. Section 4 describes our methodology, and Section 5 details our results. Finally, Sections 6 and 7 contribute a thorough discussion and outline limitations.

2 BACKGROUND

2.1 CSA and CSAM distribution

The creation, storage, and dissemination of CSAM has increased significantly in the recent decade [8, 31, 54]. Sexual exploitation of children has previously been addressed as child pornography, which fails to convey the emotional and criminal severity of the problem while also subjecting the victim to potential stigmatization and re-victimization [32]. Children safety organizations define CSAM as any content that depicts sexually explicit activities involving a person younger than 18 years old (minor) [2, 3]. A large fraction of the distribution of CSAM takes place over Electronic Service Providers (ESPs) like Facebook and WhatsApp [16]. Most widely used ESPs employ some level of end-to-end encryption in their communication services, which makes detecting CSAM inherently tricky.

2.2 Access to End-to-End Encrypted Communication

End-to-end encryption (E2EE) prevents third parties like service providers, law enforcement, and malicious actors from accessing the transmission channel to ensure the privacy of users [55]. While E2EE is beneficial for the general population, it has been abused by malicious actors for unscrupulous activities ranging from concealing communication and spreading misinformation to disseminating CSAM. Law enforcement has been pressing for more comprehensive access to encrypted communication channels to collect evidence and prohibit possible crimes.

In February 2016, following the San Bernardino shooting incident in the U.S., the Federal Bureau of Investigation (FBI) asked Apple to build and install a weakened operating system onto one of the phones recovered from the shooters [24]. In December 2016, the UK passed the 'Investigatory Powers Act,' which legally obligated communication service providers to allow intelligence agencies and law enforcement access to communications data and internet connection records [6]. In 2018, the Australian government passed the 'Telecommunication and Other Legislation Amendment,' which

required companies to grant intelligence agencies and law enforcement access to E2EE communications [23]. In 2021, the UK published a draft of its Online Safety Bill [14] intended to prevent the spread of CSAM and protect children from inappropriate content. However, in September 2023, the UK government acknowledged that the technology needed to securely scan encrypted messages without compromising users' privacy doesn't exist [36]. Recently, in an effort to combat the creation and dissemination of CSAM, the EU has proposed a regulation considering Client-Side Scanning as a potential solution [1, 13]. In the controversially discussed approach, CSS systems would utilize perceptual hashes that are more resilient to subtle alterations than conventional cryptographic hashes when scanning for content on the client.

2.3 Scanning Algorithm Types

Various technologies can be employed to scan for content, which are often synonymously described as CSS.

2.3.1 Cryptographic Hashing. A cryptographic hash algorithm generates short, fixed-size, and unique hashes for different variable-length inputs, even if two inputs have only slight differences. A cryptographic hash that produces the same output for different inputs (hash collision) is considered compromised. The concept originates from computer security, where hashes are widely used, e.g., to store passwords or safeguard data integrity. Their input is variable, i.e., cryptographic hashes can be generated for text, images, and other types of data. They are ideal for verifying data integrity but are limited regarding images, for example. Two images with slight pixel differences can produce different hashes but look the same to the viewer.

2.3.2 Perceptual Hashing. Perceptual hash algorithms originate in computer vision and are suitable for an application on visual data such as images or video frames. A perceptual hash algorithm (e.g., PhotoDNA [44]) generates the same hash output for similar inputs. This approach makes perceptual hashes resistant to small changes that do not affect the overall appearance. Numerous perceptual hashing methods and techniques are available, each with its own set of characteristics.

2.3.3 Machine Learning. Machine learning is another vast space of methods to detect inappropriate content. While hash-based techniques can only check for known content and subtle alterations, machine learning models can be trained to check for known as well as previously unknown content. Their output is the likelihood of whether a checked input matches the specified content criteria.

2.3.4 Industry Example. NeuralHash is a deep perceptual-hash-based CSS system developed by Apple to detect known CSAM by performing on-device scanning of images [26]. The NeuralHash algorithm incorporates two modules: a convolutional neural network and a perceptual hash function. To check if an image contains CSAM, the input image is first fed into the convolutional neural network, which generates an input descriptor. The descriptor is subsequently fed into the perceptual hash function. The generated hash is then compared against a list of known CSAM.

2.4 Matching Techniques

Scanning systems employ matching techniques to check for matches. Relevant matching techniques are discussed below.

2.4.1 Private Set Membership. Private Set Membership (PSM) is a technique that allows for querying a database stored on a server in a privacy-preserving way [30, 33], e.g., checking if a hash is present without revealing it to the server. In the CSS context, a device can use PSM to query if an image belongs to a database of known CSAM without gaining access to the actual database [60].

2.4.2 Homomorphic Encryption. Homomorphic Encryption is a technique that allows a system to perform calculations and check for matching data on an encrypted database [58]. In the CSS context, a device can use Homomorphic Encryption to check for near-duplicate CSAM images [60].

2.5 History of Scanning Technologies

Prior to being used for scanning CSAM, content scanning technologies have been employed in various non-CSAM contexts for a long time. Some of these contexts are discussed below.

2.5.1 Client-Side Scanning. The concept Client-Side Scanning has been used in non-CSAM contexts such as antivirus software. These programs scan files at the client's end to detect potentially harmful viruses and malware based on centrally managed fingerprints. Advertising elements on the web can contain malicious (or annoying) code that can be used to exploit unsuspecting users. Ad-blockers scan webpages as they load to identify and block advertising elements. A form of CSS is also implemented as spell checkers that come packaged with editors and word processing software to identify misspelled words. Similarly, web browsers use input validation techniques to review user input for any malicious content before sending it to the server. Copying machines and picture editing software scan for common patterns found on bank notes to inhibit their reproduction [17].

2.5.2 Perceptual Hashing. Perceptual hashing is extensively used in digital forensics [51, 52, 62] to check if images have been manipulated. It also finds considerable use in reverse image search engines, allowing users to search for an image by uploading similar images [18, 19, 21]. Additionally, perceptual hashing is a critical component in analyzing and detecting image-based misinformation in social messaging applications [62].

3 RELATED WORK

3.1 Timeline of CSAM Detection Using CSS

In 2019, the U.S., UK, and Australia expressed concerns regarding the use of strong end-to-end encryption in communication channels [7] due to the increasing dissemination of CSAM. To combat this issue, CSS was proposed as a solution that maintains end-to-end encryption. However, privacy organizations [50] and experts [49] argued that it would considerably weaken encryption [34] and leave people vulnerable to exploitation. In August 2021, Apple announced its plans to combat CSA [25] by scanning images stored on-device and matching them against a database of known CSAM provided by child safety organizations like NCMEC. Their CSS system aimed to detect identical or visually similar images to the ones

contained in a hash database of known CSAM content provided by NCMEC. The proposal received mixed reactions: While child safety organizations like NCMEC welcomed the idea [45], others were strictly against it. Well-known privacy and security advocates emphasized the need for carefulness and exercising caution, citing the potential threats associated with the mass deployment of such a technology [22]. The majorly negative feedback from experts and pushback from users [61] led to Apple announcing delays in releasing its CSS-based CSAM detection system [26]. Concurrently, to put the effectiveness and robustness of perceptual hashing-based algorithms [42] like NeuralHash to the test, the scientific community started conducting experiments and found the algorithms to be vulnerable to various adversarial attacks [38, 59]. In December 2022, Apple scrapped their CSS-based CSAM detection tool, acknowledging that "children can be protected without companies combing through personal data" [48]. Instead, Apple pivoted its efforts towards a less invasive approach, protecting children by warning them when sending or receiving images with nudity and providing resources to make safe choices [28, 29]. Furthermore, Thomas et al. [60] looked into existing blocklisting techniques and proposed design principles to implement privacy-preserving and transparent protocols for on-device blocklisting.

Since 2020, the EU has been designing a course of action to fight CSA [9], looking into technical solutions for CSAM detection in end-to-end encrypted communication channels [10]. The EU passed a temporary derogation of the ePrivacy Act [11], which asked ESPs to voluntarily scan for and report CSAM transmitted via an encrypted channel. This derogation was heavily denounced with findings that the automated tools reported non-targeted content 86% of the time [15], leading to third-party reviewers and law enforcement peeking at private encrypted content. Finally, in May 2022, in a draft regulation, the EU laid down the obligation of ESPs to deploy the so-called 'Chat Control' regulation mandatorily [1] and proposed rules to combat CSAM [12, 13]. Similar legislative acts have also been proposed in the UK [14] and the U.S. [7].

3.2 Attacks on Perceptual Hashing Algorithms

Perceptual hashing algorithms are used for scanning images due to their inherent robustness against small changes. Consequently, they are often recommended to be used in CSS built to detect CSAM. However, recent literature evaluating the efficacy of these algorithms has found them vulnerable to various adversarial attacks. Hao et al. [37] experimentally demonstrated that perceptual hash algorithms could be attacked by enlarging the hash distance between images while keeping images visually the same. They also found the attack to be transferable between different perceptual hashing algorithms. Jain et al. [42] evaluated the robustness of five widely employed perceptual hashing algorithms utilizing three adversarial attacks and found them to be extremely vulnerable to detecting avoidance attacks. Struppek et al. [59] empirically analyzed NeuralHash, a deep perceptual hashing algorithm proposed by Apple, and found it susceptible to detection evasion and hash collision attacks. In follow-up work, Jain et al. [41] showed that perceptual hashing algorithms could also have concealed secondary intentions, which could be used to identify target individuals based on face recognition. Hooda et al. [39] showed that Client-Side Scanning systems could emulate physical surveillance by way of database poisoning attacks. Due to privacy and security concerns, these researchers advised against the widespread real-world usage of perceptual hash-based Client-Side Scanning systems.

3.3 End-User Perspective

Geierhaas et al. [35] studied end-user attitudes and perspectives in Germany of CSS for CSAM, terrorism, and other crimes through a survey. They found that participants were more receptive to specific implementations of CSS to combat serious crimes such as child abuse but were also worried about the potential misuse of the technology. Further, their findings show participants' indifferent attitudes toward client-side or cloud-based scanning, with the authors stating that "the general population [sample] might not weigh the [technical] distinction in the same way as the tech-community" [35].

Despite prior research, there is still no complete explanation for CSS and little comprehension of what CSS systems are meant to achieve and their resulting consequences. It is imperative to have a general definition to facilitate a deeper understanding of the issue. Our work aims to address this gap by gathering insights from a diverse group of experts and organizing their mental models into distinct dimensions, laying the groundwork for a universal definition of CSS.

4 METHODOLOGY

We explore mental models, expectations towards CSS, and understanding of potential implications of both the technology and its application through a qualitative approach: 28 semi-structured interviews with domain experts. We decided on this approach because CSS, as proposed by the European Commission to counter child sexual abuse (CSA) [12], is not rolled out at the time of writing, and the topic is sensitive. Our approach intends to answer the following research questions.

RQ1 What are experts' mental models of Client-Side Scanning? There exists no comprehensive definition of CSS. Thus, this RQ aims to provide a joint base for future discussions on what CSS is and what it can be by analyzing experts' understanding of the topic.

RQ2 What are experts' expectations of Client-Side Scanning? This RQ aims to outline expectations that experts have for a CSS system scanning for CSAM, that is their anticipation of system's goals, adoption, usage and challenges.

RQ3 What are the potential implications of Client-Side Scanning? This RQ aims to comprehend the potential consequences of implementing a CSS system in the context of CSAM. This is essential for informed decision-making and risk mitigation.

4.1 Study Design

We utilize semi-structured interviews, which provide a guideline while simultaneously allowing us to react with pertinent questions regarding the interview topic.

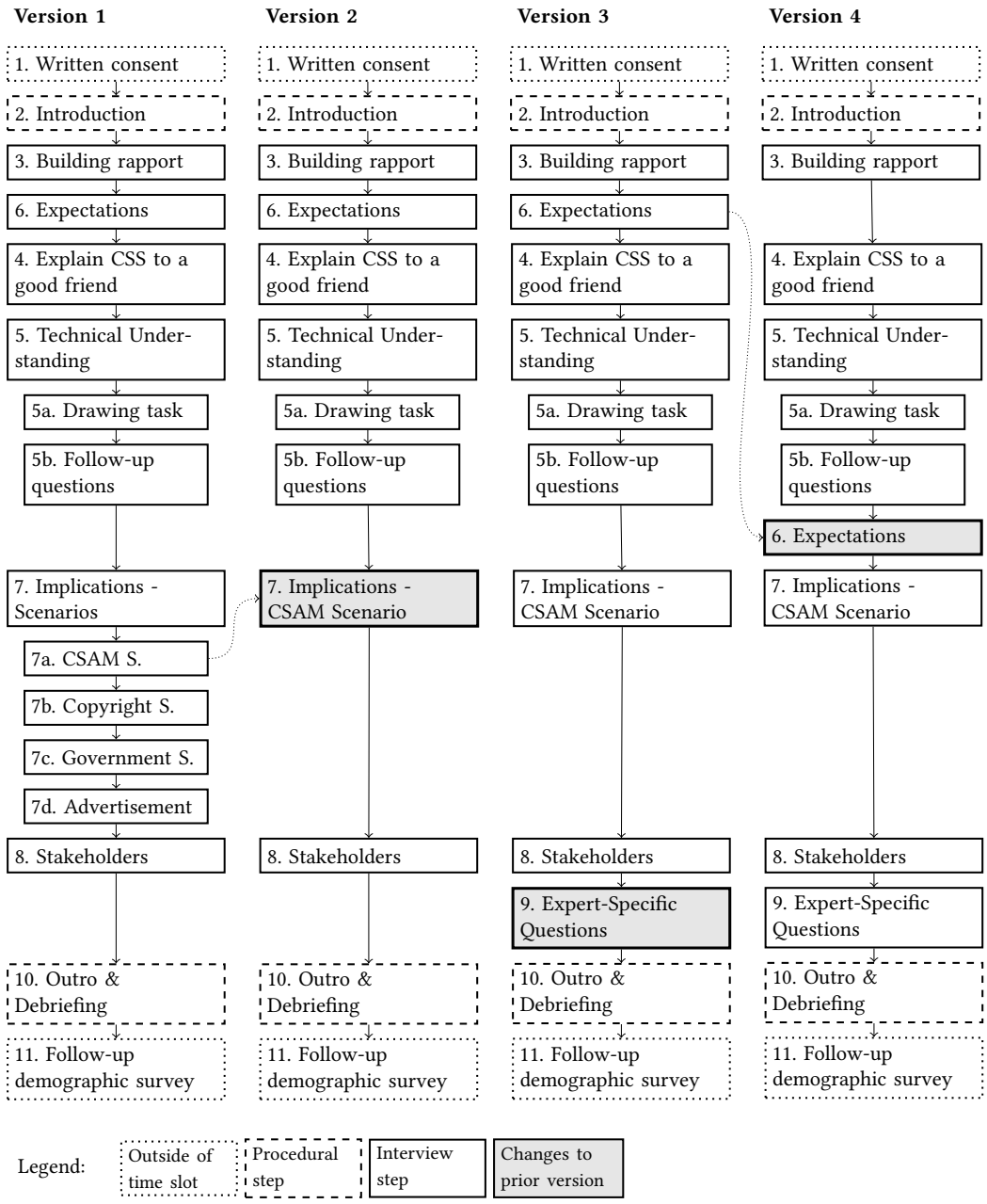

Figure 2: High-level overview of our general setup and the interview in particular, as well as its development throughout the study.

Due to the (at times) emotional and heated discussion about CSS as a potential CSA mitigation, it is impossible to decouple CSS from the CSA debate fully - neither could we. However, where possible, at several points throughout the methodology and the results, we (partially) decontextualized CSS from CSAM to collect and describe the details of the mental models while minimizing the complexity introduced by the context. Afterward, we recontextualize the topic to focus on expectations and implications within the current context of the CSA debate. We specifically mention where contextualization or decontextualization happens.

4.2 Interview Guideline

Our interview guideline explored all aspects of our research questions. Following this, we first explain the structure of the interview guideline before describing its evolution. CSA and CSAM were purposefully mentioned only in stages 1, 2, and 7. The former two

were intended as trigger warnings.

Figure 2 gives a visual overview of our interview procedure and its evolution, while Table 3 in Appendix B shows the full guideline.

4.2.1 Structure.

1. Written Consent. Before each interview, every participant gave consent via a Qualtrics survey. This was separated from the demographics survey to keep anonymity. The survey stated that CSA might be a topic of conversation during the interview.

2. Introduction. We gave a short description of the interview and a trigger warning about CSA and CSAM. We made clear that participants could terminate the interview without reason at any time.

3. Building Rapport. To build rapport, we asked general questions about CSS, such as when they first learned about it and their general attitude towards CSS.

4. Explain CSS to a Good Friend. To ease our participants into talking about their technical understanding of CSS, we tasked them with explaining CSS to a good friend who has no technical background knowledge. Afterward, we asked what they omitted due to complexity.

5. Technical Understanding. Following the simplified explanation, we let the participants draw and describe their full technical understanding of CSS. We followed up with questions about different parts of the explanation that the interviewer deemed interesting and about potential security and privacy risks.

The context given for stages 4 and 5 was finding target material on a cellphone. We did not mention CSAM, and when asked, we replied that the task was not intended within a CSA context, but the participants could explain their understanding within this context.

6. Expectations. We asked our participants about their expectations towards CSS, both on a national and a global level.

7. Implications - Four Scenarios. We asked our participants which implications they could foresee for different scanning scenarios: scanning for CSAM, scanning for copyright infringements, scanning by a government responsible for the targeted material, and a fourth scenario where CSS is used for advertisement purposes.

For each scenario, we asked the same questions about user behavior for notifications, personal implications of both a match and a false positive, and societal implications if this scenario were to be implemented on most end-user devices.

7. Implications - only CSAM Scenario. The single scenario focuses on CSAM scanning and asks analog questions to the prior version, focusing on the personal and societal implications of matches and false positives.

8. Stakeholders. We asked our participants which stakeholders should be included in the CSS discussion. It is crucial to gain a holistic overview of all stakeholders that should be heard.

9. Expert-Specific Questions. We added expert-specific questions that asked participants how the adoption of CSS would influence their work or profession. This section also provided room to ask targeted questions. For example, prior to the interview, X22(A-D) pointed to an online document that we briefly discussed with them.

10. Outro. Finally, we asked whether the participant had anything to add or any questions. We also informed about the follow-up demographic survey.

11. Follow-Up. After every interview, we sent a link to a demographic survey that included questions about the highest level of education, educational background, country of work, current field of work, and age.

4.2.2 Evolution of the Interview Guideline.

Throughout our study, we adapted our interview guideline. The used version for each interview is denoted in Table 2 in Appendix A.

Pilot Study. We piloted version 1 through four recorded in-person interviews in early 2022. Our participants gave oral consent, two were usable security experts, one was a security expert, and one had policy expertise. After each session, we reviewed all parts of the interview with the pilot participant and asked for detailed feedback.

Based on this feedback, we changed some question phrasings and the scenario sequence for improved interview flow.

Changes from Version 1 to Version 2. We cut down to one scenario, as introducing different scenarios biased our participants regarding content that could be scanned for.

Changes from Version 2 to Version 3. Several participants mentioned special confidentiality requirements of some professions possibly influenced by CSS. We took the journalist (X19) as an occasion to add expert-specific questions (9).

Changes from Version 3 to Version 4. Finally, we moved the expectation questions (6) behind the technical understanding questions (5). Some participants used technical explanations to explain their expectations, and the interviewer had difficulty asking questions about expectations without first understanding the participants' technical views.

4.3 Recruitment and Participants

We focused our recruitment on experts in their respective fields with some CSS knowledge. To recruit effectively, we first started with a stakeholder analysis.

A stakeholder is an individual or group with a vested interest in a system. They can either influence or be affected by it. Analyzing stakeholders is crucial to determining their understanding for and anticipation of potential concerns, which helps make informed decisions and address a technology's potential risks. For our study, we tried to cover as many different stakeholders as possible. We fixated on the following stakeholder groups: academia, politics, digital rights, children's rights (including child protection), law enforcement, service providers, and technology companies. We built this initial list based on a literature review about 'Client-Side Scanning' and 'Chat Control' encompassing academic papers, position papers, newspaper articles, blog posts, tweets on X (formerly Twitter), and discussions with peers.

Second, based on this list, we built a sample of persons of interest. Our selection criteria included some stake in the topic and proper background knowledge depending on the type of stakeholder. For example, children's rights stakeholders either work for a child protection organization.

Third, we compiled lists of individuals, companies, and government organizations and wrote personalized invitation emails to each individual or entity. We also used snowball recruitment by asking our participants to put us in contact with people they would choose to interview next. A couple of participants were recruited through personal contacts.

Table 2 in Appendix A gives an overview of our participants. Columns 2, 4, and 5 depict the demographic survey results. We added the type of stakeholder for each participant based on why we recruited them and their answers from the interviews. In agreement with Participant X21, we omitted any demographics for anonymity reasons due to the sensitivity of their work environment.

4.4 Data Collection

We collected 28 expert views on CSS through 25 semi-structured interviews between November 2022 and June 2023. A letter suffix to some participant's label in Table 2 denotes a joint interview session. All interviews were held by one researcher who speaks German and English on a native or near-native level. They held all interviews to avoid introducing bias through gender or cultural differences in either English or German, depending on the participant's preference.

All interviews were recorded. 19 were held via Zoom, one via another video platform, and five in person.

For the drawing task, we used Zoom's built-in whiteboard feature when operational. In two cases, pen and paper were used and shown via video, with a digital copy sent in afterward. One participant could not draw because of his surroundings, and one other declined the drawing task. In these cases, and for interview X22 with four interview partners, we asked the participants to describe their understanding verbally. In-person interviewees used pen and paper.

4.5 Data Analysis

We divided the data analysis into two parts, one focusing on RQ1 (mental models) and the other jointly on RQ2 (expectations) and RQ3 (implications). The split facilitates our efforts of decontextualizing the mental models from the debate around CSAM.

Two researchers analyzed as follows. For the part focusing on RQ1, one researcher was responsible, and the other did supportive work, and vice versa on the part focusing on RQ2 and RQ3. The analysis steps are identical for both analyses and are thus described below only once, but the steps were conducted twice (once for each part), resulting in two codebooks, which can be found in Appendix C. Only the analysis step on resolving conflicts is specific to the respective part of the analysis and, therefore, appears twice in the following description.

We started with open coding and utilized Krippendorff's alpha to find and discuss any conflicts before applying a combination of axial and selective coding.

Table 1: Results of Krippendroff's Alpha.

Codebook	Interview	Krippendorff's Alpha
RQ1	X23	$\alpha_{RQ1,X23} = 0.620$
RQ1	X24	$\alpha_{RQ1,X24} = 0.643$
RQ1	X23, X24	$\alpha_{RQ1,X23,X24} = 0.635$
RQ2, RQ3	X23	$\alpha_{RQ2,RQ3,X23} = 0.236$
RQ2, RQ3	X24	$\alpha_{RQ2,RQ3,X24} = 0.406$
RQ2, RQ3	X23, X24	$\alpha_{RQ2,RQ3,X23,X24} = 0.341$

Open Coding. Both leading and supporting researchers started by coding two interviews and building independent codebooks. Both met and merged their codebooks into one initial codebook while extensively discussing the reasoning for specific codes. Afterward, the leading researcher coded the remaining 23 interviews and added codes where they fit. Codebook changes were regularly discussed with the supporting researcher. To ensure a common understanding of the final codebook, and after finishing all 25 interviews, the supporting researcher also coded interview X23 with first-level codes.

Afterward, Krippendorff's alpha was calculated as shown in Table 1. Both researchers discussed all discrepancies and subsequently coded interview X24. Krippendorff's alpha was calculated again, and conflicts were discussed.

Conflicts Codebook RQ1. Krippendorff's alpha for codebook RQ1 only slightly changed between the first and the second iteration. We argue the stability is due to the technical nature of the topic. Minor conflicts occurred for all first-level codes, such as one person coding 'algorithm' while the other didn't. Overall, the majority of codes and our understanding of them overlapped.

Conflicts Codebook RQ2, RQ3. In contrast, Krippendorff's alpha for codebook RQ2, RQ3 improved significantly between the first and the second iteration. Particularly many discrepancies stem from coding 'expectations' and 'implications', due to their subjective interpretation. Additionally, as the implications and expectations of a system are closely intertwined, reaching a consensus on codes was sometimes challenging. To resolve, we discussed these codes extensively and defined them as follows: Anything that the participants anticipated from the CSS system was coded as 'expectations.' Any potential consequences resulting from the CSS system were coded as 'implications.' This substantially increased Krippendorff's alpha between the two iterations. We argue that the increase indicates growth in consensus between the researchers.

Axial and Selective Coding. After building the final codebooks and coding all interviews with these, we applied a combination of axial and selective coding to answer our research questions on each codebook independently.

To explain everything logically and soundly, we summarized some parts to form categories or subcategories. Both researchers discussed with each other how to report the results.

Saturation. Our sample reached saturation for higher-level codes for both codebooks within the first ten interviews. However, we discovered more nuances and diverse perspectives in lower-level

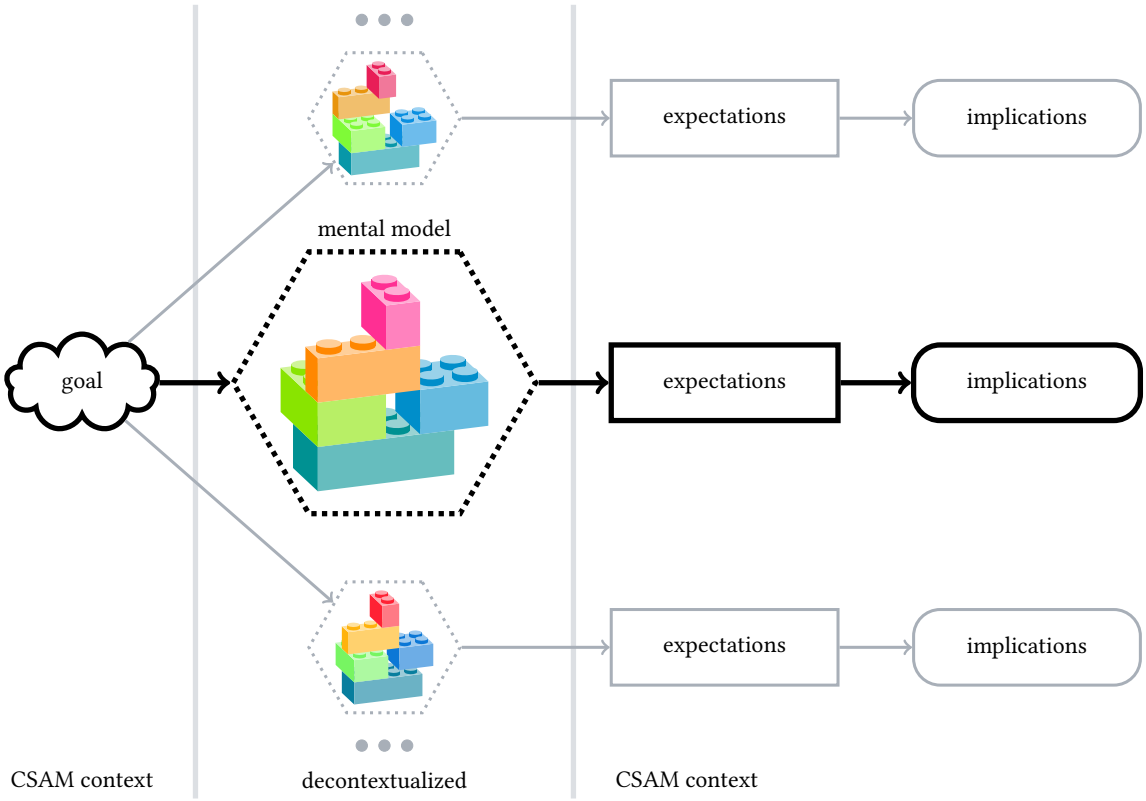

Figure 3: Overview of selecting a mental model of CSS to achieve a desired goal, where the selected mental model comes with implicit expectations, which lead to potential implications. Many different mental models can be chosen to achieve the goal, and each model accompanies different anticipated expectations and potential consequences.

codes through coding all interviews for which we might not have reached saturation.

4.6 Ethical Considerations

CSA is a very sensitive, emotional, and disturbing topic, which influenced our decision to focus on experts. Nonetheless, experts are also humans, and their mental, psychological, and physical health must be considered and is of utmost priority when discussing sensitive topics. Thus, we took some precautions: We looked up every participant before getting in touch and gave trigger warnings at several points during our interaction with participants. We asked specifically that referrals also work in the domain of CSA prevention or have another connection to CSS. We also had a list of helpful contacts and ensured participants knew they could terminate the interview at any point.

This research study was approved by the ethical review board (ERB) of the computer science department at Saarland University.

5 RESULTS

In this section, we discuss the findings of our interview study with expert stakeholders by defining the fundamental blocks of a CSS system: goals, mental models, expectations, and implications. Figure 3 gives an overview of the relationship between these. The goal informs the mental models, which then form the basis for differing expectations and implications. In the subsequent sections, we first discuss the system's goal. Next, we identify participants' mental models by decontextualizing CSS. Finally, we point out the expectations and potential implications of CSS by placing CSS in the CSAM context.

5.1 Goals

With the majority of communication channels being E2EE, it has become difficult for third parties to peer into the communication channel to analyze the contents of the transmission. Most of our participants agreed that the main goal of a CSS system is to scan for content on the client side before it is E2E-encrypted. This sustains the notion that the communication channel may be E2EE without any backdoors. The context in which the proposed system is deployed would determine the content to be scanned. In the context of CSAM, participants believed that the primary goal of a CSS system would be to prevent the dissemination of such content via messaging services.

However, participants also acknowledged that depending on the architecture of the mental model shown in Figure 4, the CSS system could be implemented to achieve a variety of goals. With the various types of content topics that can be scanned for, participants expected the goal(s) to be shifted fairly easily. X18 stated "*[...] it*

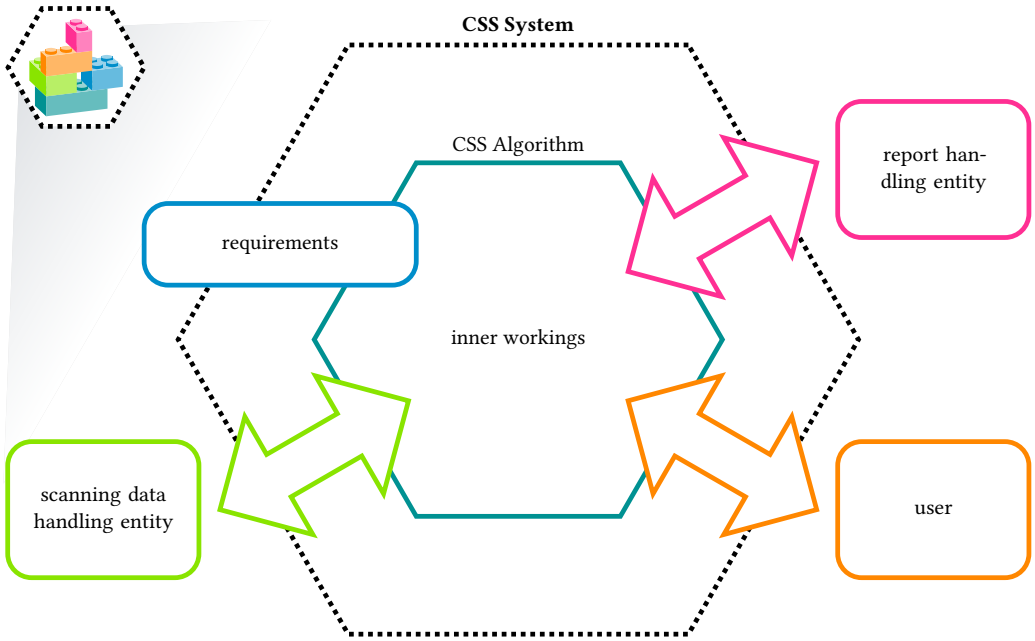

Figure 4: High-level overview of the mental models of CSS, which are comprised of inner workings that are influenced by requirements and have interactions with the scanning data handling entity, the user, and the report handling entity. A combination of all gives a full CSS system.

assesses whether the content is illegal or otherwise, more focused in recent discussions on detecting CSAM. But there's no reason why it couldn't be applied to other content."[X18] X12 further pointed out a recurring theme in the shifting of goals and expected the system to realign purposes, saying "*At the end of the day, somebody realizes, hey, that thing you built, we can also do something else with it.*"[X12]

5.2 Mental Models

The goal specifies the general direction of CSS and influences design system decisions which in turn influence a person's mental model of the system. In this section, we decontextualize CSS from the CSA debate to organize and summarize our participants' understanding of CSS systems into mental models.

5.2.1 Overview. Any form of CSS is part of a bigger context. Figure 4 gives a high-level overview of CSS as understood by our participants. The **CSS system** needs to interact with the **scanning data handling entity**, the **user**, and the **report handling entity**. The **CSS algorithm** encompasses **inner workings** defining *how*, *what*, *when*, and *where* something is scanned. This internal part is influenced by outside **requirements** that can be of a process, technical, policy, or legal nature. We will explain dependencies in the paragraphs below. For simplicity, the term "CSS algorithm" only describes the main functioning parts of the algorithm such as input, output, and crucial design decisions. In contrast, the term "CSS system" also encompasses interactions with the user, the scanning data handling entity, and the report handling entity.

5.2.2 CSS Algorithm. Figure 5 shows an overview of the **CSS algorithm** and any potential variations. Four non-negotiables are decided very early in the planning stage of a potential **CSS algorithm** that define any future architectural and engineering decisions. They are the **CSS algorithm**'s *how*, *what*, *when*, and *where*. The *when* and *where* are environmental decisions, while the *how* and *what* define the CSS algorithm. Furthermore, the algorithm has an *input* and *output*.

How. The *how* defines the concrete method of scanning for selected material. Many participants mentioned either a cryptographic or perceptual hashing algorithm, or a machine learning (ML) algorithm (cf. Section 2.3 for background information). However, some participants also mentioned that it can be any algorithm that allows the definition of search criteria.

Furthermore, the chosen *content familiarity* constrains the choice of the *algorithm type*. Hashing algorithms can only scan for known material, whereas ML algorithms can also scan for unknown material. Known material describes content that has already been vetted and deemed to be of a specific topic. For example, several participants mentioned that NCMEC maintains a database of confirmed CSAM, which could be used as comparison data.

What. The *what* of the CSS algorithm defines the *type of target material*, i.e., the media type of content that can be scanned for. According to most participants, a CSS algorithm can scan images, video, audio, or text data. X10 also mentioned scanning executables, X21 databases, and X13 behaviors. X13 focused on ML algorithms and said, "*[...] you can literally scan for every single bit on the user phone [...], it can be text, it can be multimedia content, [...] it can be certain behaviors, certain patterns of the user as well, like how the user interacts with the phone and the messenger.*"[X13] X18 and

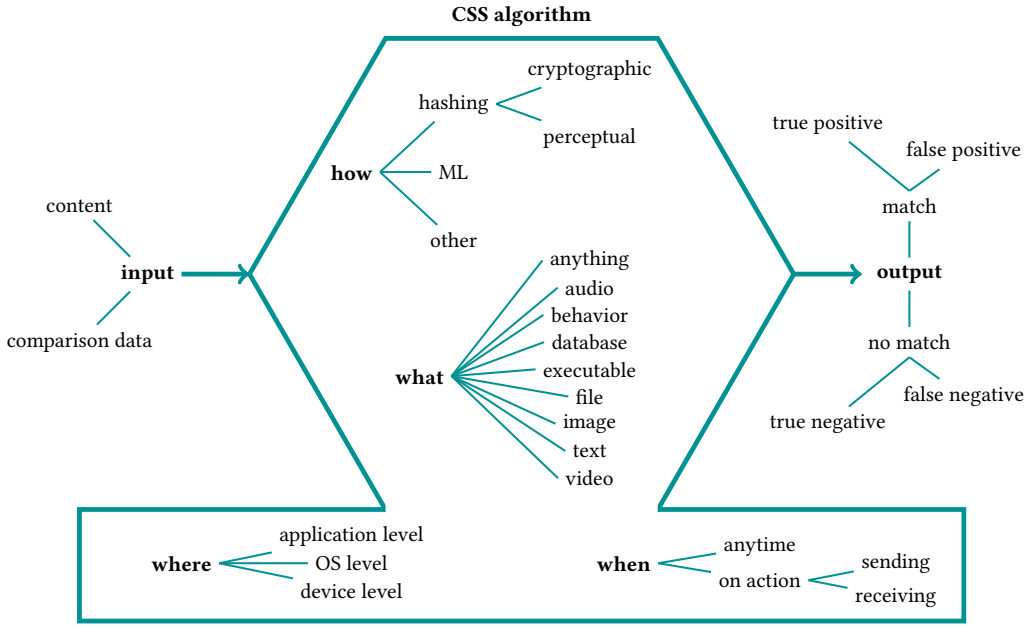

Figure 5: Overview of the inner workings of a CSS system. The *how* and *what* define the CSS algorithm while the *when* and *where* are environmental decisions, visualized through their layout. Furthermore, the *CSS algorithm* has an *inupt* and an *output*.

X23 focused on hashing algorithms and mentioned that anything hashable can be scanned for.

When. The CSS algorithm's *when* describes the *scanning schedule or trigger*, i.e., whether the algorithm scans at any time or if an action is needed to initiate scanning. This action could be receiving or sending data or creating and viewing files. Most participants mentioned uploading a file to a cloud, or messaging. X20 mentioned saving a file on the device as a trigger. Several participants specifically mentioned scanning before something is encrypted.

Where. The *where* of a CSS algorithm describes the *level* at which it will be running on the device. This could be the *application level*, the *operating system level*, or the *hardware device level*. All three come with benefits and challenges. Scanning on the *application level* is less invasive than scanning on the operating system or device level. However, it is easily circumvented by using a different application. Scanning on the *operating system or device level* is much more intrusive but more complicated to implement. The *operating system level* (e.g., system libraries) can be updated by both the operating system provider, i.e., Google or Apple, and the device manufacturer. The *device level* (e.g., the kernel) can only be updated by the device manufacturer. X20 described the distinction as well as benefits and drawbacks in detail:

> *There are three different levels. [Variant] one is in-app scanning. The app manufacturer can do that. [... In] Variant two [...] the service provider, i.e., the OS manufacturer, does that. This is the standard case. These are Apple, Google, and so on. But there is also a third version. This is the smartphone manufacturer [...] Huawei, Samsung. These are the ones who implement these security features. For example, this thing runs in the Trust Zone. Google also has no access to it. You can do Client-Side Scanning on all three levels. [...] The advantage of variant one is that you can control it better as a user. For example, if you know that WhatsApp does Client-Side Scanning, then you simply don't use WhatsApp. And I think the deeper, i.e., [variant] one, two, three, you go into the hierarchy, the more intrusive it is. On the other hand, it is also becoming technically more difficult and complex. Variant three is a little more complex than variant two. And variant two is more complex than variant one.* "[X20]

The actual matching can happen on the device or outside, either via a protocol such as Private Set Membership (see Section 2.4.1 for background information) or entirely on an external server.

Input. A CSS algorithm receives two types of *input*: the *comparison data* and the *content* it scans. The *comparison data* depends on the type of algorithm used. If a hashing algorithm is used, this input will be a database of hash values, as many of our participants mentioned. The hash values represent the known content the algorithm should search for. In contrast, for ML the comparison data input is a pre-trained classifier model, as described by a few participants.

Our participants mentioned different types of potential *content*. Section 5.3.2 covers this in more detail.

Output. A CSS algorithm can *output match* or *no match*, depending on the comparison of a hash or a threshold value of an ML classifier. Figure 5 covers the potential outputs.

True positives and true negatives are desired, while false positives and false negatives lead to complications. False positives are matches that the algorithm caught wrongly , i.e., they are benign

content, but the algorithm tagged them as unwanted. False negatives are real matches that were missed.

CCS algorithms are supposed to be deterministic, i.e., the same input produces the same output. In the case of hashing algorithms, the output is a definite yes or no decision, even if possibly false. Conversely, the output of a machine learning algorithm is a prediction (a probability of belonging to a certain group), as one participant touched on. The output value indicates the probability that the content input matches the comparison data.

5.2.3 Requirements. Any system that exists within a process and interacts with other components has **requirements** that need to be fulfilled. These are usually technical, process-related, policy-dependent, or legal. These **requirements** highly depend on the goal of the system. As we are trying to decontextualize the mental models from the debate around CSAM, we exclude a more detailed discussion from our definition of the mental models of CSS. CSAM-specific requirements will be mentioned in Section 5.3.

5.2.4 Interaction with the Scanning Data Handling Entity. A CSS algorithm needs comparison data to scan content, which is provided and maintained by the **scanning data handling entity**. This entity is responsible for the *goal* of what should be scanned for, which hinges on the *context* and the *content topic* and thus influences the *type of content* to be scanned. In turn, the *type of content* is an essential factor in choosing an appropriate *algorithm type* and building the algorithm. In most cases of CSS, the **scanning data handling entity** is a third party. Since the *comparison data* is highly *context*-specific and depends on the *content topic*, participants gave an overview of applications they have seen in the wild and applications they expect. Applications that our participants mentioned included phone number hashing, antivirus scanning, and surveillance of criminal suspects via wiretapping. We have listed some more real-world applications from the literature in Section 2.5.1. Applications expected by our participants will be covered in Section 5.3.3.

5.2.5 Interaction with the User. There are specific points in the process of deploying and using a CSS algorithm on a device that may require *interaction with the user*. These include consent from the user, disclosure of a CSS algorithm running, legalities, whether the algorithm is beneficial to the user, is targeted, or privacy-preserving.

Benefit. In a CCS system, the user can either be the *beneficiary* or the *adversary*. For example, Participant X10 mentioned that a smartphone user would benefit from an antivirus program scanning for malware. X24 gives an example for an opposite setup: "*[...] in a particular case of Client-Side Scanning [referring to the version Apple wanted to introduce in 2021], it is not just that there are certain things that I cannot do on my device, but actually things that I do on my device will be actively used against me when this device is essentially reporting on me.*"[X24]

Consent. A user can either *voluntarily* or *mandatorily* use CSS. *Voluntary* use can be realized via opt-in or opt-out models. Participant X20 describes an opt-out model for the antivirus scanner *Microsoft Defender Antivirus*, which is preinstalled on Windows PCs nowadays.

> "*There is Microsoft Defender. And, to my knowledge, it scans every file. Then there are always these warnings when you download data via a network drive and want to run it. And I believe that these files are always scanned by Windows Defender in any case. I think you can configure that away. So, I have quite a lot of malware on my system because we use it in teaching. And that's why I have to be careful that Windows Defender doesn't quarantine files all the time.*"[X20]

Mandatory use means that to use a service or a device, the execution of a CSS algorithm has to be tolerated and cannot be declined.

Disclosure and Notification. Disclosure describes whether the device user is informed of the deployment of CSS before it is used. Conversely, a *notification* denotes informing the user after a (potential) match. This highly depends on the legal aspects of the goal of the CSS system, e.g., whether a user is suspected of breaking the law due to a match. However, if a *notification* is not given immediately after a match, our participants said that it should still happen at some point for transparency reasons. Most of what our participants said about *disclosures* and *notifications* was contextualized by the CSAM debate. Thus, further details can be found in Section 5.3.2.

Legal Basis. The decision on the *legal basis* defines whether a warrant is needed to apply CSS or not. A few participants mentioned the need for a warrant when wiretapping. Others said that any scanning with the intention of reporting a person due to potential criminal activity needs to be done with a warrant based on the current legal situation.

Target. Some participants differentiated between CSS systems applied to a very *targeted clientele* or to the *general public*. An example of the former is using a CSS algorithm to surveil a specific person. Conversely, if a messaging service routinely hashes the contents of its users' phone books with the user's permission to match contacts using the same service, it would be a non-targeted application.

Privacy. Our participants said that some CSS systems run solely on an end user's device, while others scan content and send certain information to a third party. For example, if a messaging service hashes the contents of its users' phone books, then these hashes need to be sent to a server and compared to the hashes of phone numbers for the service to work and be beneficial. This can lead to a breach of privacy for the user.

5.2.6 Interaction with Report Handling Entity. After the CSS algorithm assumes it has found a match, it might be passed to a **report-handling component**. This interaction can happen with a different device component or a third party outside the device. Expected consequences within the CSAM context regarding reporting are discussed in Section 5.3.2

Threshold. For some applications of CSS, it is necessary to have a *threshold* that limits when an interaction with another component happens. For example, some participants mentioned that when Apple introduced its plans to deploy its CSS system NeuralHash, it included a threshold to pass before a report was made to Apple [27].

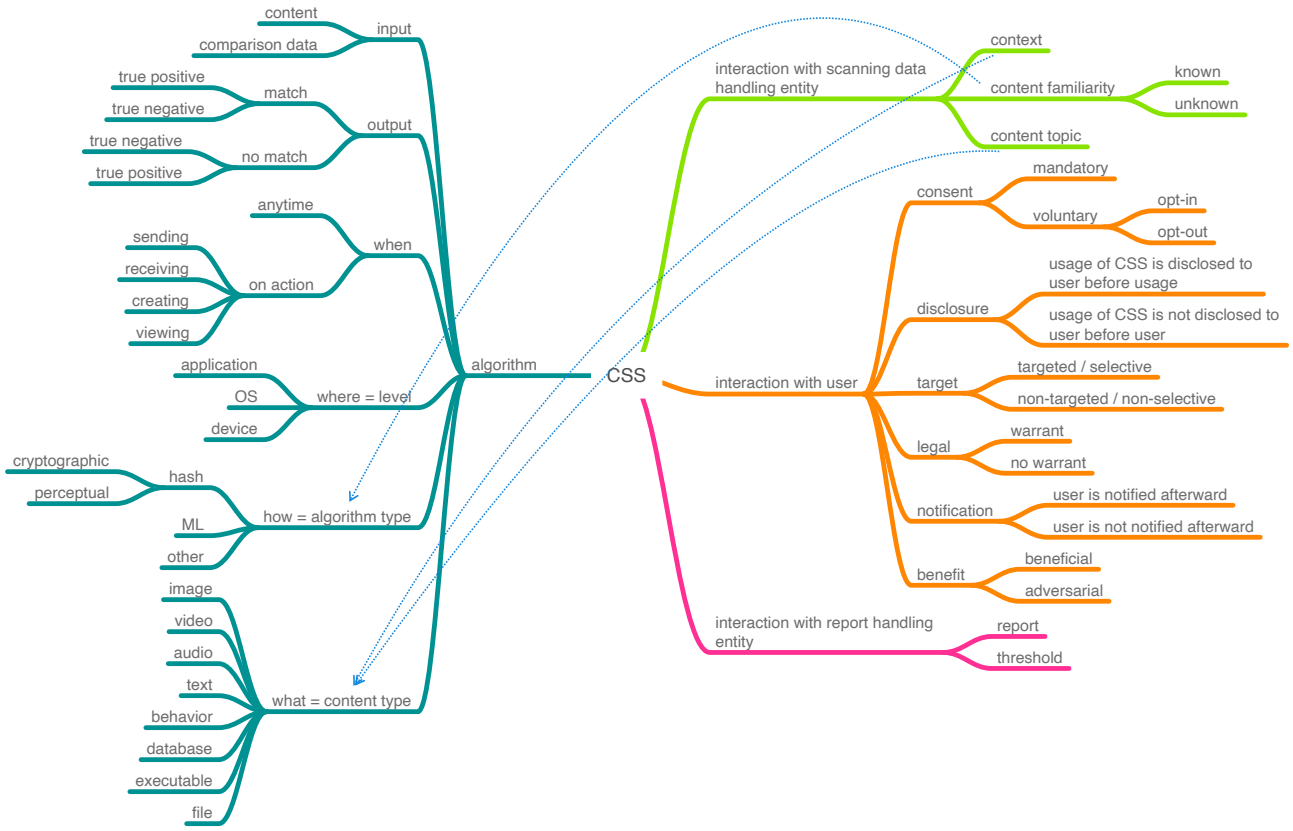

Figure 6: Overview of possible dimensions of CSS as described by our participants. The blue dotted lines depict interdependencies between dimensions.

Report. Some CSS applications aim to *report* information to a third party. This *report* can depend on a positive evaluation by the CSS algorithm and whether a threshold is passed. For example, several participants mentioned that if a CSS system is built to detect CSAM and needs to inform the service provider in case of a match, a report needs to be sent. This *report* can include various information. However, depending on national law, certain things may be prohibited. Participant X01 stated, "*they [referring to Apple as the service provider] can't legally attach that kind of imagery [referring to CSAM] and send it to other people, even themselves.*"X01

5.2.7 The Dimensions of CSS. The previous sections have explained the mental models of our participants in detail. They usually do not only have a one-dimensional mental model of CSS, but a multi-dimensional one that depends on the goal of the system and the building blocks used to explain a specific model. A visual overview of all dimensions of CSS is shown in Figure 6. The first level gives categories similar to the ones discussed above, while the second level describes the dimensions. All subsequent levels give different options for the dimensions. A CSS system can be defined by selecting one or more options for each dimension while minding dependencies, e.g., scanning for unknown material requires the use of an ML algorithm. Each possible combination describes a different mental model of CSS.

5.2.8 Expert's Understanding of CSS. We interviewed experts from different backgrounds with different technical background knowledge levels. Interestingly, the better the technical background knowledge of a participant, the more they decontextualized CSS from the debate of fighting CSA. For them, CSS was only a means to an end, and often, they said it was a terrible solution. Participant X10 described it as:

> "*In many ways, I think it [CSS] is a bad solution that's been created by a different problem.*"X10

They referred to CSS as a solution to the problem that big tech companies introduce E2EE on communication channels to protect themselves while disregarding potential bad consequences for, e.g., minors. On another note, participant X24 mentioned that:

> "*I think an underlying problem of some of the Client-Side Scanning is this belief that technology will solve our societal problems.*"X24

They described the entire discussion as not solution-oriented.

If participants lacked technical background, they saw CSS as a potential solution while acknowledging that they might not have enough knowledge to judge this properly, and it was harder for them to distinguish between CSS and its goal. This shows the importance of our structuring into dimensions of experts' mental models

because it can vastly improve stakeholders' understanding of the technical dimensions of Client-Side Scanning.

5.3 Expectations

This section delves into participants' expectations of using CSS to scan for CSAM, as shown in Figure 7. We explore adoption and reception expectations of CSS, its anticipated implementation, expected usage, and potential challenges that may arise in its operation.

5.3.1 Adoption and Reception. Adopting and deploying technology is tied directly to how well the general population receives it. A few participants expected some form of CSS to be adopted soon due to the ever-rising circulation of CSAM content online and CSS being proposed as a way to counter it. According to X02, as is the case with most technologies, once a CSS system is proven to work, it is inevitably deployed. They anticipate that the next few years will bring frequent deployments of CSS. Along the same lines, X14 suggested that CSS would be implemented initially for sensitive issues that are universally condemned, such as child sexual abuse or potential terrorist activities. X07 pointed to the lack of alternatives as a likely adoption criterion and stated *"[...] the lack of a good alternative then makes it attractive."*[X07] They explained that in the absence of good alternatives, taking any action rather than doing nothing at all can seem appealing.

Some participants outlined potential ways to ease the expected adoption of a CSS system in the CSAM context, with X05 stating, *"I think if CSS were to be adopted more broadly, I would like to ensure that there is a mechanism for certifying the provenance (source) of whatever we're scanning against."*[X05] According to X09, the technology may serve as a viable option in specific scenarios as a band-aid measure. X13 echoed this sentiment, saying that they anticipated selective use-case-based adoption of the system. However, most participants strongly objected to implementing a CSS system, likening it to a *"massive surveillance system with serious privacy implications."*[X13] X02 hopes the point will never be reached where CSS is installed on people's devices without consent. X17 rebuked CSS, saying they were very critical of the technology and that they believed the effects could be catastrophic if widely implemented. Some participants drew parallels between the public response to Apple's NeuralHash announcement and how they expected people to raise their voices and come out against the technology. X01 predicted a similar fate for a CSS system, stating, *"I think that there are a lot of elements of civil society and corporations, etc. that will fight back very hard against any proposals [...]."*[X01] X22 expressed concerns about the validity of implementing a CSS system based on legal and fundamental rights. They noted that the regulations concerning the proposed system are too vague to determine which rights are being restricted, which is highly problematic from a legal standpoint. Finally, X11 stated their opposition to introducing the system but acknowledged that if implemented, they would demand CSS to work without producing any false positives. They also emphasized the need for the system not to be abused for other purposes.

Summary

The short answer to RQ1 is that people do not only have one mental model for CSS. It depends on certain factors like the type of input and goal or the level of technological understanding. A mental model of CSS can thus change depending on the context within which it will be applied. This makes communication about specific versions of CSS important.

5.3.2 Implementation. Participants described how they expected the CSS system to be implemented within the CSAM context. When asked to explain their expected implementation to a non-technical person, some participants used real-world analogies. When describing the system, some participants used the 'post' analogy, where, before a person sends a letter, somebody opens the envelope to check its contents. The content checking does not happen somewhere during the transport of the letter but rather just after it is written. If the content inside the envelope is benign, nothing happens. However, if the content is nefarious, a report is sent out. Other participants expected the system to function like a 'wiretap' or 'snitch,' looking up everything to check if something meets the criteria. X08 compared the system to speed cameras, which monitor all cars, but are only triggered by those going too fast. A few participants used 'bank' and 'walls of a room' analogies to describe how the system would fit into established privacy norms.

We organized participants' expectations for implementing a CSS system scanning for CSAM content into categories based on the mental model dimensions from Section 5.2, which we now elaborate.

Disclosure. Some participants expected people to be informed that a CSS system scanning for CSAM was active on their device before they started using it and that it could result in severe consequences. They also demanded that people get a disclosure of consequences.

Content. The scanned content can be split by content type and content topic as mentioned in Section 5.2.2. In the CSAM context, participants anticipated the content type to be mostly for image, audio, video, and text formats. Apart from CSAM, the expected content topics were abortion, activism, commercial, copyright infringement, criminal, grooming, hate speech, human rights, malware, oppressive content, political content, religious content, sexual content, and terrorism.

Algorithm. As mentioned in Section 5.2.2, depending on whether the system would only scan for known CSAM or also for unknown CSAM, participants expected the system to be implemented either using hashing or ML. In the case of hashing, some participants expected people to receive a database of known CSAM hashes on their devices, which would be regularly updated to account for the increasing amount of CSAM. Content would then be checked against this hash database.

Scanning. Content is primarily checked when sending and receiving data or when a downloaded file triggers the scanning, as mentioned in Section 5.2.2. When scanning for CSAM, two types of algorithms can be used: hashing and ML. Hashing algorithms generate hashes of the content and match them to a known CSAM hash database, while ML algorithms leverage a model trained on CSAM to classify the content.

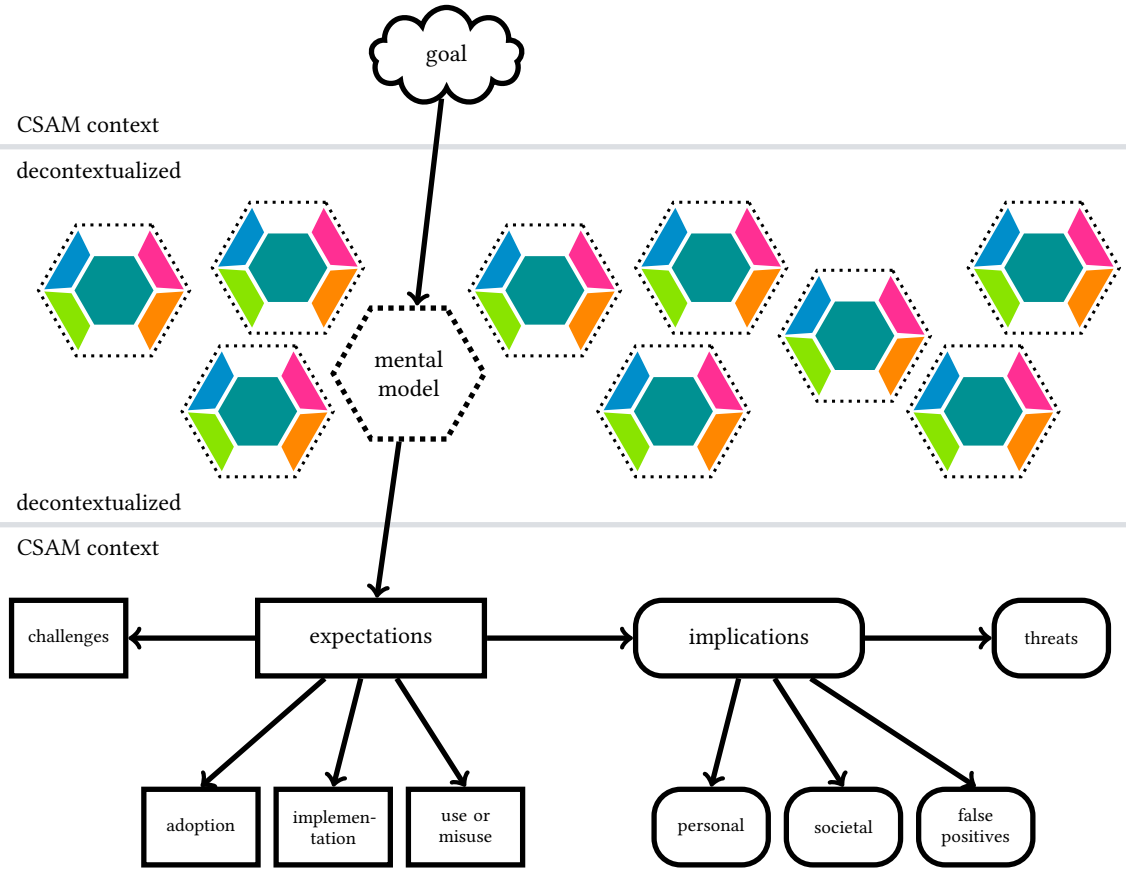

Figure 7: Detailed overview of the relationship between mental models, expectations, and implications of a CSS system in the CSAM context.

Match. If the input did not match the targeted CSAM, participants expected the current action (e.g., communication) to proceed unimpeded. Conversely, if the content matched the targeted CSAM, some participants expected it to be stopped from transmission and potentially deleted from the device, depending on the implementation. Other participants expected reports to be sent out.

Notification. In the CSAM context, most participants expected notification of a match, except for suspects or those with large amounts of CSAM to prevent evidence destruction. To aid transparency, participants demanded people to be always notified of a match but acknowledged that the notifications could be potentially delayed.

Report. Some participants expected reports to be sent out to service providers or law enforcement if a match was found with the CSAM database. Several participants expressed the need for a human review of the reported content to prevent false positives and reduce inaccurate reports that could overwhelm the system. A few participants did expect a certain threshold to be met before reporting people. Based on the implementation, some participants anticipated that if a report was filed against an individual, their personal accounts would be completely shut down. Some participants expected the individual in question to face legal consequences and have law enforcement officials show up at their doorstep.

Participants also addressed various factors that could impact the implementation of a CSS system. X07 recognized that the system could be tailored to consider different cultural and religious beliefs to determine if content is illegal. X01 anticipated that there could be regional differences in public opinion about the system, stating *"it wouldn't necessarily be consistent from place to place."*[X01] X02 predicted that companies would find ways to profit from the system and capitalize on its value. Thus, when implementing such a system, X05 emphasized transparency and auditability and called for a *"thoughtful, minimal, and very tightly bounded system."*[X05] In general, participants advocated for a risk-based approach and advised against a one-size-fits-all implementation.

5.3.3 Use and Misuse. A CSS system is technologically only bound by the content type, not the content topic. This modularity of the content topic allows the system to be used for multiple purposes. Participants listed different applications that CSS could be used for when discussing the dimensions of their mental model in Section 5.2.4. When appropriately implemented and limited to the

domain of CSAM, some participants expected the system to help reduce dissemination. X02 stated, **"***On a good level, it could be used to stop drug transactions, child sex trafficking, et cetera.***"** X05 added that a cryptographic matching-based system could be used to block dissemination, safeguard victims against revictimization, and identify and potentially prosecute bad actors.

In contrast, many participants expected the system's domain to be expanded to suppress non-CSAM content, referenced in Section 5.3.2. X03 pointed out, **"***It's not linked to a particular domain at all, which is dangerous.***"**[X03] X24 and X07 stated that non-democratic countries could misuse the system to scan for political activism, sexuality, and religion.

5.3.4 Challenges. Participants expected certain challenges associated with deploying a CSS system, as discussed below.

Differentiating CSAM from non-CSAM context. For a CSS system to perform effectively, it must accurately distinguish between CSAM and non-CSAM. Failure to do so could result in a significant number of false positives and negatives. Participants acknowledged that defining CSAM can be challenging and requires extensive contextual knowledge. To illustrate this challenge, several participants highlighted a real-world example where an innocent individual lost their account by sharing a medical image of their child [46]. They argued that the challenge would be even more significant if the CSS system were also to identify unknown CSAM. Some participants noted that differentiating between CSAM and non-CSAM contexts can be arduous in cases of consensual sharing of nude content among teenagers and nude photos of children within the family.

Integrity of the CSS system. Deploying a CSS system in the real world poses a challenge in ensuring its robustness and integrity. Participants discussed the possibility of malicious actors attempting to circumvent the system by using non-CSS-compliant apps or devices and continue CSAM trading. To ensure the integrity of the system, participants expected the reliability of algorithms and the database, its technical robustness, and ease of scalability to be the key factors. X07 believed that more reliable algorithms would lead to fewer false positives and expected the current scanning algorithms not to work because they were not robust enough. X04 highlighted the database of hashes as a single point of failure and warned against making it susceptible to manipulation by malicious actors. X16 mentioned the substantial logistical effort required to maintain and frequently update this database. A few participants also discussed the expected challenges in scaling the system, considering the variation in people's device capabilities when implementing it.

Overwhelmed legal system. With the large-scale deployment of a CSS system, people are bound to be reported. The volume of reports is compounded by the large number of people using applications to communicate online. Participants thus expected the human reviewers, law enforcement, and judiciary to become overwhelmed. They also pointed out that if the system also scanned for unknown CSAM, the number of false positives would rise exponentially, thereby rendering the legal systems dysfunctional.

Summary

In response to RQ2, participants had mixed expectations about the deployment of CSS. Some anticipated backlash and others foresaw its misuse. Mental models were deeply interconnected with the anticipations of how the system would be implemented. Regardless, everyone expected technical challenges once CSS systems were deployed.

5.4 Implications

Participants' expectations influence the potential implications of using a CSS system in the CSAM context, as shown in Figure 7. This section examines the personal and societal consequences, impact of false positives, and potential threats associated with the system.

5.4.1 Personal. During the interview, participants shared their views on the personal consequences and effects of a CSS system. Some participants were concerned about the potential privacy breach resulting from the human review layer. They also considered the emotional impact such a system could have on people, imagining that individuals may feel confused and scared. Additionally, participants highlighted that the loss of an account is an implication often overlooked and underestimated. However, X10 argued that account loss is not a consequence of CSAM but rather the service provider's process of reinstating accounts. Some participants also touched upon the reputational harm that can occur when someone is falsely accused, leading to a negative perception of the individual by society. Generally speaking, participants feared that individuals would feel like they were being watched, and would worry about how much of their communication was private. To protect themselves, they would repeatedly explain themselves to the authorities. As a result, some participants expressed skepticism towards the technology, which could lead to their withdrawal from digital spaces.

5.4.2 Societal. Participants expressed their opinions on the overarching effects of a CSS system and expressed concern about the potentially negative impact of the system on society. They feared that people would not have enough information about the issue and how to defend against it, which could lead to a *chilling effect* [56]. X04 believed that the chilling effect would make people feel constantly watched and scrutinized, hindering free social communication. Others pointed out that a lack of public discourse would only add to the chilling effect. X01 worried that certain groups in society would be at greater risk if the system's domain were expanded. X13 added that it could lead to **"***mind control and propaganda.***"**[X13] Some participants also noted that the system would result in people having to self-edit and self-censor regularly, ultimately leading to a decline in societal empowerment and creativity. Contrarily, X08 argued that the system would make it harder for sex offenders to engage in criminal activity and potentially discourage new offenders.

5.4.3 False Positives. Several participants expressed concern that false reports would divert resources and overwhelm the legal system. They also believed that false positives would cause mistrust in the CSS system to grow. X08 felt that people would perceive the system as arbitrary and unjust, leading to resentment. Conversely,

X10 argued that false positives occur in all aspects of life and can be mitigated by a properly functioning legal system. X01 added that reducing false positives would positively impact the system by reducing confusion and fear among people.

5.4.4 Threats. Participants believed that a CSS system could pose a threat to people in authoritarian countries. They feared that corresponding authorities could use the system to target individuals who oppose the state or hold differing opinions. They also identified a threat model in which an individual's account could be wrongfully shut down if they receive either CSAM or falsely-detected CSAM. X03 suspected "an army of trolls" could be leveraged to spam CSAM content to cross the minimum threshold of hits required to generate a report. X05 raised concerns that abusive spouses could use the system to spy on their partners. Participants were also worried about the possibility of malicious actors inserting non-CSAM content into the CSAM database to expand the scanning domain. For the most part, participants agreed that the system would increase the attack surface, leaving devices more vulnerable to exploitation.

5.4.5 Stakeholder Groups. Apart from analyzing the general implications of a CSS system, we also examined its impact on specific stakeholder groups.

Children and Teenagers. It is crucial to understand the impact of a CSS system on children, as it heavily focuses on them. Children have their own privacy rights that extend into their communication [5]. They also have a right to free will [4]. Children often experience harm from individuals within their social environment. Under such premises, participants were uncertain whether the system offered adequate protection. They also believed that as part of youth culture, children and teenagers frequently exchange intimate images among themselves with consent. This is crucial for youth to understand their bodies and sexuality, as well as develop independent opinions [47]. Criminalizing this behavior could be detrimental to their growth and may cause them to become rebellious. Some participants feared that a CSS system could leave children with difficult family relationships or communication barriers, such as LGBTQ children, even more vulnerable. Several participants felt that such a system would end up punishing more children and teenagers than actual perpetrators. To best protect the interests of children, X05 suggested that the system provide appropriate educational resources and warnings about sharing content online. They remarked "*[...] we're not calling the cops on you, but there are bad folks out there, let's brush you up on why you probably shouldn't be doing this.*"[X05] X10 advocated for a system that prevents the dissemination of known CSAM, thereby preventing revictimization.

Companies and Service Providers. Companies and service providers tasked with implementing and maintaining a CSS system have a challenge at hand. X03 acknowledged that companies would be pressured to have an extra-safe system or risk extreme backlash. The system would have to be consistent with the laws and policies of the region in which it is deployed. Therefore, stakeholders must determine the cost of business and where their compliance lies with the laws and policies of the region. Some participants expressed concerns that countries may pressure stakeholders with existing infrastructure to manipulate the hash database. X07 also highlighted these stakeholders' moral and ethical responsibilities towards protecting people. They further emphasized the conflicting situation these stakeholders face, stating, "*The same users who now have an interest in not being prosecuted themselves naturally have an interest in making sure that illegal activities can be prosecuted and investigated.*"[X07]

Summary

In response to RQ3, participants expressed concern about the personal and societal implications of a CSS system. It could harm individuals emotionally and reputationally, hinder free communication, and cause self-censorship. Participants feared that children and teenagers are particularly vulnerable to these adverse effects. Furthermore, false positives can irrevocably erode trust in the system.

6 DISCUSSION

6.1 Technology Working for the User vs. Technology Working Against the User

In Section 5.2.5 while answering RQ1, we touched on the user being a beneficiary or adversary of CSS usage. Especially in versions of CSS where the user is seen as a potential adversary, the use must be discussed in depth for two reasons. First, placing the user into an adversarial role can lead to a contentious reaction by the user before any potential benefit can be discussed. Second, by making the user an adversary, the technology will be more prone to misuse because the user is more inclined to circumvent the technology.

We have seen the first reason discussed extensively within the public CSAM debate, but for a holistic discussion, the second reason must also be considered in the public discourse.

Nonetheless, in our opinion, technology where the user is seen as an adversary should be used sparingly and targeted to make sure it does not run into the above-mentioned problems.

6.2 The Least Invasive Approach

While a CSS system with a built-in reporting mechanism carries the potential for grave consequences, there are less intrusive methods that can be employed to prevent the spread of CSAM. One such measure proposed by X05 is to use cryptographic hashes to scan for known CSAM. In case of a match, instead of reporting to law enforcement, softer interventions would be applied, which include blocking the matched content from dissemination and deleting it from the device. The use of cryptographic hashes ensures that only identical matches are considered, avoiding any false positives.

Another less invasive approach also proposed by our participants is how Apple's "Communications Safety" feature [28] intervenes when children try to send or receive any content containing nudity and provides them with helpful resources. This feature scans not only still images but also video content. Additionally, Apple has developed a "Sensitive Content Warning" feature [29] that prevents the automatic display of nude content and allows users to choose whether they want to view such content. Both these features are turned on by default for child accounts. Apple emphasizes that these features are intended only to prevent people from viewing nude content and not intended for reporting.

We argue that it is imperative to implement less invasive measures to provide individuals with protection against inadvertent nudity while ensuring privacy is maintained, thereby mitigating extreme consequences. Future work should focus on improving the presentation of intervention messages, identifying helpful resources, and determining useful self-reporting measures.

6.3 The Long-term Threat of Scanning

Our findings from Section 5.4, outlining potential implications of CSS (RQ3), suggest that deploying a CSS system to scan for content could have long-term effects. Although the primary purpose of the system would be to detect CSAM, there is a possibility that its domain could be expanded to scan for other content. This raises concerns of a *"slippery slope"* situation, where adopting the CSS system could set a precedent and ultimately normalize a lower level of privacy, leading to a general lack of trust among people in the long run. The possibility of non-democratic states using this system to perform mass surveillance and manipulate their population is also alarming. Some participants also acknowledged that there is no way to know what future technology will be capable of. Moreover, there is a lack of understanding regarding the evolution of the CSS system after its adoption. While it is currently suggested as a way to safeguard vulnerable groups, there is a likelihood that it could cause more harm than good in the long term.

6.4 Technology is not the Solution to Societal Problems

In Section 5.2.8, we briefly touched on the fact that the more technical understanding the participants had, the more they decontextualized CSS. Decontextualizing a technology from its purpose is essential in grasping its innovative capability, just like its potential for harm. Technology was, is, and will be essential in solving societal problems. However, it will not be the solution to those problems. One participant mentioned the following:

> *"I think an underlying problem of some of the Client-Side Scanning is this belief that technology will solve our societal problems. It won't. [...] In this particular case, I find it really problematic that there's a lot of technology being pushed with pretty severe ramifications and potential for abuse. Whereas [it is] also not clear whether it actually will be particularly helpful. [...] What fascinates me about this is [...] the drive for techno-solutionism. So to believe that we can solve this serious societal problem of child abuse through the use of technology without necessarily a lot of discussions about how effective such technology would be in actually solving such problems."* [X24]

We are aware that this is not solely a CSS problem. Nonetheless, we need to reiterate this because it is crucial to know when planning, conceptualizing, and building any technology, especially for people without a technical background. On that note, future work needs to look into technological solutions that help fight the dissemination of CSAM and revictimization while not being privacy intrusive to the general population. This includes the consideration of any potential contexts in which the dissemination of CSAM should be fought. One example of such a tool is Apple's "Communications Safety" feature [28], mentioned above.

7 LIMITATIONS

We used qualitative methods to gather data. On one hand, we only captured the views of our interviewees, and the results are not necessarily generalizable. On the other hand, we may have missed some stakeholders due to our interpretation of the topic or inability to reach them for an interview.

Due to all authors being computer scientists, we are privacy conscious and trained to think of worst-case scenarios. This might have introduced bias during any part of the study. Nonetheless, we tried our best to stay objective and discussed repeatedly whether our views influenced our judgment.

Lastly, given that Germany is a privacy-conscious country and that a significant portion of our participants were German, there is a potential for geographic and cultural bias.

8 CONCLUSION

Client-Side Scanning systems can be created to detect any arbitrary content with unspecified reliability. Proposed safeguards are policy-wise, not technological. Zero percent false positive rates will never exist, and even a minuscule false positive rate over a vast number of benign cases can lead to substantially more suspects than actual perpetrators. The discussion of CSS as a measure to limit the dissemination of CSAM is currently in a standstill with hardened fronts. Proponents would like to see a tightly bound CSS to fight CSAM, while opponents always focus on the worst versions of CSS. The discussion on fighting the dissemination of CSAM sometimes loses track of the fight on the origin of such material. The best way to help the fight against CSA is to research different perspectives that will aid the cause while being as non-invasive to privacy as possible. The dimensions of the mental models presented in this paper are foundational to any solution-oriented discussion.

ACKNOWLEDGMENTS

We sincerely thank our participants for their time and expertise, among them Lloyd Richardson from the Canadian Centre for Child Protection (C3P). We also thank the anonymous reviewers for their valuable and constructive feedback, which was very useful in improving our paper.

REFERENCES

[1] [n. d.]. Chat Control: The EU's CSEM Scanner Proposal. Retrieved 2023-09-12 from https://www.patrick-breyer.de/en/posts/chat-control/
[2] [n. d.]. NCMEC Child Sexual Abuse Material. Retrieved 2023-09-12 from http://www.missingkids.org/theissues/csam.html
[3] [n. d.]. Thorn Child Pornography and Sexual Abuse Statistics. Retrieved 2023-09-12 from https://www.thorn.org/child-pornography-and-abuse-statistics/
[4] 2012. Charter of Fundamental Rights of the European Union. Retrieved 2023-09-12 from http://data.europa.eu/eli/treaty/char_2012/oj/eng
[5] 2013. Children's Online Privacy Protection Rule ("COPPA"). Retrieved 2023-09-12 from https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa
[6] 2016. Investigatory Powers Act 2016. Retrieved 2023-09-12 from https://www.legislation.gov.uk/ukpga/2016/25/contents/enacted
[7] 2019. Attorney General Bill Barr Will Ask Zuckerberg To Halt Plans For End-To-End Encryption Across Facebook's Apps. Retrieved 2023-09-12 from https://www.buzzfeednews.com/article/ryanmac/bill-barr-facebook-letter-halt-encryption
[8] 2020. Curbing the Surge in Online Child Abuse: The Dual Role of Digital Technology in Fighting and Facilitating Its Proliferation | Think Tank | European Parliament. Retrieved 2023-09-12 from https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2020)659360

[9] 2020. EU Strategy for a More Effective Fight against Child Sexual Abuse. Retrieved 2023-09-12 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0607

[10] 2020. EU Technical Solutions to Detect Child Sexual Abuse in E2EE Communications. Retrieved 2023-09-12 from https://www.politico.eu/wp-content/uploads/2020/09/SKM_C45820090717470-1_new.pdf

[11] 2020. EU Temporary Derogation of Directive. Retrieved 2023-09-12 from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0568

[12] 2022. Assessment of Proposal for a Regulation Laying down Rules to Prevent and Combat Child Sexual Abuse | Think Tank | European Parliament. Retrieved 2023-09-12 from https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2023)740248

[13] 2022. EU Proposal for Laying down Rules to Prevent and Combat Child Sexual Abuse. Retrieved 2023-09-12 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A209%3AFIN

[14] 2022. Online Safety Bill - Parliamentary Bills - UK Parliament. Retrieved 2023-09-12 from https://bills.parliament.uk/bills/3137

[15] 2023. Chat Control - The End of the Privacy of Digital Correspondence. Retrieved 2023-09-07 from https://www.patrick-breyer.de/en/posts/messaging-and-chat-control/

[16] 2023. CyberTipline 2022 Report. Retrieved 2023-09-04 from https://www.missingkids.org/cybertiplinedata

[17] 2023. EURion Constellation. Retrieved 2023-09-12 from https://en.wikipedia.org/w/index.php?title=EURion_constellation&oldid=1174760675

[18] 2023. Google Image Search. Retrieved 2023-09-04 from https://www.google.com/imghp

[19] 2023. Microsoft Bing. Retrieved 2023-09-04 from https://www.bing.com/images/

[20] 2023. Text - S.1207 - 118th Congress (2023-2024): EARN IT Act of 2023. Retrieved 2023-09-12 from https://www.congress.gov/bill/118th-congress/senate-bill/1207/text

[21] 2023. TinEye. Retrieved 2023-09-04 from https://tineye.com/

[22] Hal Abelson, Ross Anderson, Steven M. Bellovin, Josh Benaloh, Matt Blaze, Jon Callas, Whitfield Diffie, Susan Landau, Peter G. Neumann, Ronald L. Rivest, Jeffrey I. Schiller, Bruce Schneier, Vanessa Teague, and Carmela Troncoso. 2021. Bugs in Our Pockets: The Risks of Client-Side Scanning. *arXiv:2110.07450 [cs]* (Oct. 2021). arXiv:2110.07450 [cs] Retrieved 2023-09-12 from http://arxiv.org/abs/2110.07450

[23] Home Affairs. 2018. Telecommunications and Other Legislation Amendment (Assistance and Access) Act 2018. Retrieved 2023-09-12 from https://www.legislation.gov.au/Details/C2018A00148/Html/Text,http://www.legislation.gov.au/Details/C2018A00148

[24] Apple. 2016. Customer Letter: The San Bernardino Case. Retrieved 2023-09-12 from http://www.apple.com/customer-letter/

[25] Apple. 2021. Child Safety. Retrieved 2023-09-12 from https://www.apple.com/child-safety/

[26] Apple. 2021. CSAM Detection - Technical Summary. Retrieved 2023-09-12 from https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf

[27] Apple. 2021. Security Threat Model Review of Apple's Child Safety Features. Retrieved 2023-12-01 from https://www.apple.com/child-safety/pdf/Security_Threat_Model_Review_of_Apple_Child_Safety_Features.pdf

[28] Apple. 2023. About Communication Safety in Messages. Retrieved 2023-09-12 from https://support.apple.com/en-us/HT212850

[29] Apple. 2023. Sensitive Content Analysis. Retrieved 2023-09-12 from https://developer.apple.com/documentation/sensitivecontentanalysis

[30] Benny Chor, Eyal Kushilevitz, Oded Goldreich, and Madhu Sudan. 1998. Private Information Retrieval. *J. ACM* 45, 6 (Nov. 1998), 965–981. https://doi.org/10.1145/293347.293350

[31] Larissa S. Christensen, Susan Rayment-McHugh, Tim Prenzler, Yi-Ning Chiu, and Julianne Webster. 2021. The Theory and Evidence behind Law Enforcement Strategies That Combat Child Sexual Abuse Material. *International Journal of Police Science & Management* 23, 4 (Dec. 2021), 392–405. https://doi.org/10.1177/14613557211026935

[32] Danijela Frangež, Anton Toni Klančnik, Mojca Žagar Karer, Bjørn-Erik Ludvigsen, Jarosław Kończyk, Fernando Ruiz Perez, Mikko Veijalainen, and Maurine Lewin. 2015. The Importance of Terminology Related to Child Sexual Exploitation. In *Revija Za Kriminalistiko in Kriminologijo*. Retrieved 2023-09-12 from https://www.policija.si/images/stories/Publikacije/RKK/PDF/2015/04/RKK2015-04_DanijelaFrangez_TheImportanceOfTerminology.pdf

[33] Michael J. Freedman, Kobbi Nissim, and Benny Pinkas. 2004. Efficient Private Matching and Set Intersection. In *Advances in Cryptology - EUROCRYPT 2004*, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Christian Cachin, and Jan L. Camenisch (Eds.). Vol. 3027. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–19. https://doi.org/10.1007/978-3-540-24676-3_1

[34] Andrew Crocker and Gennie Gebhart. 2019. Don't Let Encrypted Messaging Become a Hollow Promise. Retrieved 2023-09-12 from https://www.eff.org/deeplinks/2019/07/dont-let-encrypted-messaging-become-hollow-promise

[35] 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. https://doi.org/10.1109/SP46215.2023.00178

[36] Peter Guest. 2023. Britain Admits Defeat in Controversial Fight to Break Encryption. Retrieved 2023-09-12 from https://www.wired.com/story/britain-admits-defeat-online-safety-bill-encryption/

[37] Qingying Hao, Licheng Luo, Steve T.K. Jan, and Gang Wang. 2021. It's Not What It Looks Like: Manipulating Perceptual Hashing Based Applications. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. ACM, Virtual Event Republic of Korea, 69–85. https://doi.org/10.1145/3460120.3484559

[38] Dominik Hintersdorf, Lukas Struppek, Daniel Neider, and Kristian Kersting. 2022. Investigating the Risks of Client-Side Scanning for the Use Case NeuralHash. In *6th Workshop on Technology and Consumer Protection (ConPro)*. Retrieved 2023-09-12 from https://www.ml.informatik.tu-darmstadt.de/papers/hintersdorf2022conpro_learning.pdf

[39] Ashish Hooda, Andrey Labunets, Tadayoshi Kohno, and Earlence Fernandes. 2022. Re-Purposing Perceptual Hashing Based Client Side Scanning for Physical Surveillance. arXiv:2212.04107 [cs] Retrieved 2023-09-12 from http://arxiv.org/abs/2212.04107

[40] Ian Levy, Crispin Robinson. 2018. Principles for a More Informed Exceptional Access Debate. Retrieved 2023-09-11 from https://www.lawfaremedia.org/article/principles-more-informed-exceptional-access-debate

[41] Shubham Jain, Ana-Maria Cretu, Antoine Cully, and Yves-Alexandre de Montjoye. 2023. Deep Perceptual Hashing Algorithms with Hidden Dual Purpose: When Client-Side Scanning Does Facial Recognition. https://doi.org/10.1109/SP46215.2023.00179 arXiv:2306.11924 [cs]

[42] Shubham Jain, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. 2022. Adversarial Detection Avoidance Attacks: Evaluating the Robustness of Perceptual Hashing-Based Client-Side Scanning. In *31st USENIX Security Symposium (USENIX Security 22)*. 2317–2334. Retrieved 2023-09-12 from https://www.usenix.org/conference/usenixsecurity22/presentation/jain

[43] Joe Mullin. 2023. The EARN IT Bill Is Back, Seeking To Scan Our Messages and Photos. Retrieved 2023-09-11 from https://www.eff.org/de/deeplinks/2023/04/earn-it-bill-back-again-seeking-scan-our-messages-and-photos

[44] Jennifer Langston. 2018. How PhotoDNA for Video Is Being Used to Fight Online Child Exploitation – On the Issues. Retrieved 2023-09-12 from https://news.microsoft.com/on-the-issues/2018/09/12/how-photodna-for-video-is-being-used-to-fight-online-child-exploitation/

[45] Chance Miller. 2021. In Internal Memo, Apple Addresses Concerns around New Photo Scanning Features, Doubles down on the Need to Protect Children. Retrieved 2023-09-12 from https://9to5mac.com/2021/08/06/apple-internal-memo-icloud-photo-scanning-concerns/

[46] Joe Mullin. 2022. Google's Scans of Private Photos Led to False Accusations of Child Abuse. Retrieved 2023-09-12 from https://www.eff.org/deeplinks/2022/08/googles-scans-private-photos-led-false-accusations-child-abuse

[47] Dawn Marie Murphy and Becky Spencer. 2021. Teens' Experiences with Sexting: A Grounded Theory Study. *Journal of Pediatric Health Care* 35, 4 (July 2021), 387–400. https://doi.org/10.1016/j.pedhc.2020.11.010

[48] Lily Hay Newman. 2022. Apple Kills Its Plan to Scan Your Photos for CSAM. Here's What's Next. Retrieved 2023-09-12 from https://www.wired.com/story/apple-photo-scanning-csam-communication-safety-messages/

[49] Riana Pfefferkorn. 2019. William Barr and Winnie the Pooh. Retrieved 2023-09-12 from https://cyberlaw.stanford.edu/blog/2019/10/william-barr-and-winnie-pooh

[50] Erica Portnoy. 2019. Why Adding Client-Side Scanning Breaks End-To-End Encryption. Retrieved 2023-09-12 from https://www.eff.org/deeplinks/2019/11/why-adding-client-side-scanning-breaks-end-end-encryption

[51] Jathushan Rajasegaran, Naveen Karunanayake, Ashanie Gunathillake, Suranga Seneviratne, and Guillaume Jourjon. 2019. A Multi-modal Neural Embeddings Approach for Detecting Mobile Counterfeit Apps. In *The World Wide Web Conference*. ACM, San Francisco CA USA, 3165–3171. https://doi.org/10.1145/3308558.3313427

[52] Julio C. S. Reis, Philipe Melo, Kiran Garimella, Jussara M. Almeida, Dean Eckles, and Fabrício Benevenuto. 2020. A Dataset of Fact-Checked Images Shared on WhatsApp During the Brazilian and Indian Elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 903–908. https://doi.org/10.1609/icwsm.v14i1.7356

[53] Rosa Golijan. 2013. Is your cloud drive really private? Not according to fine print. Retrieved 2023-09-11 from https://www.nbcnews.com/technolog/your-cloud-drive-really-private-not-according-fine-print-1c8881731

[54] Michael Salter and Tyson Whitten. 2022. A Comparative Content Analysis of Pre-Internet and Contemporary Child Sexual Abuse Material. *Deviant Behavior* 43, 9 (Sept. 2022), 1120–1134. https://doi.org/10.1080/01639625.2021.1967707

[55] J. H. Saltzer, D. P. Reed, and D. D. Clark. 1984. End-to-End Arguments in System Design. *ACM Transactions on Computer Systems* 2, 4 (Nov. 1984), 277–288. https:

//doi.org/10.1145/357401.357402

[56] Frederick Schauer. 1978. Fear, Risk and the First Amendment: Unraveling the Chilling Effect. *58 Boston University Law Review 685-732 (1978)* (Jan. 1978). Retrieved 2023-09-12 from https://scholarship.law.wm.edu/facpubs/879

[57] Sharon Bradford Franklin, Andi Wilson Thompson. 2019. Open Letter to GCHQ on the Threats Posed by the Ghost Proposal. Retrieved 2023-09-11 from https://www.lawfaremedia.org/article/open-letter-gchq-threats-posed-ghost-proposal

[58] Priyanka Singh and Hany Farid. 2019. Robust Homomorphic Image Hashing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 11–18. Retrieved 2023-09-12 from https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Singh_Robust_Homomorphic_Image_Hashing_CVPRW_2019_paper.html

[59] Lukas Struppek, Dominik Hintersdorf, Daniel Neider, and Kristian Kersting. 2022. Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 58–69. https://doi.org/10.1145/3531146.3533073

[60] Kurt Thomas, Sarah Meiklejohn, Michael A. Specter, Xiang Wang, Xavier Llorà, Stephan Somogyi, and David Kleidermacher. 2023. Robust, Privacy-Preserving, Transparent, and Auditable on-Device Blocklisting. arXiv:2304.02810 [cs] Retrieved 2023-09-12 from http://arxiv.org/abs/2304.02810

[61] Zack Whittaker. 2021. Apple Delays Plans to Roll out CSAM Detection in iOS 15 after Privacy Backlash. Retrieved 2023-09-12 from https://techcrunch.com/2021/09/03/apple-csam-detection-delayed/

[62] Savvas Zannettou, Tristan Caulfield, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Characterizing the Use of Images in State-Sponsored Information Warfare Operations by Russian Trolls on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 774–785. https://doi.org/10.1609/icwsm.v14i1.7342

A DEMOGRAPHICS

Table 2 shows our participants' demographics.

Table 2: Interview participants. Country codes follow ISO 3166 alpha-2. n/a = not available

Label	educational background	stakeholder	current field of work	country of work	interview language	interview guideline (version)
X01	law	academia	research	US	EN	1
X02	law	consulting	national security consulting	US	EN	1
X03	computer science	digital rights	engineering and policy	US	EN	2
X04	law	digital rights, academia	research	DE	DE	2
X05	computer science	child protection	engineering	US	EN	2
X06	civil infrastructure	child protection	IT	NL	EN	2
X07	computer science	technology	machine learning	DE	DE	2
X08	psychology	child protection	child and youth welfare	DE	DE	2
X09	economics	consulting, digital rights	consulting	BE	DE	2
X10	n/a	child protection	child protection, technology	CA	EN	2
X11	n/a	child protection	child protection	DE	DE	2
X12	computer science	technology	security research (academic and commercial)	DE	EN	2
X13	computer science	academia	research	DE	EN	2
X14	computer science	academia	research	DE	EN	2
X15	law	law enforcement	prosecution	DE	DE	2
X16	law	law enforcement	prosecution	DE	DE	2
X17	n/a	digital rights	n/a	n/a	DE	2
X18	international relations	digital rights	digital rights	UK	EN	2
X19	journalism	journalism	journalism and IT security consulting	DE	DE	3
X20	computer science	academia	computer security research	DE	DE	4
X21	n/a	n/a	n/a	n/a	n/a	4
X22A	law	data protection	data protection	DE	DE	4
X22B	law	data protection	data protection	DE	DE	4
X22C	law	data protection	data protection	DE	DE	4
X22D	IT security	data protection	data protection	DE	DE	4
X23	philosophy and computer science	academia, ethics	research	DE	EN	4
X24	computer science	academia	research	DE	EN	4
X25	n/a	academia	n/a	n/a	EN	4

B INTERVIEW GUIDELINE

Figure 2 gives an overview of the interview guideline while the full interview guideline can be found in Table 3.

Table 3: **Interview Guideline. Rows with dotted lines happened before or after the interview, dashed lines indicate procedural steps at the beginning or end of the interview. Gray rows indicate that they were deprecated at some point.**

Part	Explanations, Questions	*Comments*, Further Explanations
1. Written Consent		
2. Introduction & Oral Consent	Thank you for your participation in this interview study. This interview will be about client-side scanning or CSS abbreviated. It's divided into several parts. We will start out with some general questions and your thoughts on client-side scanning (CSS). Then we'll do a short drawing task about your technical understanding before talking about some expectations, implications and stakeholders. [Trigger Warning] As a trigger warning, this interview may include discussion around CSAM (child sexual abuse material). We will not be showing any CSAM material. You can terminate this interview at any time for any reason without stating that reason. Are you ok to continue? You can say anything that comes to your mind. Please remember that there are no right or wrong answers or opinions. You are the expert in this interview. Do you still agree to recording this interview? Do you have any questions before we start?	
3. General Questions	We'll start out with some general questions about CSS. • When and how did you first come to know about client side scanning (CSS)? • What is the goal of client side scanning (CSS)? • What is your point of view about CSS and how did you get to it? How would you explain CSS to a good friend who has no technical background knowledge? Did you intentionally leave out parts in this explanation that you thought were too complex? Why?	
4. Explain CSS to a good friend	• How would you explain CSS to a good friend with no technical background knowledge? • Did you intentionally leave out parts in this explanation that you thought were too complex? Why?	
5. Technical Understanding		
5a. Drawing Task	Next, I ask you to visualize what you understand by CSS. The drawing task mainly helps you to explain this. I'll ask some follow-up questions and the focal point of this task is client-side scanning without a primary focus on a certain context. Please draw what client-side scanning is for you. You can draw the following scenario. Imagine that client-side scanning is running on a phone. This phone contains data that CSS is looking for. Please describe what CSS is doing and who is informed about what at what time. Please consider all relevant components, both your cell phone and other external components that are part of the process. Please say what you think.	*If Apple was or is mentioned, say that this is not only about the proposal of Apple, but their general understanding.*
5b. Follow-up Questions	• How does CSS work? • Can you go into more detail of your technical understanding of CSS? • What types of content could be scanned? • Where does the matching happen? • Who is responsible for the matching data? • Do you see potential for misuse? If yes, what and where? • Do you see privacy threats? If yes, what and where? • So you see any security threats? • Any other security or privacy issues that come to mind regarding CSS? • How feasible are the mentioned algorithms? • What types of CSS do you think are feasible in the wild? • What are your thoughts on the scalability? • What are technical standards that have to be met in order for CSS to be used widely? • What are other standards or regulation and rules that have to be given in order for CSS to run? • What about other enduser devices, like computers? What are challenges that come to mind?	*List of questions about interesting topics concerning technical understanding.*

Continued on next page

Table 3: **Interview Guideline. Rows with dotted lines happened before or after the interview, dashed lines indicate procedural steps at the beginning or end of the interview. Gray rows indicate that they were deprecated at some point.** (Continued)

Part	Explanations, Questions	*Comments*, Further Explanations
6. Expectations	• From what you've told me, I gather that you are for/against CSS. With the next question, I want to neutrally shed light on the boundaries. • What applications of CSS can you imagine? • What are your thoughts on implications through the use of CSS? • You might have mentioned parts of this already: What are your expectations towards CSS? • Is there something that fascinates you the most about CSS? If yes, please elaborate. • Is there something that concerns you the most about CSS? If yes, please elaborate. • Looking at this from a global perspective, what is the best that could happen if CSS is used? • Also looking from a global perspective, what is the worst that could happen if CSS is used?	
7. Implications - Scenarios	I will now ask you some questions about hypothetical contexts. Targeted data as mentioned in a very well-known paper about CSS is data that matches through CSS. [Trigger Warning] As mentioned before, we will talk about child sexual abuse material during these hypothetical scenarios. We will not be showing any CSAM material. We in no way want to implicate you or anyone else. Please remember that you can terminate this interview at any time for any reason without stating a reason. Are you ok to continue?	
7a. CSAM Scenario	In the first context, the goal of CSS is the fight against CSAM. Let's say we are talking about CSS running on a person's phone. • If CSS were to find targeted data, do you think the owner should be informed? Why? • How do you think they would react? Why? • What is information that they should receive? • What are personal implications you see in this context? • How would your answers to the previous questions change if we were talking about false positives? Now let's say that CSS is running on a device you own and regularly use. • If CSS finds targeted data on your device, what should happen? *(This is on purpose the same question as above. If the interviewee says, that it's the same as above, move directly on to B.)* • A – If CSS now finds some targeted material on your phone, what should happen? – Would you want to be informed? Why? – How would you react? Why? – What is information that they should receive? – Is there anything they shouldn't be made aware of? – Again, how would your answers to the previous questions change if we were talking about false positives? • B – Is there a difference between handling you and someone else? If yes, why? Now let's say, CSS is being applied to most devices. • What are possible societal implications? • What are possible societal implications of false positives? • What is the interest of law enforcement in this context? • What is the interest of lawmakers in this context? • What is the interest of governments in this context?	
7b. Copyright Scenario	For the second context, the goal of CSS is to catch copyright infringements.	*Ask the same question as for CSAM Scenario.*
7c. Government Scenario	In the third context, the responsibility of the matching database is in control of the government. They are supplying the dataset.	*Ask the same question as for CSAM Scenario.*
7d. Advertisement Scenario	In the fourth context, the goal of CSS is to collect information for advertisement providers.	*Ask the same question as for CSAM Scenario.*

Continued on next page

Table 3: **Interview Guideline. Rows with dotted lines happened before or after the interview, dashed lines indicate procedural steps at the beginning or end of the interview. Gray rows indicate that they were deprecated at some point.** (Continued)

Part	**Explanations, Questions**	***Comments*, Further Explanations**
7. Implications - Scenario	I will now ask you some questions about expectations, implications, and stakeholders. Targeted data as mentioned in a very well-known paper about CSS is data that matches through CSS. [Trigger Warning] As previously mentioned, the following parts may include the mention of CSAM. Are you ok to continue? Let's say we are talking about CSS running on a person's phone and the intention of CSS is to find CSAM. • If CSS were to find targeted data, what should happen? Why? • Is there something that should happen in any case? • Should the owner be informed? Why? • Suppose the cell phone finds sth., the content is forwarded to the planned EU center and it turns out to be a false positive. Should the person be informed in this case? What is information that they should receive or not receive? • What are personal implications you see in this context? • What happens if an individual is falsely accused of possessing CSAM by this algorithm? • What are personal implications of false positives? How should this be handled? Now let's say, CSS is being applied to most devices. • Are there groups of people that should be handled differently? • What are possible societal implications? • What are possible societal implications of false positives?	
8. Stakeholders	• Can you think of any other stakeholders we should consider? What are their interests? • What type of stakeholder do you consider yourself?	
9. Expert-Specific Questions	• Would the adoption of CSS influence your work? If yes, how? • Would the introduction of CSS influence your profession? If yes, how? • You as a [insert profession], do you have specific topics in mind regarding your area of expertise and CSS that we haven't touched on yet? • Do you see benefits or drawbacks specific to your profession? If yes, which?	
9a. Ethics	• Which role does ethics play in the introduction of CSS? • Do ethical constructs exist that should definitely be considered in the discussion around CSS? If yes, which and why? • Are there ethical dilemmas that can be transferred to this problem?	
9b. Data Protection	• Can CSS be used to compromise the privacy of a single person? • How can an organization that is responsible for introducing and maintaining CSS show transparency and fulfill its accountability obligations?	
10. Outro & Debriefing	Do you have any questions, or comments? Would you like to add anything else? Thank you for your participation. If you have any questions or want to change your consent to the usage of this data, please contact me.	*Remember to mention demographic survey and token of appreciation.* *List of help websites:* • *for potential victims of sexual abuse (list of URLs)* • *for people with a relevant inclination (list of URLs omitted for space)*
11. Follow-up demographic survey		

C CODEBOOKS

Figure 8 shows the codebook used to answer RQ1 while Figure 9 shows the codebook used to answer RQ2 and RQ3.

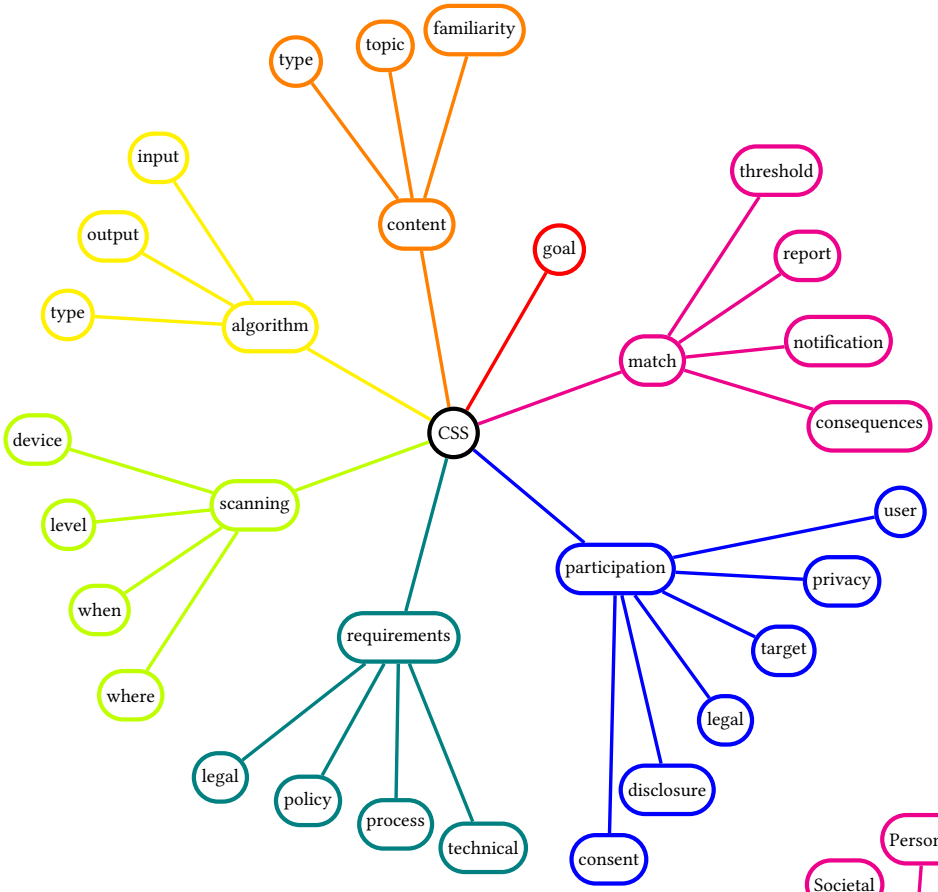

Figure 8: Visual representation of the codebook used to answer RQ1.

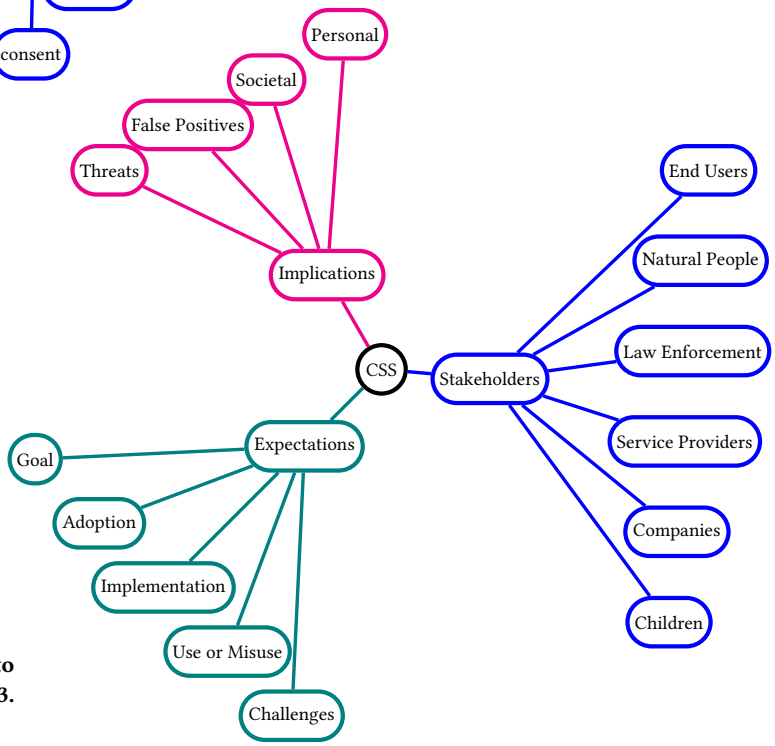

Figure 9: Visual representation of the codebook used to answer RQ2 and RQ3.