Gold Standard or Gold-Plated? Human Practices of Triple Verification in CSAM Takedown

Melissa Rottier
Delft University of Technology
Delft, Netherlands
m.j.rottier@student.tudelft.nl

Michel van Eeten
Delft University of Technology
Delft, Netherlands
m.j.g.vaneeten@tudelft.nl

Savvas Zannettou
Delft University of Technology
Delft, Netherlands
s.zannettou@tudelft.nl

Abstract

Child sexual abuse material (CSAM) presents a critical challenge for online safety, yet the verification procedures that determine which items are classified as CSAM remain poorly understood. Triple verification (requiring three reviewers to agree) is promoted as a safeguard, but little is known about how it is implemented, how it is perceived by experts, and how voting conditions affect reliability. We address this gap through a mixed-methods study. We interviewed 14 experts from seven organizations (e.g., law enforcement, hotlines, etc.) to map current verification practices, then ran an inter-reliability experiment with Dutch National Police experts who reviewed 2,031 images and videos under different voting conditions (blind vs. non-blind, varied order). Finally, we held a focus group to explore the reasons behind disagreements. We find that practices vary widely, perceptions of triple verification reflect both safeguards and burdens, and expert agreement depends on voting conditions and content type.

CCS Concepts

• **Human-centered computing** → **User studies**; **Empirical studies in HCI**; • **Security and privacy** → **Social aspects of security and privacy**.

Keywords

CSAM, triple verification, hash databases

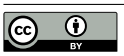
ACM Reference Format:
Melissa Rottier, Michel van Eeten, and Savvas Zannettou. 2026. Gold Standard or Gold-Plated? Human Practices of Triple Verification in CSAM Takedown. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3772318.3791039

1 Introduction

Child sexual abuse material (CSAM) represents one of the most harmful forms of online content. To combat its dissemination, trained professionals in national hotlines – private, non-profit organizations – evaluate millions of reports of potential CSAM every year [36]. If verified as CSAM, they issue a Notice-and-Takedown request (NTD) to the platform or hosting provider where the content is hosted [52]. The hotlines collaborate in a global network called INHOPE [24]. The bulk of all NTDs are sent from this network. Law enforcement professionals also evaluate CSAM in the course of their criminal investigations. They occasionally issue NTDs, but the day-to-day bulk processing of public reports and the generation of removal requests rest with the hotline network. In addition to triggering NTDs, verified CSAM is added to so-called hash databases, which are used by online providers to scan for known CSAM and remove it, without requiring further verification [54].

A core challenge for the evaluation is to establish a highly accurate outcome. Triple verification – where every instance of CSAM is evaluated and confirmed by three independent trained professionals – has emerged as the gold standard for this process [21, 37]. This helps in establishing trust with providers; they typically act on NTDs from INHOPE hotlines because they are seen as highly accurate. Accuracy also plays a key role in whether a notifier, such as a CSAM hotline, will be given the designation of so-called "trusted flagger" under the EU's Digital Services Act (DSA) [39]. NTDs from trusted flaggers "are given priority and are processed and decided upon without undue delay" [39]. In short, accuracy is at the core of the current NTD regimes. Of course, it is even more critical in criminal investigations. Yet, relying on triple verification also increases the workload for professionals in a situation that already faces severe scaling problems because of the amount of CSAM that is detected online [21, 37]. The increased workload also imposes more emotional burden on the evaluators. Moreover, if the evaluation has difficulty scaling, then less material can be covered, meaning more false negatives. This allows more CSAM to continue spreading.

This tension raises the question of how to balance accuracy and efficiency in the context of CSAM evaluation. There is very little research in this area. Prior work around CSAM has focused on the detection methods [20, 22, 32, 33] and on user perceptions of the use of client-side scanning (using hash databases) to combat CSAM [2, 18, 19]. The latter work did touch on accuracy as an important issue. To the best of our knowledge, no prior work in HCI has been done on the verification procedures; how they are operationalized in different contexts, what the impact is of different conditions, such as blind or non-blind verification (blind verification refers to the setting where reviewers assess a case without access to previous labels and non-blind verification allows reviewers to see prior labels from other reviewers), or how disagreements are resolved by the professionals. The closest related work is outside of HCI, in forensic science, where researchers have conducted measurements around inter-rater reliability for CSAM [29, 30]. Contrary to our study, this was only focused on applications in criminal prosecution; plus, they used reviewers who are not professionals routinely classifying CSAM, nor did they test different evaluation conditions. Taken together, there is an important knowledge gap in understanding how verification is structured, experienced, and evaluated in practice in organizations that maintain CSAM hash



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791039

databases, as well as empirically assessing the inter-rater reliability in CSAM classification under realistic conditions.

Our study aims to address this knowledge gap by providing answers to the following research questions:

- **RQ1:** What verification procedures are in place in organizations that create and maintain CSAM hash databases?
- **RQ2:** How do experts working in organizations aiming to identify illegal content perceive various verification procedures, including triple verification?
- **RQ3:** How do different verification conditions (e.g., blind vs. non-blind, voting order) affect inter-rater reliability in CSAM classification?
- **RQ4:** What challenges cause experts to disagree when classifying CSAM?

To shed light on the above-mentioned research questions, we performed a mixed-methods study. First, we conducted 14 interviews with experts working in seven organizations, including law enforcement, CSAM hotlines, and governmental organizations that aim to tackle the spread of illegal content online. Second, in collaboration with the Dutch National Police, we conducted an experiment where three CSAM evaluators reviewed 2,031 images and videos. The aim is to understand how the inter-rater agreement is impacted by double versus triple verification and by various voting conditions that were reported in the interviews. Finally, to understand the underlying reasons for the disagreements between the expert reviewers, we conducted a focus group, where we had an open discussion with the evaluators to surface the main challenges that exist and cause experts to disagree when annotating CSAM material.

Our study makes the following main findings:

- **Verification practices vary across organizations.** While triple verification is often treated as the standard, some organizations rely on double verification or reserve triple verification for ambiguous cases. Systems also vary in whether votes are blind or non-blind, reflecting trade-offs between independence and efficiency. Finally, classification frameworks range from simple three-label schemes to detailed taxonomies with over 20 tags, highlighting a trade-off between database richness and required time to create and maintain (**RQ1**).
- **Expert perceptions of triple verification reveal a trade-off.** Participants viewed it as an important safeguard against misclassification and for legal compliance, but also noted its costly nature in terms of time and emotional toll on the reviewers. Experts favored flexible applications based on case complexity and stressed that, while automation can support filtering, human judgment remains essential for final decisions (**RQ2**).
- **Voting conditions have an effect on inter-rater agreement.** Expert agreement increased under non-blind conditions (Cohen's Kappa score of 0.893 vs. 0.670), and it was higher for videos than for images (Cohen's Kappa score of 0.812 vs. 0.601 under blind conditions). In addition, we find that voting order also influenced the agreement of the experts. Also, while two reviewers already achieved high-agreement, adding a third reviewer revealed disagreements in 8% of the cases, highlighting that uncertainty persists even with three expert reviewers (**RQ3**).
- **Focus group insights highlights key challenges around disagreement in CSAM classification.** Experts often relied on series recognition, meaning that they judged CSAM based on prior knowledge that it belonged to a known series (e.g., recurring setups, backgrounds, or logos), even if the individual image itself did not depict explicit abuse. They also noted the difficulty of age estimation, especially across ethnicities, and emphasized that sexualized framing rather than nudity alone often drove classification. Finally, while non-blind conditions provided useful peer support, raters stressed the importance of retaining individual responsibility for final decisions (**RQ4**).

2 Background & Related Work

Our study is located on the intersection of different areas of research: tools to detect CSAM, verification of CSAM, content moderation practices and experiences, and user perception of anti-CSAM measures, most notably client-side scanning. Our work draws insights from the HCI community regarding the psychological costs of human verification of disturbing content [14, 47, 50, 55], user perception of automated scanning [2], and the theory of "data cascades" in automated detection systems [45].

2.1 Detecting CSAM

CSAM detection relies on a multi-layered ecosystem involving public reporting, expert verification, and downstream technical enforcement. Most suspected material enters this ecosystem through public reports submitted to hotlines, which receive millions of reports each year from individuals, hosting providers, and online platforms [36]. A smaller but critical share originates from law enforcement, where investigators encounter potential CSAM in the course of investigations and forward items for expert review. Verified material serves two core functions. First, hotlines issue NTD requests to online platforms or hosting providers, who remove the content from their services [52]. Second, verified items are added to curated hash databases maintained by organizations such as NCMEC, INTERPOL, and international hotline networks [54]. These databases are used by hosting providers and online platforms for proactive scanning (either server-side or client-side) to detect and block known CSAM at scale [32]. Across organizations, items that do not yet receive consistent votes are not entered into these databases; instead, they remain in a provisional state and are re-reviewed until sufficient agreement is reached, rather than being discarded.

Hash databases effectively encode professional judgments into a reusable infrastructure for automated enforcement. Consequently, much of the CSAM detection literature focuses on how to match content against these databases efficiently and robustly. Specifically, research on tools for CSAM detection has mostly focused on matching content against curated hash databases, using cryptographic or perceptual hashes of confirmed illegal material. Cryptographic hashes (e.g., MD5, SHA-1) reliably detect exact duplicates but fail when content is altered [33]. Perceptual hashing addresses this by encoding visual features, enabling the identification of modified

but visually similar CSAM [13, 15, 33]. Microsoft's PhotoDNA operationalizes this approach [32]. More recently, hash databases underpin *client-side scanning (CSS)* systems [1, 2, 19], such as Apple's Neuralhash [15, 43]. The effectiveness of these systems ultimately depends on the integrity of the underlying hash databases, which hinges on professional judgment. Our study aims to provide more insight into the human verification.

2.2 Verification of CSAM

In the CSAM ecosystem, verification outcomes carry exceptionally high stakes because they affect both immediate takedown actions and the long-term accuracy of hash databases used for automated detection. In this context, an "error" can take two forms. First, a false negative, in which illegal material is not detected, allowing continued circulation and further victimization. Second, a false positive, in which benign content is misclassified as CSAM, potentially triggering inaccurate reporting and wrongful inclusion in hash databases. Hotlines receive millions of reports each year [36], and their decisions determine which items proceed to NTD procedures and which are added to hash databases, creating the basis for proposals for client-side or server-side scanning by platforms. Because these hashes propagate globally and across providers, even a single false positive can contaminate downstream detection systems, while false negatives reduce the reach of CSAM interventions. In this context, an error is not merely a local mistake but a phenomenon known in the HCI community as a "data cascade" [45]; a compounding event causing negative downstream effects from data/verification issues. Overall, this creates a structural tension: high verification accuracy is essential for trusted-flagger status under the DSA and for maintaining providers' confidence in hotline decisions, yet verification must operate at the scale of unprecedented volumes of reported material [36]. Therefore, it is crucial to study and understand the strengths and limitations of human decision-making when verifying highly sensitive material like CSAM.

Despite the centrality of human judgment in preventing false positives and false negatives from cascading downstream, empirical evidence about CSAM verification remains limited. The closest body of work comes from forensic science and criminal investigations, which has begun to quantify reliability and sources of disagreement in CSAM classification. Specifically, prior research has explored the strengths and limitations of human decision-making in verifying CSAM. Kloess et al. [29, 30] studied interrater reliability and decision-making in classifying CSAM material using law enforcement raters in the UK. They find a substantial agreement rate and highlight some notable disagreements related to challenges about the ambiguity around age estimation and contextual cues. Our work builds on this prior work in three ways. First, we explore verification procedures in other contexts than criminal prosecutions. Second, Kloess et al. used experts, but not the real evaluators. Our experiment works with the actual professionals doing the verification as their main task. Third, we test verification results under different evaluation conditions, like blind vs. non-blind.

2.3 Psychological burden on content moderators and verifiers

The human costs of verification are significant as moderators and reviewers report psychological strain from the repeated exposure to traumatic material, pointing to the necessity of rotation, trauma-informed support, and interface designs that reduce exposure without degrading the quality of the verification process [34, 48]. In the HCI community this phenomenon is characterized as intense "emotional labor" [14, 50]. A complementary lens is *vicarious traumatization*, which suggests that repeated, work-related exposure to traumatic material can have cumulative and enduring psychological effects, originally documented in trauma-facing professions such as therapy [40]. Together, these lenses separate the short-term work of managing emotions from the longer-term risks of repeated exposure. This labor is not limited to commercial or organizational settings; it also heavily impacts volunteer moderators who often lack institutional support [55]. Scott et al. [47] argue that current platform policies often fail to address this trauma, which affects moderators exposed to sensitive content like CSAM. They propose a "trauma-informed" approach to social media architectures, governed by six core principles: safety, trustworthiness, peer support, collaboration, empowerment, and cultural humility. This framework suggests that interventions must actively prevent re-traumatization. In the context of interventions, HCI researchers have proposed various sociotechnical interventions. Interface modifications, such as greyscale filtering and content blurring, have been shown to reduce immediate emotional impact without significantly degrading verification accuracy [11, 28]. Recent work has extended these techniques to text moderation, using AI to paraphrase hate speech before being presented to moderators and allowing moderators to see the original content only if they opt to see it [38]. Beyond interface design, positive psychological interventions, such as positive stimuli during breaks (e.g., positive images), have been explored [9] to reduce the negative side-effects of moderators. Overall, in line with previous work, our interviews also include a focus on the psychological tolls. Our study advances this area by exploring the tradeoff between accuracy and efficiency, where more efficient procedures (e.g., double verification) would lower the burden on moderators.

2.4 User and Expert Perceptions

Legal scholarship has examined expert perspectives on the effectiveness of CSAM regulation across borders [3, 10, 17], while the rise of AI-generated child sexual content has introduced new legal and detection gray zones [31, 53]. Client-side scanning has sparked particular debate. Bhardwaj et al. [2] show that experts across child protection, law enforcement, data protection, and academia often view CSS as promising, yet highlight risks such as false positives and children's rights concerns. Public studies reveal similar ambivalence: Deldari et al. [12] find users worry about opaque detection and inappropriate data access, while surveys in Germany show strong support for CSAM scanning relative to other offenses [19]. A follow-up cross-national study reports comparable support in the U.S., though shaped by cultural differences [18]. This research sets the context for our work, but it does not study the verification approaches underlying hash databases and the NTD regimes.

3 Methodology

To investigate how verification processes for CSAM are operationalized and experienced in practice, we adopt a multiphase mixed-methods design. This approach includes: (1) semi-structured interviews with experts to uncover organizational workflows and perceptions of multi-rater verification; (2) a controlled inter-reliability experiment with expert reviewers to empirically assess how verification conditions affect inter-rater agreement in CSAM classification; and (3) a follow-up focus group to contextualize quantitative findings through in-depth discussion of disagreement cases. This sequential design allowed qualitative insights to inform the experimental setup and enabled iterative interpretation of the results. Our mixed-methods design was intentionally sequential: insights from the qualitative phase directly shaped the structure of the quantitative experiment, and the quantitative findings were subsequently re-embedded in qualitative explanation. First, the interviews revealed substantial variation in verification workflows across organizations, particularly around blind vs. non-blind voting and voting order. These informed the experimental conditions we tested. Second, the focus group was used explicitly to interpret quantitative disagreements from the experiment: cases in which reviewers disagreed, particularly around borderline CSAM and "Other" decisions. These were revisited collectively, allowing us to link statistical disagreement patterns to the underlying cognitive and contextual factors identified qualitatively (e.g., series recognition, cultural differences in age estimation, etc). Together, these phases created a coherent qualitative–quantitative–qualitative loop that kept the experiment closely aligned with real-world verification challenges. Below, we describe the three phases of our methodological approach.

3.1 Interviews

To understand how verification is organized and experienced in practice, we conducted 14 semi-structured interviews with experts from law enforcement, hotlines, and organizations managing or using hash databases for CSAM. Below, we elaborate on how we conducted the interviews.

3.1.1 Recruitment and Participants. Our aim was to interview professionals from organizations that verify harmful content. We focused primarily on CSAM, but we also tried to recruit some professionals who deal with terrorist content to get a wider sense of verification practices.

Participants were recruited through purposive sampling [4] to ensure coverage of diverse organizational roles and expertise. For the purposes of identifying and recruiting participants, we collaborated with the Authority for the Prevention of Online Terrorist Content and Child Sexual Abuse Material (ATKM) in the Netherlands. ATKM is the Dutch national authority responsible for detecting and assessing online terrorist and CSAM and for sending removal orders to providers. ATKM has an extensive network of people and organizations working on CSAM hash databases. In collaboration with the ATKM and its existing connections with various organizations, we recruited 14 participants. Our participants' sample includes both verification experts (i.e., people who are responsible for classifying content) and hash database managers (responsible for both content classification and maintenance of the hash database). Table 1 provides an overview of the recruited participants. Overall, we recruited participants from seven different organizations:

- **ATKM:** The Dutch Authority for the Prevention of Online Terrorist Content and Child Sexual Abuse Material is a national authority that coordinates the removal of illegal content.
- **C3P:** The Canadian Centre for Child Protection is a non-profit organization that operates national child protection hotlines, manages a hash database of CSAM, and collaborates internationally to support law enforcement and prevent child exploitation.
- **Dutch National Police:** Law enforcement agency that is responsible for investigating CSAM-related offences, classifying material for inclusion in hash databases, and supporting criminal prosecutions through expert verification.
- **Offlimits:** Offlimits is a Dutch hotline and NGO specializing in the assessment and reporting of CSAM, operating at the intersection of victim protection, public reporting, and collaboration with law enforcement.
- **INTERPOL**: INTERPOL operates the International Child Sexual Exploitation (ICSE) database, facilitating the exchange of verified CSAM hashes and investigative intelligence among law enforcement agencies worldwide.
- **IWF:** The Internet Watch Foundation is a UK-based hotline that identifies and removes CSAM globally, curates a hash database, and collaborates with both industry and law enforcement to reduce CSAM.
- **GIFCT:** The Global Internet Forum to Counter Terrorism is an industry-led consortium that maintains a shared hash database for terrorist content and coordinates collaborative moderation efforts among technology platforms.

During our participant recruitment, we reached out to other relevant organizations as well, specifically the National Center for Missing and Exploited Children (NCMEC) and Tech Against Terrorism; however, they did not respond to our invitation to participate in our interviews.

3.1.2 Interview Procedure. Interviews were conducted in a semi-structured format to explore organizational workflows and expert perspectives on verification, especially on triple verification procedures. In particular, to accommodate differences across organizations, we prepared a separate set of interview questions for law enforcement and for the non-governmental organizations (see Appendix B for our interview protocol). Before the interviews, we informed the participants about our research objectives and the intended use of the interview data, obtained their informed consent, and emphasized their voluntary participation and right to withdraw at any time without providing a reason. Then, we proceeded with a semi-structured question guide covering topics related to the participants' background, queries about verification procedures, perceptions of triple verification, and contextual factors specific to CSAM. Each interview lasted between 45 and 60 minutes, was conducted in English or Dutch (based on the participant's preference), and took place either in person or over a secure call. The interviews were conducted between February and March 2025. With our participants' consent, all sessions were audio-recorded and locally transcribed using OpenAI's Whisper model [42]. Transcripts were

Table 1: Overview of our interview participants.

Label	Stakeholder	Field of work	Role of organization	Country	Language	Gender	Years of experience
P01	Hotline	Child protection	Verification	NL	NL	Male	0-5
P02	Hotline	Child protection	Verification	NL	NL	Female	0-5
P03	Hotline	Child protection	Verification	NL	NL	Male	0-5
P04	Hotline	Child protection	Verification	NL	NL	Male	0-5
P05	Law enforcement	Criminal investigation	Verification and hash database management	NL	NL	Male	20+
P06	Law enforcement	Criminal investigation	Verification and hash database management	NL	NL	Male	20+
P07	NGO	Internet governance	Verification and hash database management	Worldwide	EN	Female	15-20
P08	Hotline	Child protection	Verification	NL	NL	Male	5-10
P09	Hotline	Child protection	Verification	NL	NL	Female	5-10
P10	Hotline	Child protection	Verification	NL	NL	Male	10-15
P11	Hotline	Child protection	Verification	NL	NL	Female	5-10
P12	Law enforcement	Criminal investigation	Verification and hash database management	Worldwide	EN	Male	20+
P13	Hotline	Child protection	Verification and hash database management	UK	EN	Male	10-15
P14	Hotline	Child protection	Verification and hash database management	CA	EN	Male	20+

manually reviewed to ensure accuracy and anonymization before being stored on our secure servers. While our initial goal was to conduct all interviews individually, due to scheduling constraints and the preference of a single organization, we had a single interview with four participants. Each participant answered all questions.

3.1.3 Data Analysis. All interview transcriptions were imported into ATLAS.ti for inductive, thematic analysis, which is a data-driven approach that begins with generating open codes directly from the transcripts to simplify and conceptualize large volumes of text [5, 27]. Following our initial coding, we continuously refined and iterated on our code set (i.e., merging overlapping codes, clarifying definitions, and reorganizing related codes) to organize the codes into coherent main themes and sub-themes. Our iterative thematic analysis led to the generation of 39 codes. The codes were then grouped into four main themes: (1) Benefits of triple verification; (2) Challenges of triple verification; (3) Necessity of triple verification; and (4) Future perspectives and opportunities. Our final codebook, including the main themes, sub-themes, and their codes, is in Appendix C.

3.2 Inter-Rater Reliability Experiment

To empirically examine how verification practices influence CSAM classification accuracy, we designed an annotation experiment grounded in insights from both prior literature and our qualitative interviews. Our interviews revealed significant variation in organizational workflows (see Section 4.1), with some organizations using blind triple voting, others employing sequential non-blind voting, and some with no fixed procedure, accepting double verification in some situations. Based on these insights, we operationalized two experimental conditions that mirror real-world practices: *blind verification*, where reviewers classify items independently without access to prior votes, and *non-blind verification*, where reviewers sequentially review content with visibility into earlier votes.

To conduct the experiment, we first created a set of 2,031 samples that were initially reviewed by an expert working for the Dutch National Police (single verification). The samples were selected by the law enforcement professional to provide diversity in age ranges and severity while minimizing, but not fully excluding, so-called 'series' content to preserve ecological validity. Series content refers to the identification of whether an image or video belongs to a broader set of files that share substantive features and/or a common manner of creation. Based on this, it might be the case that a specific image does not show anything that formally meets the CSAM definitions, but the series as a whole does. Thus, the whole series would be classified as CSAM, including that specific image.

For the double and triple verification measurements, we recruited two additional expert reviewers from the Dutch National Police, whose daily work involves classifying potential CSAM images and videos. Reviewers classified the content into the three categories used in Dutch law enforcement practice: CSAM, Animal Pornography, and Other. Ground truth was defined procedurally as agreement among three votes, reflecting how items are finalized for inclusion in the CSAM hash database.

The experiment was conducted over a two-week period in May 2025, scheduled to fit within the expert reviewers' regular workload. We conducted the experiment in two phases. In Phase 1 (blind verification), each reviewer independently classified 1,000 items (images and videos) in their standard work environment without access to the other reviewer's votes, although they knew that a prior classification existed. In Phase 2 (non-blind verification), the reviewers assessed a total of 1,031 items, divided over four rounds (292, 293, 228, and 218 items, respectively). The variation in set size was due to technical limitations in how items could be retrieved

from the database. Selecting exactly 250 items per round would have required extensive additional handling, which was not feasible within the available timeframe. In this condition, each reviewer could view prior votes through a temporary system configuration that was enabled for the experiment by the Dutch National Police. Specifically, the reviewers were instructed to open a designated tab within their system that provided visibility into prior votes. This normally restricted functionality was explicitly enabled to facilitate the non-blind condition.[1] Both reviewers completed all classifications within the secure environment of the Dutch National Police using operational tools, over a two-week period that accommodated their regular workload. For each phase, we generated output files that recorded the individual votes, relevant file metadata (e.g., size, dimensions), and a "master category" indicating whether the three votes converged. The output files contained a larger number of rows compared to the number of images/videos because the system groups visually similar files through perceptual hashing, as grouped by perceptual once, while these similar files are reported in different rows in our output file. Therefore, the blind condition output comprised 1,140 rows and the non-blind condition 1,373. Having obtained the outputs of our experiment, we then calculated the agreement in double vs. triple verification settings, as well as calculated the inter-reliability agreement of our experts across different conditions using Cohen's Kappa score [8].

3.3 Focus Group

To shed light on the reasons for disagreements and challenges faced by the reviewers during our inter-reliability experiment, we conducted a two-hour post-assessment focus group with both Dutch National Police reviewers. The session was conducted in a hybrid mode, with one reviewer being present on-site while the other joined remotely. The session centered around the reasons for the disagreements for 49 items (45 images and 4 videos). The sample was chosen to ensure that less frequent but sensitive cases would not be overlooked. First, we included all six items with disagreements on animal pornography versus CSAM or Other. The remaining 43 items were then drawn manually from the pool of CSAM versus Other disagreements. We aimed to include both images and videos. As disagreements were rarer in videos, fewer of these were included in the final sample. During the focus group, participants were prompted to articulate the rationale for their classification decisions on each disputed item, enabling us to link quantitative disagreement patterns to the cognitive and contextual factors influencing rater judgment. The session was audio-recorded and subsequently annotated. The recording was transcribed locally using the Whisper model and then coded to identify recurrent reasons for disagreements. These qualitative insights enriched our understanding of verification dynamics in the context of CSAM under realistic environments.

[1]The authors of this paper do not have access to or any control over making changes to the system from the Dutch National Police. The reviewers were instructed to review samples as part of their regular CSAM verification workload. Because this is a national government system operating under strict security and confidentiality protocols, we cannot provide screenshots or detailed depictions of the workflow for either the blind or non-blind conditions. These security and confidentiality constraints are also the reason why we anonymized the participating organizations in this study.

3.4 Ethical Considerations

Our study focuses on CSAM, which is a highly sensitive topic that requires careful ethical considerations. Before conducting our study, we obtained approval from our institution's Ethical Review Board. All participants involved in the interviews were experienced professionals who regularly work with CSAM in their official capacities. We believe that their expertise and familiarity with CSAM helped minimize the potential psychological and mental harm that could arise from participating in the study. At the same time, we recognize that experts are still humans and may be affected by exposure to distressing material and topics. We therefore designed our study to align with their daily work routine. Specifically, before conducting the interviews, we obtained all participants' explicit verbal consent to participate and informed them that they could withdraw at any time if they felt uncomfortable. With respect to our annotation experiment, we carefully designed the experiment so as not to introduce any harm. In particular, the raters in our annotation experiment were experts from the Dutch National Police whose daily responsibilities include the review and categorization of CSAM material. Given their training and their daily job activities, we believe that our study did not subject them to psychological or mental risks beyond those encountered in their regular work. Also, we emphasize that members of the research team (i.e., the authors of this work) did not have access to and were not exposed to any CSAM material, ensuring that this research did not affect their well-being. Finally, to safeguard our participants' privacy and anonymity, all interview transcripts were anonymized and stored in secure servers.

3.5 Researcher Positionality and Reflexivity

Given the close collaboration with the Dutch National Police, we acknowledge the possibility of positional bias, particularly the risk that institutional perspectives might shape data interpretation and reporting. To mitigate this, we used some reflexive practices. First, our police collaborators, who are not authors of this paper, were asked to review two specific parts of this paper: (1) the methodology description of the inter-rate reliability experiment (Section 3.2) and (2) the results for the experiment and the focus group (Section 4.3 and 4.4). Their review was limited strictly to verifying that no operationally sensitive information is included in this paper, and they did not provide any input on data interpretation and framing of the results, with a sole exception on the definition of "series recognition," to ensure terminological accuracy. Second, all qualitative coding and analyses were conducted independently by the research team, without feedback from our law enforcement collaborators. These steps ensured interpretive neutrality while maintaining the security and confidentiality protocols required for research involving a national law-enforcement system.

4 Findings

4.1 RQ1: What Verification Procedures are in Place in Organizations?

In this section, we describe our findings based on the interviews and related to the verification procedures used across various organizations. For the purposes of this analysis, we focus only on

the five organizations that are responsible for verification and hash database management (see Table 3 in Appendix A for the list of organizations). The remaining two organizations, ATKM and Offlimits, were excluded because they do not operate a hash database and therefore do not have independent processes relevant to this analysis. To safeguard the confidentiality of the procedures and workflow of these organizations, we anonymize the organizations and label them as A to E. Table 2 outlines the key factors related to the verification procedures of each organization such as the type of verification mandated, the number of reviewers involved, classification criteria for content, reviewing setup, and the size of the databases each organization manages. Below, we provide more details on each organization.

4.1.1 Organization A. Organization A employs a mandatory triple verification process to ensure high confidence in the classification of illegal material before inclusion in the permanent hash database. The review begins with an automated hash check that filters out any pre-verified content. From an initial batch, which may contain up to 2 million files, only unrecognized images are routed for human verification. Reviewers are instructed to categorize each image into one of three fixed categories: child pornography, animal pornography, or other. Importantly, the system tracks how many votes an image has already received, though not the content of those votes. Reviewers thus know whether they are casting the first or second/third vote, but not how previous reviewers classified the image. This partial transparency renders the process blind, as reviewers are aware of their order in the voting sequence, though they are not influenced by prior votes. Images that receive three identical votes are considered verified and are transferred to the permanent database. Items that have been reviewed once or twice remain in a provisional state, awaiting the required number of classifications. Items with disagreements stay in the provisional queue and may receive additional reviews. The item is only added to the hash database once three positive identifications have been recorded.

4.1.2 Organization B. Unlike other organizations, Organization B does not prescribe a fixed verification process. Instead, it facilitates a collaborative hash-sharing framework where participation is conditional on adherence to a strict set of membership criteria, including alignment with human rights principles and six additional technical and ethical requirements. Member organizations are granted the autonomy to put hashes directly into the shared database, provided that the content meets the consortium's clearly defined inclusion standards. These standards are supported by a transparent taxonomy that includes criteria for what qualifies as harmful material. The taxonomy ensures consistency, even in the absence of a uniform verification protocol. While the number of reviewers involved and the reviewing setup (e.g., blind or non-blind) are unspecified, the system includes built-in dispute mechanisms. Member organizations can flag and request a re-review of any hash they believe is mislabeled. This provides a form of post-hoc verification that prioritizes consensus and database integrity.

4.1.3 Organization C. Organization C adheres to a strict triple verification model, underpinned by narrowly defined classification criteria. Only images that depict real children under the age of 13 and that are considered to involve severe forms of abuse are eligible for inclusion in the hash list. This narrow scope results in a small database. Each image must be independently reviewed and approved by three trained reviewers. The review workflow is carefully structured to reduce redundancy and bias: once a reviewer has seen an image, it is automatically routed to a different colleague, never returning to the same person. In this organization, reviewers can see whether an image has received 0 or 1 votes. The final vote is cast in a distinct interface, clearly signaling to the reviewer that they are providing the decisive judgment.

4.1.4 Organization D. This organization implements a flexible, context-dependent verification process. Depending on the nature and severity of the content, images may undergo single, double, triple, or even more reviews. While there is no fixed threshold for verification, ambiguous content, particularly material involving older minors (ages 14+), is mandatorily subjected to multiple assessments to reduce false positives. The review process begins with the submission of suspicious URLs, often by the public or victims themselves, which is a unique feature among the surveyed organizations. Reviewers then extract images and assign rich metadata, using up to 21 classification tags such as estimated age, gender, number of individuals, and specific sexual acts. The voting system may be blind or non-blind, depending on the workflow. Although the system includes features such as clustering and historical context that allow reviewers to infer how others may have voted, there is a strong emphasis on independent assessment. Reviewers are encouraged to base their decision solely on the content and contextual tags presented. Verification relies on achieving a predefined number of consistent votes, but voting order is not fixed.

4.1.5 Organization E. Organization E transitioned from a triple to a double verification process, following an internal evaluation that demonstrated limited added value from a third reviewer. The classification system distinguishes between two primary categories: "CSAM" and "Harmful to Children." Each of these categories includes a range of sub-classifications to capture nuance. The non-blind voting process is supported by a queue management system that prioritizes images based on both severity and prior classification. For instance, images depicting prepubescent children are expedited to ensure rapid review by a second reviewer. Similarly, content that is potentially illegal but ambiguous is also prioritized, albeit slightly lower in the queue. The workflow is segmented such that some reviewers focus on initial classification, while others specialize in follow-up validation. This division allows for efficient processing of high volumes of content, especially given the organization's very large database. By reducing the required number of votes per image, the organization aims to minimize reviewer exposure to harmful content while maintaining classification reliability and ethical standards.

4.1.6 Main Takeaways. In sum, the main takeaways from the interviews related to the verification procedures employed by various organizations are:

- **Triple verification is a strong norm, but it is not universally applied.** While some organizations require three identical votes before including a hash in their database,

Table 2: Overview of processes across different organizations.

Organization	Verification Process	Number of Reviewers	Classification Criteria	Reviewing Setup	Database Size
Org. A	Triple verification mandated	> 50	3 categories, including child pornography, animal pornography, and other	Blind	2M–10M
Org. B	No prescribed verification process	Unknown	Transparent taxonomy based on human rights and specific criteria for membership	Unknown	2M–10M
Org. C	Triple verification mandated	< 20	Severe content only, real images of children under 13	Non-blind	< 500K
Org. D	Variable, based on context (single, double, triple, or more)	20–50	3 categories with extensive metadata, 21 tags including age, gender, specific acts, etc.	Blind/Non-blind	2M–10M
Org. E	Switched from triple to double	> 50	CSAM and harmful to children, multiple categories within both buckets	Non-blind	> 10M

others have moved to double verification or apply triple verification only for ambiguous cases.

- **Blindness to prior votes depends on system configuration and organizational priorities.** While triple verification implies independence, most organizations operate non-blind or mixed systems where reviewers can see earlier votes. This choice reflects a trade-off between independent assessment and workflow optimization.
- **Classification detail varies according to intended database function.** Some organizations employ a streamlined three-category workflow optimized for rapid verification. Other organizations capture over 20 metadata tags, encoding details like age and nature of acts. This higher granularity supports richer searching and analytics, but requires more time and expertise.

4.2 RQ2: What are Experts' Perceptions on Triple Verification?

In this section, we describe our findings on the thematic analysis of the interviews. The analysis revealed five themes: benefits, challenges, necessity, as well as future perspectives and opportunities.

4.2.1 Benefits of Triple Verification. We examined how participants perceived the benefits of triple verification. Participants described a wide range of positive aspects. They highlighted how triple verification strengthens legal certainty, supports difficult assessments, and improves database quality.

Legal Considerations. Participants consistently highlighted the legal dimension. They explained that while INTERPOL's "baseline" criteria[2] is treated as illegal across jurisdictions, national laws differ. Several participants said this was the main reason to use multiple reviewers, to make sure the content meets the criminal threshold in their country. P02 gave the example of the Netherlands, where reviewers must check content against Article 252 of the Criminal Code. P10 highlighted *"It really needs to be legit. The child must be under 13, it must be a real child, and it should contain a sexual pose."*

Content Assessment. Ten participants said triple verification helps in cases where content is difficult to judge. They mentioned challenges such as estimating age or determining whether a pose is sexual. P12 explained *"If I show a picture to my colleagues and say: can you estimate the age of this child? The result might be different. Verification is [...] so hard."* P14 added that cultural, situational, and emotional factors should be considered when evaluating content.

Inconsistent Assessment. Nine participants discussed inconsistency. They said that even with good training, reviewers sometimes reach different conclusions. P13 observed that despite a *"solid foundation of knowledge"* individual perspectives still led to variation. P06 referred to more practical mistakes *"We also have people who misprint, who make mistakes, and who click the wrong picture. Yes, there will be those too, but just for that there are already three pairs of eyes that take out those rotten ones."* P05 also described how distractions, such as accidentally pressing the wrong button, could result in misclassification.

Quality. Seven participants stated that triple verification improves the accuracy and reliability of hash databases. P02 stressed, *"Just for the purity of the database. The moment three people look at it, the chances of false positives are minimal."* P01 and P04 agreed that multiple checks could be important for maintaining confidence in the system.

Operations. Four participants reflected on the operational side of triple verification. P05 explained that assessing illegal content is not an exact science. While single classifications may not take much time, participants stressed that making decisions legally sound and consistent is more complex. Training and education were seen as essential to reach this consistency. P13, who works with international hotline reviewers, described how differences in national law require ongoing instruction *"Some of the base training related to sexual maturation rates is pretty universal. But the law within your own country, you might start to learn with that. That is the reason why we have extensive training with them to start to explain the different categories."*

[2]Baseline refers to a set of strict INTERPOL criteria used for evaluating CSAM. To meet these criteria, the content must show a real child who appears to be under the age of 13 and is either involved in or witnessing sexual activities, or where there is a focus on the child's genital or anal area.

4.2.2 Challenges of Triple Verification. We examined the challenges participants perceived with triple verification. While they valued its safeguards, they also described burdens that make implementation difficult. The most common challenges were the emotional toll of viewing CSAM, the pressure of large caseloads, and the strain on staff or budgets. Participants also spoke about the limits of technology and doubted whether three reviewers can ever guarantee accuracy.

Human Impact. All participants emphasized the psychological strain of repeated exposure. They said that triple verification increases this burden, especially in clear cases. P14 explained that this was why their organization moved from triple to double verification. P13 also observed that for children under 14, three reviewers were often excessive.

Content Volume. Thirteen participants described the growing amount of possible CSAM content. P6 compared the task to *"rolling the stone up the hill."* P10 noted that verification processes were already *"overloaded."* P14, whose organization uses a web crawler, explained that the inflow had become too much. P14 described triple verification as a *"massive waste of time and reviewers."* Other participants, such as P05 and P09, argued that while the process is costly, the investment is worthwhile given the protection it offers to its victims.

Technological Limitations. Six participants pointed out the subtleties of digital content. P01 explained that small changes, such as a single pixel or Snapchat filter, can create entirely new hashes. P5, P6, P12, and P13 added that duplicate checking slows down growth, but at the same time prevents uncontrolled expansion of the database.

Accuracy. Several participants questioned whether three reviews guarantee correctness. P06 reflected that subjectivity and different national laws make 100% accuracy impossible. P05 stated that *"Even though three individuals have checked the content, there is no guarantee that no mistakes are made."* P01 and P02 agreed that errors can still occur, even with multiple checks.

4.2.3 Necessity of Triple Verification. We examined how participants viewed the necessity of triple verification. Their perspectives show both support and doubt. Some considered the process important for accuracy and accountability, while others argued for a more flexible approach.

Process. Eight participants expressed clear support for triple verification, especially when human or legal consequences were at stake. P03 said that it was essential for accuracy in databases that affect human rights and judicial integrity. P08 added *"If it is so invasive that you are going to use censorship or intervene on human lives, I can imagine triple verifying is a good idea."* Six participants argued for a *"flexible eyes principle,"* suggesting a flexible approach to the number of verifications based on the case's complexity and content's clarity. P02 supported this approach; however, they asked how many reviews are *"too much."* P13 explained that obvious cases do not need three checks, while difficult cases sometimes need more.

Utility and Impact. Participants stressed that errors in databases can have serious consequences. P06 explained *"It is highly unlikely that an offender is sentenced based on just one image. The proof in court will always be based on solid classifications and images or videos that pose no questions."* P05 preferred *"a smaller more accurate database over a larger, less reliable one."* In contrast, P14 questioned the value of triple verification *"What are we actually doing this for? For this tiny fraction of it maybe we will catch some errors?"*

Human Factors. Several participants explained that reviewers bring different experiences to the table. P03, P05, and P13 said this shapes interpretation. P05 gave the example of disagreements over whether a child in the bath should be seen as erotic. For him, triple verification was needed to reduce such subjective differences.

Ethical Considerations. Training and ethical considerations were also mentioned during the need for triple verification. Seven participants stressed the importance of proper training to ensure reliable and consistent content assessment and to prevent potential over-censorship. When asked what training or education they received on how to assess and identify CSAM content, all participants said that they received the same training from Interpol and in-house training at their organization.

4.2.4 Future Perspectives and Opportunities. We examined how participants envisioned the future of verification processes. They described opportunities for technology to reduce workload and improve efficiency, but they also stressed the importance of human oversight remaining central.

AI and ML. Twelve participants mentioned artificial intelligence (AI) and machine learning (ML). P01-P04, and P06 saw opportunities for these technologies to support the first or second review. P04 described how AI could help at the initial stage, reducing the workload on human reviewers and increasing speed. P06 suggested a layered approach where AI handles early assessments to make the process more scalable. Others disagreed. *P09-P14* only saw a role for AI in prioritizing cases, not in making classifications or votes.

Oversight. Ten participants stressed the importance of maintaining human oversight. They worried about systems making autonomous decisions without interventions. P06 explained *"You see, people are afraid that AI systems will make automated decisions without human intervention. You don't want a robot determining for you that you're going to jail. So, you want the data to be accurate. I always want it to be at least one, but preferably two human pairs of eyes who say I've seen it and it's correct."* Although current practice requires triple verification, this participant suggested that in the future, a double-check system might suffice, provided that human oversight remains central. P12 reinforced this point, saying that AI can assist, but humans must control the process and ensure ethical handling. Overall, participants were divided on the future role of verification. Seven supported a double-check system, while seven supported triple verification.

Technological Diversification. Five participants discussed the value of diversifying technological methods. Four of them advocated perceptual hashing as an effective approach. P13 explained the advantage of working with companies that train classifiers. By running their software against tagged datasets, these systems could identify specific content, such as images of children within a certain age range.

4.2.5 Main Takeaways. In sum, the main takeaways from the interviews related to our participants' perceptions of triple verification are:

- **Triple verification is seen as a legal and ethical safeguard.** It helps prevent misclassification in ambiguous cases and supports compliance with diverse international laws.
- **The process improves quality but strains capacity.** While it boosts accuracy and reduces false positives, triple verification is costly, time-consuming, and emotionally taxing for reviewers.
- **Necessity depends on case complexity.** Participants supported flexible verification, with more eyes on unclear content and fewer on baseline material.
- **Future solutions should balance tech and human oversight.** AI and automation are welcome for filtering and prioritizing, but human judgment remains essential in final decisions.

4.3 RQ3: How do Different Verification Conditions Affect Inter-Rater Reliability in CSAM Classification?

To examine the accuracy of expert classifications, we conducted an inter-rater agreement experiment with CSAM. All items had been initially reviewed by a Dutch National Police reviewer as part of their operational workflow. For the purposes of this study, we recruited two additional expert reviewers from the Dutch National Police, who independently reclassified a total of 2,031 items consisting of images and videos. We analyze agreement across four conditions: (1) blind versus non-blind reviewing (i.e., whether prior votes were hidden or visible), (2) differences by file type, (3) the effect of voting order, and (4) outcomes when classification relies on double versus triple verification.

4.3.1 Blind vs. Non-Blind Conditions. Across both conditions, raters classified a set of 2,031 items. The exported logs include more rows because perceptual hashing groups near-duplicates to be verified as one item. In the system, these are then recorded again as separate entries; this yields 1,140 rows in the blind conditions and 1,373 in the non-blind conditions. In the blind conditions, raw agreement was 89.4% and Cohen's Kappa was 0.670, which is interpreted as *substantial* agreement. In the non-blind phase, agreement rose to 97.1% and Kappa increased to 0.893, which is *almost perfect*. The number of disagreements dropped from 117 files in the blind phase to 37 files in the non-blind phase, indicating that visibility of prior labels in the workflow helped reviewers converge. In both conditions, the dominant disagreement was around whether items belonged in *CSAM* or in the *Other* category. Overall, while higher agreement in the non-blind condition suggests that prior labels support convergence, it also raises the possibility of biases [35], where reviewers align with earlier votes ("follow the leader") rather than making fully independent judgements. Such influence would constitute a validity concern, as it reduces the independence of ratings and may mask underlying uncertainty. Therefore, one should interpret the non-blind gains in agreement with caution.

4.3.2 Impact of File Type. File type influenced the reliability of classification. In the blind condition, image files (n=938) yielded a Cohen's Kappa of 0.601, indicating *moderate to substantial* agreement, whereas video files (n=202) reached 0.812, corresponding to *almost perfect* agreement. In the non-blind condition, agreement for images improved (Kappa 0.893; n=1,138), while videos remained consistently high (Kappa 0.829; n=235). These results show that videos tend to provide clearer cues for classification across conditions.

4.3.3 Impact of Voting Order. We examined whether the order in the voting sequence influenced the inter-rater agreement in the non-blind phase. When reviewer A voted second, agreement occurred in 96.2% of the cases. When reviewer A voted third, agreement increased to 98.0% of the cases. A chi-square test indicated this difference was statistically significant, $\chi^2(1, N = 1373) = 4.151$, $p = 0.042$. Although the effect size is small, the pattern suggests that reviewer A may have been more sensitive to prior votes than reviewer B, leading to a slightly higher likelihood of convergence when A voted last. This shows that there is variance in the degree in which reviewers are influenced by prior votes.

4.3.4 CSAM Classification in Double vs. Triple Verification Conditions. Finally, we analyzed how agreement levels differ when CSAM classification is based on double versus triple verification. This comparison allows us to assess whether high agreement between two reviewers also extends to a broader consensus when a third reviewer is included. Due to limitations in the Dutch National Police system, the first review in the blind condition was not accessible, which prevented us from conducting this analysis in the blind setting.

Across the 1,373 files in the non-blind condition, the third rater agreed with the previous two raters in 1,257 cases (91.6%), while in 126 cases (8.4%), at least one rater disagreed. These results show that while two-rater agreement was already very high, but the inclusion of a third reviewer still surfaced a small but meaningful proportion of contested files. This highlights both the value of triple verification for judging difficult items and the persistence of residual uncertainty even under favorable conditions. The types of cases that triggered the divergent third rating were discussed in the later focus group.

4.3.5 Main Takeaways. In sum, the main takeaways from our annotation experiments with experts are:

- **Non-blind conditions yield significantly higher agreement.** Cohen's Kappa rose from 0.670 (blind) to 0.893 (non-blind), suggesting that access to previous labels enhances consistency among the reviewers, though this may partly reflect social influence bias.
- **Video content is classified more reliably than images.** Agreement was consistently higher for videos, indicating clearer classification cues.
- **Voting order subtly affects agreement.** Agreement was slightly higher when reviewer A voted third rather than second ($p = 0.042$), suggesting that reviewer A may have been more influenced by prior votes than reviewer B.

4.4 RQ4: What Challenges Cause Experts to Disagree When Classifying CSAM?

Here, we present our results from our focus group, where we aim to shed light on the challenges and potential reasons for disagreements during our inter-rater reliability experiment.

4.4.1 Series Recognition. Series recognition (22 mentions) emerged as the most critical challenge and reason for disagreements between the expert reviewers. In law enforcement practice, "series recognition" refers to the identification of whether an image or video belongs to a broader set of files that share substantive features and/or a common manner of creation. Dutch Supreme Court jurisprudence has established that a series may only be classified as CSAM if at least one item contains explicit child sexual abuse [51]. Once such an item is present, the connection with the other items allows the series to be deemed CSAM. The reviewers explained that single images can appear neutral when viewed in isolation, but they were classified as CSAM once recognized as part of a known illegal series. One reviewer noted, *"On its own, this image is not punishable, but I know it is part of a series. That makes it illegal."* The other reviewer added, "This is part of a known child pornographic series. We recognize the setup and the logo." Disagreements arose when only one reviewer recognized the broader series context, while another treated the file as a standalone item. This reliance on retrospective reasoning shows how CSAM classification is shaped by institutional workflows in police investigations, where reviewers often have access to large volumes of material rather than single items in isolation. We suspect that in practice, this means this source of disagreement is smaller than in our experiment.

4.4.2 Age Estimation. Age estimation (20 mentions) was another recurring theme that emerged in our focus group. Reviewers highlighted the challenge in judging physical maturity, particularly in low-resolution files or when ethnic variation complicated visual cues. One reviewer mentioned, *"I can't estimate the age properly. It's also someone of Asian appearance, which makes it more difficult."* Another relied on skin clarity: *"You can see it's a minor from the even, clean skin. No blemishes. That's how I recognize youth."* In several cases, reviewers referred to dental development as a clue: *"For me, it was the teeth. The development stage of the teeth gave it away."* When uncertainty persisted, they defaulted to the Other category: *"Even if I suspect she's underage, I can't say it with certainty. So I won't classify it as CSAM."* This strategy reflects an effort to avoid false positives, but it also helps explain why disagreements occurred: not all reviewers drew the threshold at the same point. While one reviewer might err on the side of caution and select "Other," another might classify the same file as CSAM based on the same ambiguous cues. In this way, individual differences in applying a conservative approach became a direct source of disagreement.

4.4.3 Pose, Framing, and Intent. Disagreements also stemmed from pose, framing, and intent (17 mentions). Images of minors who were clothed or partially clothed were interpreted differently depending on how the body was positioned and what the camera emphasized. One rater explained, *"She is sitting with her legs apart. The way it's photographed really draws attention to her underwear."* Others stressed the possibility of benign interpretations, such as *"It could just be a girl standing at a nudist campsite."* And for another item explained, *"If a parent takes a photo of their child in a playful moment, it doesn't mean it's sexual."* These differences illustrate how sexualization is often inferred rather than directly observed.

4.4.4 Animal Pornography. Animal pornography (6 mentions) introduced a distinct layer of complexity. While cases of direct sexual contact between children and animals were consistently classified as CSAM, several ambiguous examples created disagreement. In some, the framing or poor resolution made it unclear whether actual contact was taking place. Reviewers pointed out the difficulty of determining whether an image was photographic, virtual, or drawn, which affected whether it could be deemed punishable. In one contested video, a dog licked a young girl. One rater judged it as CSAM, citing visible signs of youth, while another hesitated, noting *"there is pubic hair, so the age is not clear."* In two cases, what was first classified as Animal Pornography was later reclassified as CSAM after closer inspection.

4.4.5 Textual Cues. Textual cues (7 mentions) also had a strong impact on CSAM classification. Captions or embedded text sometimes transformed an otherwise neutral image into CSAM. For example, one reviewer described an item where a child held a banana, accompanied by a caption: *"She's learning to take it deep in her mouth."* They explained, *"Without the text, I'd classify it as neutral. But the caption makes it sexual."* Yet the other reviewer expressed uncertainty about whether text alone was sufficient for legal classification. As one reflected, *"If someone types that in a chat, it makes the image CSAM. But if it's embedded text, I'm not sure."*

4.4.6 Conditions of Classification. Finally, reviewers reflected on the conditions of classification. In the non-blind setting, one acknowledged consulting peers' votes: *"In some doubt cases, I looked at what the others answered."* Yet professional responsibility remained paramount: *"Even if others say CSAM, if I can't defend that in court, I won't follow."* Reviewers also emphasized that trust in peers was selective, depending on familiarity: *"If it's someone I know and trust, I'm more likely to align. But if it's a new reviewer, I'm more critical."* Reviewers also raised concerns about volume pressure, with one remarking, *"I reviewed 120,000 images yesterday. I probably missed something, but it's unavoidable."* Another added, *"When it's a million pictures, you don't have time to zoom in on everything."* This suggests that in the face of the large scale of the task, reviewers try to avoid false positives but acknowledge that some false negatives may occur.

Overall, the focus group revealed that disagreements in CSAM classification by experts were not random errors but stemmed from structural interpretive challenges: contextual knowledge (series), subjective inference (age and intent), and working conditions (blindness, workload). Furthermore, in various situations, we found reviewers followed conservative approaches: not classifying an item as CSAM if they were not confident. This means they avoid false positives at the cost of some false negatives. In sum, CSAM classification depends strongly on context and working conditions, rather than being a straightforward technical task.

4.4.7 Main Takeaways. In sum, the main takeaways from the focus group and our analysis are:

- **Series recognition plays an important role in classifying CSAM in law enforcement.** Expert reviewers classified individual images as CSAM based on their experience and expertise that the image itself is part of a known CSAM series (e.g., based on the setup or logos), regardless of whether the image itself had CSAM material.

- **Age estimation is highly uncertain, especially across ethnicities.** Experts emphasized the difficulty in assessing the age of the depicted person, especially when they have a different ethnicity (e.g., Asian ethnicity). In such cases, raters relied on proxies like skin texture, dental development, and body proportions. Despite this, raters emphasized that image quality and ethnicity complicated their ratings.
- **Sexualized framing, not nudity, often drove CSAM classification.** Pose, angle, and visual focus influenced perception of intent. Disagreement between the experts arose when there was divergence in the interpretation of natural vs. sexualized imagery.
- **Non-blind conditions provided support, but not at the cost of autonomy.** Expert reviewers used peer input to resolve doubts, but retained the final responsibility for their decisions, also visible in some remaining level of disagreement.

5 Discussion

CSAM verification is a sociotechnical process shaped by three interconnected factors: (1) *Human Factors:* the judgements and limitations of human reviewers; (2) *Process Factors:* the organizational processes that structure and perform verification; and (3) *Technological Factors:* Technological systems that filter, cluster, or otherwise shape what humans see and decide. Together, these human, process, and technological factors determine both the accuracy and consistency of CSAM verification, as well as the emotional, operational, and legal consequences that follow from verification decisions. Below, we discuss our work around these three factors, allowing us to articulate concrete recommendations and implications that reflect the complexity of real-world CSAM governance. In addition, motivated by the rapid adoption of Generative AI, we discuss how AI-generated CSAM may shift this landscape. Finally, we discuss the limitations of our work.

5.1 Human Factors

Our findings reveal that disagreements in CSAM verification cluster around inherently human interpretive challenges. Reviewers struggled most with uncertain age estimation, ambiguous sexualized framing, and the recognition of series content based on contextual familiarity with prior cases. These disagreements are not random but reflect the reviewers' situated expertise, cultural background, and prior knowledge. Importantly, the consequences of such disagreements vary significantly across organizations. Specifically, our findings show that triple verification surfaced disagreements in 8% of the cases when compared to double verification. This non-negligible percentage of disagreements emphasizes that triple verification adds significant value in scenarios where a false positive rate of 8% will be deemed unacceptable. For instance, in criminal prosecutions related to CSAM, an 8% false positive rate is deeply problematic, as wrongful classification can compromise legal proceedings and infringe on due process. In such high-stakes contexts, triple verification provides the necessary safeguard to ensure evidentiary reliability. Yet, in other contexts, like content moderation on online platforms, the same 8% disagreement must be weighed not only against the risk of false positives (removing lawful content) but also against the risk of false negatives (failing to detect and remove abusive material). Both outcomes are harmful, though in different ways: one threatens freedom of expression, while the other undermines child protection. This creates a central tension for platforms: stricter verification increases certainty in each individual case but reduces overall throughput, leaving more material unreviewed. Conversely, prioritizing speed may increase coverage but at the cost of occasional wrongful removals – or wrongful non-removals, as indicated by reviewers indicating they did not have time to study images in more detail to look for cues of abuse.

Given these human factors, verification practices should adjust the level of verification depending on how the resulting classification will be used, the potential consequences of error, and how ambiguous the content is. Content that may be used for criminal investigations should undergo blind triple review by independent experts, while content used for preventive moderation or automated takedown (in the form of hashes) can typically rely on double review. Review environments should also be designed to minimize unnecessary exposure and psychological harm. Also, another concrete recommendation from our work is that reviewers should be able to flag uncertain items so that additional colleagues are consulted only when needed. These practices prioritize human expertise while reducing avoidable cognitive and emotional burden.

5.2 Process Factors

Process design plays a major role in shaping verification outcomes because it determines how reviewers encounter information, what cues they see, and how disagreements are handled. Our experiment shows that non-blind workflows substantially increase agreement (Cohen's Kappa 0.893 vs. 0.670), and that voting order can subtly influence decisions. These findings make it clear that workflows are not neutral administrative steps: they actively shape convergence and must be designed deliberately. This aligns with research from the HCI community that shows that content moderation is fundamentally a management of trade-offs between competing harms and values (precision vs. coverage, speed vs. due process) rather than a one-size-fits-all pipeline design. Jiang et al.'s [26] trade-off framework makes this explicit, showing how moderation choices operationalize different risk tolerances; our results extend that work to expert CSAM verification by offering empirical insights on disagreement patterns in CSAM classification workflows. Overall, how platforms resolve this tension is ultimately shaped by their policies and the legal environment in which they operate. Under the EU's Digital Services Act, for instance, trusted flaggers are expected to meet high accuracy standards, pushing platforms toward more conservative approaches. In contrast, in jurisdictions where the priority is rapid takedown of harmful material, efficiency may take precedence even at the risk of some over-removal. Our findings thus illustrate that verification procedures are not purely technical choices but reflect broader societal judgments about which form of harm (censorship or victimization) should be more heavily guarded against.

Also, our findings show that disagreements are not random but arise under specific, predictable conditions. Expert reviewers most

often struggled with cases such as uncertain age estimation or ambiguous contextual cues. This predictability opens up a path toward more adaptive verification procedures. Instead of applying triple verification universally, reviewers themselves could flag uncertain items for escalation to a third reviewer. In cases where the classification is clear, two independent evaluations would be sufficient to achieve high reliability. Such a selective escalation approach resonates with prior HCI research on content moderation. Jhaver et al. [25] showed that automation is most effective for handling routine tasks but fails on contextual details, underscoring the need for human review in ambiguous cases and a clearer division of labor. Similarly, Chandrasekharan et al. [6] demonstrate an example of a system on Reddit that can augment moderators by flagging difficult-to-detect cases for oversight, rather than replacing human judgment altogether. In addition, Hartmann et al. [23] highlight how algorithmic systems often over- or under-moderate nuanced content, and call for human oversight and collaboration with affected communities to ensure ethical decision-making. Finally, Bhardwaj et al. [2] found in their study of client-side scanning for CSAM that experts explicitly expect systems to include escalation mechanisms and safeguards against errors to maintain trust. Building on these, our work on the context of CSAM suggests that moderation workflows should integrate uncertainty signaling and tools for ranking and prioritizing ambiguous cases. This would reduce the emotional toll on professionals by limiting exposure to disturbing material [50], increase throughput by avoiding unnecessary third reviews, and maintain accuracy by focusing scrutiny where disagreement is most likely to occur. In doing so, organizations could preserve the trustworthiness of their notice-and-takedown requests while also addressing the pressing scalability challenges that currently threaten CSAM governance.

Process design also affects how and when non-blind review should be used. Although non-blind workflows are common in some settings where reviewers check each other's work, they introduce known risks. Seeing another reviewer's decision can lead to biases, especially when reviewers are tired or unsure [35]. This is a concern in ambiguous cases where independent judgment is essential. Because of these risks, non-blind review should not be the default approach. In law enforcement, where independence is vital for evidentiary integrity and legal consequences are significant, non-blind review is best limited to training or post-hoc calibration. In hotlines and platforms, it may be acceptable when speed is important, and safeguards are in place. When non-blind review is used, systems should include features that encourage independent thinking, allow reviewers to briefly explain their decisions, and support periodic audits to identify patterns that may indicate undue influence rather than genuine agreement.

5.3 Technological Factors

Technology shapes much of the verification process, from perceptual hashing and near-duplicate detection to series clustering, metadata analysis, and automated triage. However, our interviews show a clear pattern: although organizations welcome automation to reduce workload, none trust automated systems to determine illegality on their own. This mirrors HCI research showing that automated moderation often struggles with nuanced content and therefore requires strong human oversight [2, 23]. Our findings point to several useful technological opportunities, such as clustering tools that group related images to limit repeated exposure, ranking models that highlight ambiguous items for closer review, and provenance-aware metadata extraction that offers important contextual clues (e.g., systems that perform automatic series recognition). Note that some organizations already use tools such as clustering and series recognition to group related images and reduce repeated exposure, but this support is not consistently available across the ecosystem. We therefore recommend broader, more uniform adoption of such tools. To be effective, these tools must be transparent and accountable: automated decisions should be logged, auditable, and always reversible, with clear ways for reviewers to disagree. Finally, interfaces should be designed so that automation supports, and not replaces, expert human judgment, while still accommodating the differing workflows and legal responsibilities across organizations.

5.4 AI-Generated CSAM

The emergence of generative AI complicates CSAM verification in ways that manifest differently across organizations [7]. For hotlines, AI-generated CSAM risks overwhelming existing workflows with synthetic content that perceptual hashing can not reliably detect [41, 44]. For platforms, synthetic content blurs the boundary between illegal and harmful-but-legal material, raising new legal questions and increasing the difficulty of meeting notice-and-takedown obligations [31, 46]. For law enforcement, distinguishing synthetic from real child victims becomes critical for safeguarding evidentiary integrity and avoiding misclassification of fabricated content [7]. Preparing for these emerging challenges requires: (1) new taxonomies that differentiate real, synthetic, and manipulated content; (2) provenance-tracking mechanisms that record how an image or video was created and transformed [16]; and (3) systems capable of detecting AI-generated content. These capabilities will carry different stakes for different organizations: hotlines need these tools for workload management, online platforms need them for compliance and liability mitigation, and law enforcement needs them for accurate identification of victims. Overall, generative AI amplifies the need for robust human oversight, adaptive processes, and transparent procedures that aim to enhance the importance of the human, process, and technological factors discussed in this study [49].

5.5 Limitations

Our study brings together qualitative insights and experimental design in a domain where access to data, participants, and systems is inherently restricted. As such, our study has some limitations worth acknowledging. First, our interview sample was built through purposive recruitment of experts directly involved in CSAM detection and response. While this ensured high relevance and operational depth, it limited the diversity of perspectives. For instance, key organizations like NCMEC were not included due to non-response in our inquiry. Second, most interviews were conducted in Dutch, and although translations were carefully verified, some nuance may have been lost in presenting participants' quotes into English. Third, one interview took place in a group setting with four individuals. Although we ensured that each participant responded individually

to every question, the group context may still have shaped the openness, depth, or framing of their contributions due to social dynamics or institutional hierarchies. Our analysis did not explicitly model these potential sources of bias. Fourth, our experimental component was tightly constrained by the legal and operational environment of the Dutch National Police. The classification categories were predefined by law, and the dataset consisted of real-world items selected by experts working for the police rather than being randomly sampled. While this increased ecological validity, it reduced experimental control over distribution and item ordering. The design also relied on content that had already received an initial classification, which could subtly shape rater expectations. Finally, the experimental design was situated within the Dutch legal and operational context, meaning that the findings primarily reflect practices and standards in the Netherlands. While this setting offered unique access to real-world procedures, the results may not directly generalize to other jurisdictions where classification categories, legal definitions, and institutional norms around CSAM differ.

6 Conclusion

This paper investigated how CSAM verification practices are structured and experienced in relevant organizations, as well as investigating the inter-reliability of experts annotating CSAM content under different conditions. Through 14 expert interviews across seven organizations, we mapped the diversity of verification procedures and expert perceptions of double and triple verification. In collaboration with the Dutch National Police, we then conducted an inter-rater reliability experiment on 2,031 images and videos, testing how blind versus non-blind conditions and voting order influenced agreement among three reviewers. A follow-up focus group contextualized these results by surfacing the main sources of disagreement, including challenges of age estimation, series recognition, and sexualized framing. Taken together, our study shows that verification is shaped by trade-offs between accuracy, efficiency, and reviewer well-being. By empirically grounding these debates, our work contributes to ongoing discussions in HCI and policy about how to ensure CSAM verification quality while addressing operational and human costs.

Acknowledgments

This work is part of the project "EGOS: Effective Governance for cybersecurity and Online Safety" with file number KICH1.VE05.23.001 of the research programme "Cybersecurity Research Program 2023," which is financed by the Dutch Research Council (NWO) under the grant ID https://doi.org/10.61686/GXRYB36907. Also, this work was partially funded by an unrestricted gift from Google. We would like to thank Arda Gerkens (ATKM) and Ellen Janssen (ATKM) for their support and fruitful discussions that helped shape this work. Finally, we would like to thank our collaborators from the Dutch National Police for their help and support in conducting the inter-reliability experiment and for the focus group.

References

[1] Harold Abelson, Ross Anderson, Steven M Bellovin, Josh Benaloh, Matt Blaze, Jon Callas, Whitfield Diffie, Susan Landau, Peter G Neumann, Ronald L Rivest, et al. 2024. Bugs in our pockets: the risks of client-side scanning. *Journal of Cybersecurity* 10, 1 (2024), tyad020. doi:10.1093/cybsec/tyad020

[2] 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 2024 CHI Conference on Human Factors in Computing Systems*. 1–24. doi:10.1145/3613904.3642310

[3] Paul Bleakley, Elena Martellozzo, Ruth Spence, and Jeffrey DeMarco. 2024. Moderating online child sexual abuse material (CSAM): Does self-regulation work, or is greater state regulation needed? *European Journal of Criminology* 21, 2 (2024), 231–250. doi:10.1177/14773708231181361

[4] Steve Campbell, Melanie Greenwood, Sarah Prior, Toniele Shearer, Kerrie Walkem, Sarah Young, Danielle Bywaters, and Kim Walker. 2020. Purposive sampling: complex or simple? Research case examples. *Journal of research in Nursing* 25, 8 (2020), 652–661. doi:10.1177/1744987120927206

[5] Barry Chametzky et al. 2016. Coding in classic grounded theory: I've done an interview; now what? *Sociology Mind* 6, 04 (2016), 163. doi:10.4236/sm.2016.64014

[6] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30. doi:10.1145/3359276

[7] Caoilte Ó Ciardha, John Buckley, and Rebecca S Portnoff. 2025. AI Generated Child Sexual Abuse Material–What's the Harm? *arXiv preprint arXiv:2510.02978* (2025). doi:10.48550/arXiv.2510.02978

[8] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46. doi:10.1177/001316446002000104

[9] Christine L Cook, Jie Cai, and Donghee Yvette Wohn. 2022. Awe Versus Aww: The Effectiveness of Two Kinds of Positive Emotional Stimulation on Stress Reduction for Online Content Moderators. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–19. doi:10.1145/3555168

[10] Olivia Cullen, Keri Zug Ernst, Natalie Dawes, Warren Binford, and Gina Dimitropoulos. 2020. "Our laws have not caught up with the technology": Understanding challenges and facilitators in investigating and prosecuting child sexual abuse materials in the United States. *Laws* 9, 4 (2020), 28. doi:10.3390/laws9040028

[11] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 33–42. doi:10.1609/hcomp.v8i1.7461

[12] Elmira Deldari, Parth Thakkar, and Yaxing Yao. 2024. Users' Perceptions of Online Child Abuse Detection Mechanisms. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–26. doi:10.1145/3637424

[13] Hericson Dos Santos, Tiago S Martins, Jorge AD Barreto, Luis HV Nakamura, Caetano M Ranieri, E Robson, PR Geraldo Filho, and Rodolfo I Meneguette. 2024. ChaSAM: An Architecture Based on Perceptual Hashing for Image Detection in Computer Forensics. *IEEE Access* (2024). doi:10.1109/ACCESS.2024.3435027

[14] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3290605.3300372

[15] Hany Farid. 2021. An overview of perceptual hashing. *Journal of Online Trust and Safety* 1, 1 (2021). doi:10.54501/jots.v1i1.24

[16] KJ Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X Zhang. 2023. Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–42. doi:10.1145/3610061

[17] Konstantinos Kosmas Gaitis, Chrystala Fakonti, Zoe Lonard, Mengyao Lu, Jessica Schidlow, James Stevenson, and Deborah Fry. 2025. Legal challenges in tackling AI-generated CSAM across the UK, USA, Canada, Australia and New Zealand: Who is accountable according to the law? In *Searchlight 2025–Who Benefits? Shining a Light on the Business of Child Sexual Exploitation and Abuse*. 50–59.

[18] Lisa Geierhaas, Florin Martius, Arthi Arumugam, and Matthew Smith. 2025. "Not the Right Question?" A Study on Attitudes Toward Client-Side Scanning with Security and Privacy Researchers and a US Population Sample. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2246–2263. doi:10.1109/SP61157.2025.00086

[19] 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, 217–233. doi:10.1109/SP46215.2023.10179417

[20] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data &Society* 7, 2 (7 2020). doi:10.1177/2053951720943234

[21] Google. 2025. How image hashing technology helps NCMEC - Google Safety Center. https://safety.google/stories/hash-matching-to-help-ncmec/

[22] Enrique Guerra and Bryce G. Westlake. 2021. Detecting child sexual abuse images: Traits of child sexual exploitation hosting and displaying websites. *Child Abuse &Neglect* 122 (9 2021), 105336. doi:10.1016/j.chiabu.2021.105336

[23] David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and

Linguistic Variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26. doi:10.1145/3706598.3713998
[24] INHOPE. 2025. INHOPE Network. https://www.inhope.org/EN.
[25] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35. doi:10.1145/3338243
[26] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2023. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–34. doi:10.1145/3534929
[27] Stephanie Jones. 2022. Interpreting themes from qualitative data: thematic analysis. *Eval Academy* (2022).
[28] Sowmya Karunakaran and Rashmi Ramakrishan. 2019. Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 50–58. doi:10.1609/hcomp.v7i1.5270
[29] Juliane A Kloess, Jessica Woodhams, and Catherine E Hamilton-Giachritsis. 2021. The challenges of identifying and classifying child sexual exploitation material: moving towards a more ecologically valid pilot study with digital forensics analysts. *Child abuse & neglect* 118 (2021), 105166. doi:10.1016/j.chiabu.2021.105166
[30] Juliane A Kloess, Jessica Woodhams, Helen Whittle, Tim Grant, and Catherine E Hamilton-Giachritsis. 2019. The challenges of identifying and classifying child sexual abuse material. *Sexual Abuse* 31, 2 (2019), 173–196. doi:10.1177/1079063217724768
[31] Emmanouela Kokolaki and Paraskevi Fragopoulou. 2025. Unveiling AI's Threats to Child Protection: Regulatory efforts to Criminalize AI-Generated CSAM and Emerging Children's Rights Violations. *arXiv preprint arXiv:2503.00433* (2025). doi:10.48550/arXiv.2503.00433
[32] Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. 2020. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation* 34 (2020), 301022. doi:10.1016/j.fsidi.2020.301022
[33] Jessica McGarvie. 2023. From Hashtag to Hash Value: Using the Hash Value Model to Report Child Sex Abuse Material. *Seattle Journal of Technology, Environmental & Innovation Law* 13, 2 (2023), 4.
[34] Kimberly J Mitchell, Ateret Gewirtz-Meydan, David Finkelhor, Jennifer E O'brien, and Lisa M Jones. 2023. The mental health of officials who regularly examine child sexual abuse material: strategies for harm mitigation. *BMC psychiatry* 23, 1 (2023), 940. doi:10.1186/s12888-023-05445-w
[35] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651. doi:10.1126/science.1240466
[36] National Center for Missing & Exploited Children. 2022. *2021 CyberTipline reports by electronic service providers*. Technical Report. https://www.missingkids.org/content/dam/missingkids/pdfs/2021-reports-by-esp.pdf
[37] National Center for Missing & Exploited Children. 2023. CyberTipline Data. https://www.missingkids.org/cybertiplinedata
[38] Subin Park, Jeonghyun Kim, Jeanne Choi, Joseph Seering, Uichin Lee, and Sung-Ju Lee. 2025. HateBuffer: Safeguarding Content Moderators' Mental Well-Being through Hate Speech Content Modification. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–39. doi:10.1145/3757609
[39] European Parliament and Council of the European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act). 102 pages. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065
[40] Laurie Anne Pearlman and Paula S Mac Ian. 1995. Vicarious traumatization: An empirical study of the effects of trauma work on trauma therapists. *Professional psychology: Research and practice* 26, 6 (1995), 558. doi:10.1037/0735-7028.26.6.558
[41] Jonathan Prokos, Neil Fendley, Matthew Green, Roei Schuster, Eran Tromer, Tushar Jois, and Yinzhi Cao. 2023. Squint Hard Enough: Attacking Perceptual Hashing with Adversarial Machine Learning. In *32nd USENIX Security Symposium (USENIX Security 23)*. 211–228.
[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.
[43] Paul Rosenzweig. 2023. The Apple Client-Side Scanning System. *Lawfare* (2023).
[44] Priyanka Samanta and Shweta Jain. 2021. Analysis of Perceptual Hashing Algorithms in Image Manipulation Detection. *Procedia Computer Science* 185 (2021), 203–212. doi:10.1016/j.procs.2021.05.021
[45] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15. doi:10.1145/3411764.3445518
[46] Jessica Schidlow, Konstantinos Kosmas Gaitis, Mengyao Lu, James Stevenson, and Deborah Fry. 2025. Legal challenges in tackling AI-generated child sexual abuse material within the USA-REPORT. (2025).
[47] Carol F Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W Newman, and Sarita Schoenebeck. 2023. Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Article 341, 20 pages. doi:10.1145/3544548.3581512
[48] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17, 4 (2023). doi:10.5817/CP2023-4-8
[49] Chad MS Steel. 2024. Artificial intelligence and CSEM-A research agenda. *Child Protection and Practice* 2 (2024), 100043. doi:10.1016/j.chipro.2024.100043
[50] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14. doi:10.1145/3411764.3445092
[51] Supreme Court. 2010. The Case Law. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:HR:2010:BO6446
[52] Pat Walshe, Eija Hietavuo, Amaya Gorostiaga, Natasha Jackson, Jenny Jones, and Catherine Rutgers. 2016. Notice and Takedown: Company policies and practices to remove online child sexual abuse material. https://www.unicef.org/childrightsandbusiness/media/321/file/Notice-and-Takedown.pdf.
[53] Miranda Wei, Christina Yeung, Franziska Roesner, and Tadayoshi Kohno. 2025. "We're utterly ill-prepared to deal with something like this": Teachers' Perspectives on Student Generation of Synthetic Nonconsensual Explicit Imagery. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3706598.3713226
[54] Bryce Westlake, Martin Bouchard, and Richard Frank. 2012. Comparing methods for detecting child exploitation content online. In *2012 European intelligence and security informatics conference*. IEEE, 156–163. doi:10.1109/EISIC.2012.25
[55] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3290605.3300390

A Organization Details

Table 3 summarizes the number of interviews per organization, along with the roles and responsibilities of each organization.

B General Interview Protocol

Our interviews were conducted based on the following procedure. First, we included some questions aimed at better understanding the organizations and their verification procedures (see Introduction). Then, we had a specific set of questions for Law Enforcement and a different set for the other organizations (see Law Enforcement Agency Interview Protocol and Other Organizations Interview Protocol, respectively). Finally, all interviews were concluded with some questions regarding the future and participant comments (see Closing Remarks). We provide all the questions for each of the above-mentioned parts below.

Introduction

(1) Could you please introduce yourself and describe your organizational role?
(2) What are your primary responsibilities, and how do they relate to content moderation/assessment?
(3) Could you explain how your organization's content verification processes are structured?
(4) Does your organisation differentiate verification processes based on the type of content? If so, how?
(5) How much capacity do you have for the verification processes of content moderation in your organisation?
(6) Are you familiar with the costs of content moderation processes? If so, how much are they? (Or could you explain if the costs are justified for the results?)

Table 3: Overview of interview participants by organization, number of interviews, and roles.

Organization	Number of interviews	Roles and Responsibilities
ATKM	4	Verification
C3P	1	Verification and hash database management
Dutch National Police	2	Verification and hash database management
GIFCT	1	Hash database management
Offlimits	4	Verification
INTERPOL	1	Verification and hash database management
IWF	1	Verification and hash database management

(7) What benefits have you observed from implementing multiple layers of verification in content moderation?
(8) Have you identified any specific limitations or drawbacks with these verification processes?
(9) Can you describe any differences in content verification accuracy or efficiency between your verification processes?
(10) Are you familiar with triple verification? Could you maybe tell me your understanding of this method?

Law Enforcement Agency Interview Protocol

(1) In cases involving CSAM or other serious offenses, how does the verification process impact your work?
(2) How do you collaborate with organisations in content moderation during investigations? Are there formal processes or informal arrangements?
(3) What role, if any, does your agency play in the verification process of illegal content before it reaches a judicial setting?
(4) What do you think are the benefits or drawbacks of having multiple stages of verification for content like CSAM?
(5) From your experience, how does the verification of such content affect the outcomes of legal cases involving CSAM?
(6) In cases where verification is handled by another entity, how do you see that affecting your own process or responsibilities?
(7) What are some ethical considerations you think are important in the verification of sensitive content?
(8) What challenges do you face in the context of content verification as it relates to law enforcement?
(9) How do you think verification processes for online content could be improved, especially from a law enforcement perspective?
(10) Do you have an idea why the call for triple verification is stronger for CSAM than for terrorist content?

Non-Governmental Organizations Interview Protocol

(1) What impact has triple verification had on the accuracy and reliability of your content moderation outcomes?
(2) Triple verification is often resource-intensive. How does your organisation handle the resource demands of this process? Have there been any strategies to mitigate these demands?
(3) What specific challenges arise solely from the triple verification aspect of your content moderation? How do these challenges affect your overall moderation workflow?
(4) Based on your experience, what improvements or optimisations would you suggest for verification processes to make it more effective or less resource-intensive?
(5) In your opinion, is triple verification necessary for all types of content your organization handles, or are there certain types where it's more crucial? How do you decide?
(6) Do you have an idea why the call for triple verification is stronger for CSAM than for terrorist content?
(7) Are there any alternative verification methods your organization has considered or implemented to maintain or improve content verification quality?
(8) What innovations or technologies are being looked at to potentially enhance the verification process without escalating resource commitments?

Closing Remarks

(1) How do you foresee the verification and assessment process evolving in the next few years?
(2) Is there anything else you believe is important to discuss regarding this topic that we haven't covered yet?

C Codebook

Table 4 provides an overview of our codebook, including the main themes, sub-themes, codes, and examples of quotes supporting them.

Table 4: Overview of our codebook, including the main themes, sub-themes, the codes, and example of quotes supporting them.

Main Theme	Sub-Theme	Code	Example Quotes
Benefits	Quality	Accuracy of hash database, Quality of hash database	*"We want to ensure that we have an as clean database as possible"* (P07), *"For accuracy for their database as a standing argument, I get it"* (P01)
	Content Assessment	Age estimation in CSAM is difficult, Estimation of sexual act is difficult, Sexual pose is sometimes hard to determine	*"When sometimes also in the team, if I have like three pictures and I show these three pictures to my colleagues and said can you estimate the age of these children? The result might be different. Verification is much, it is so hard. When are we talking about it is the child preparation or not. Do we also see the sexual activity? What is sexual activity touching? To genital area or touching the belly?"* (P12)
	Legal Considerations	Law can sometimes be indistinct, Judgement accordingly to the law and regulation is essential, On what grounds do you add a hash	*"You do have to look at the letter of the law"* (P05), *"We do triple verification to have with some certainty, say you have looked at least 3 times to determine the legal basis from different perspectives. What exactly is it?"* (P06)
	Inconsistent assessment	Humans make mistakes, Context is important and hard to classify	*"Because yes, people make mistakes too"* (P02), *"People just make mistakes, just stupid mistakes"* (P05)
	Operations	Assessing illegal content does not take excessive time; assessing illegal content is not an exact science	*"It is people's work. There is a danger that you might find something CSAM that I don't. It can be because of your beliefs, your childhood, how you think about sex.... It is just not an exact science. It's still about what I find and whether you find it CSAM"* (P05)
Challenges	Resource and Cost	Costs a lot of (human) resources, Expensive process, Continuous work of assessing illegal content	*"I think it is mainly just that teams don't have human resource pipelines like for doing triple verification, only the very largest companies have so much resources that they can put towards certain things"* (P07)
	Technological Limitations	Different hash value if picture changes, Database grew very little and took time to update	*"The need for the database to be good, so it has to be checked three times. So we have done that now, but we saw because of that is that our database grew very little"* (P05)
	Content Volume	The amount of content is growing, A lot of content is waiting to be checked, Lot of duplicates, There will always be a backlog	*"No and so that's actually also a problem in CSAM moderation then. You don't have people to kin of, to do this in a way that you can deal with that volume"* (P01)
	Human Impact	The burden to have three people watch the same content, Impact assessing CSAM content on moderators, Not enough analysts to moderate the content	*"And the disadvantages? You burden three people with it"* (P02)
	Accuracy	There is no 100% certainty in triple verification, After matching hash database, some content needs rechecking, The process of classifying content is difficult	"The downside is; it is complex. That is definitely true, the downside is that you almost never get the 100 percent accuracy and certainty and you have to live with that" (P06)
Necessity	Utility and Impact	Hash database is used by other parties, Hash databases can have major consequences, What happens if we have a larger database?	*"It has quite a lot of of consequences"* (P04), *"I don't want people to suffer from this"* (P05)
	Process	Multiple eye principle should be the baseline, In ambiguous cases triple verification is useful, Understanding for the triple verification process	*"But if it is so invasive that you are going to use censorship or intervene on human lives, I can imagine that triple verifying is a good idea"* (P11)

Continued on next page

Table 4 Continued from previous page

Main Theme	Sub-Theme	Code	Example Quotes
	Human Factors	Assessing content remains subjective, Experience of analyst can influence outcomes, Experience can overrule triple verification	*"Yes you know. It's just always going to be subjective anyway. In my opinion. Assessing content. I find it very vulnerable"* (P04)
	Educational and Ethical Considerations	Training and education is of importance, I don't want people to suffer because of what I've seen, People depend on how we categorize content	*"It is important to have really well-trained, motivated officers doing the 1st and 2nd if you only have like double verification"* (P12)
Opportunities	AI and ML	AI could be an addition to the current system, AI models can estimate age and detect specific content, AI or ML models trained well enough could take over, More research on AI and ML models	*"I do think AI can start helping us with triple verification. You could say, let the first or let the second verification be done by an AI tool, huh?"* (P06)
	Automation	Automation is going to happen, Combine different models to enhance accuracy, Preselection of CSAM content to streamline processes	*"And so I think in the case of AI, you can also do something with that. If you could recognize AI on the front end, then, of course, you could also start classifying in a very good way"* (P08)
	Oversight	There should always be human oversight, Double verification could be sufficient in many cases, No future for double verification	*"You should not remove the human eye of the process"* (P02)
	Technological Diversification	Different types of hashing techniques could be explored	*"I guess that is also the difference. I guess everyone is now moving in the direction of perceptual hashing, but not all forms of hashing are open"* (P07)