Beyond Accuracy: A Mixed-Method Exploration of Hash Database Verification

Focusing on the Detection of Child Sexual Abuse Material and Terrorist Content Online

Melissa Rottier (6080693)

Delft University of Technology

Beyond Accuracy: A Mixed-Method Exploration of Hash Database Verification

Focusing on the Detection of Child Sexual Abuse Material and Terrorist Content Online

by

Melissa Rottier

to obtain the degree of Master of Science in Complex Systems Engineering and Management

at the Delft University of Technology,

to be defended publicly on July 17th, 2025

Student number:	6080693	
Project duration:	February 3rd, 2025 – July 17th, 2025	
Thesis committee:	Prof. dr. M.J.G Van Eeten,	TU Delft, Chair
	Dr. S. Zannettou,	TU Delft, First Supervisor
	Dr. M. Kroesen,	TU Delft, Second Supervisor
	A. Gerkens,	ATKM, External Supervisor
	E. Janssen,	ATKM, External Supervisor

On behalf of the Authority for the prevention of online Terrorist Content and Child Sexual Abuse Material

We never forget that each one [hash] is an image of a real child being sexually abused;
an online record of suffering and pain;
the documentation of a crime scene and a place of terror for that child.

— Internet Watch Foundation (2025) [1]

[1] Retrieved and cited from paragraph 5: https://www.iwf.org.uk/our-technology/our-services/image-hash-list/

Preface

Dear reader,

As I reach the conclusion of this journey, I find myself reflecting on my academic career, which has been a rollercoaster of experiences, from navigating the challenges of the COVID pandemic to the rewarding culmination of my thesis.

This research exposed me to a subject that was entirely new to me. It has been truly inspiring to meet and engage with dedicated individuals working in this field. Their passion, expertise, and commitment have been truly inspiring and have significantly shaped my understanding of the topic. I am grateful for the opportunity to contribute, in a small way, to this important cause, and I will carry these lessons learned with me into my future.

I would like to express my gratitude to Prof. Dr. Michel van Eeten for facilitating this internship opportunity. I am also deeply thankful to my first supervisor, Dr. Savvas Zannettou, whose weekly meetings and support were invaluable. Your guidance kept me focused and provided a space for me to discuss all my thoughts and concerns.

A special mention goes to my colleagues in ATKM, whose genuine interest in my research and participation made this journey even more meaningful. I owe a debt of gratitude to Arda Gerkens and Ellen Janssen, who consistently supported me. Our discussions about the project were not just instrumental in helping me navigate this process, but also motivating and full of enthusiasm and excitement.

I would also like to extend my genuine thanks to the police department for their invaluable assistance in conducting my experiment. The quantitative part of my research could not have been realized without their support, and I am extremely grateful for the opportunity to collaborate with them.

Finally, I want to acknowledge my family and friends. A special shout-out to my boyfriend, Morris, who patiently listened to me talk about my thesis every day. I can only imagine how much that must have been! Your support has meant the world to me, and I couldn't have done this without all of you.

Melissa Rottier
Rotterdam, July 2025

Summary

The spread of Child Sexual Abuse Material (CSAM) and Terrorist Content Online (TCO) remains a pressing societal issue. Various organizations rely on hash databases to detect, flag, and remove harmful content. These databases function as storage of digital fingerprints of previously identified illegal material, enabling automated platform filtering. However, the effectiveness and reliability of such databases rely on the verification processes used to determine what content qualifies for inclusion.

This thesis investigates the characteristics of verification processes in CSAM and TCO hash databases, with a particular focus on triple verification. Using a multiphase mixed-methods design, the study integrates qualitative insights from stakeholder interviews, an annotation experiment, and a follow-up focus group with annotators.

The interviews with experts highlighted variations in verification workflows, ranging from single-rater decisions to triple verification models. While triple verification is seen as a standard for increasing trust and minimizing false positives, its feasibility in terms of emotional toll and volume has been questioned. Thematic insights centered around benefits (e.g., legal considerations), challenges (e.g., emotional toll, inconsistent thresholds), necessity (e.g., utility and impact), future opportunities (e.g., automation), and differences between CSAM and TCO workflows.

In the experiment, two raters from the Dutch National Police classified 2,031 real potentially illegal items under two different conditions. In the blind phase, raters voted independently, whereas in the non-blind phase, prior votes were visible. Overall inter-rater agreement rose from 89.4% in the blind condition to 97.1% in the non-blind condition. A statistically significant association was found between voting order and agreement rates, suggesting that seeing one or two prior votes can subtly influence rater alignment.

The focus group offered further insight into the found disagreements. A key theme was the importance of recognizing image series: individual images were often reclassified as illegal when identified as part of a known CSAM series. Age estimation was also a recurring source of ambiguity, particularly when visual quality was poor or when victims' physical development and ethnicity made assessment difficult. Raters relied on indicators such as skin texture, body proportions, and dental features, though these cues were often interpreted differently.

The findings emphasize the need for verification systems that are both flexible and context-sensitive. Not all cases require the same level of scrutiny: while baseline CSAM could be classified with fewer checks, ambiguous cases require more checks. Rather than enforcing uniformity, organizations should accommodate interpretive differences while safeguarding consistency and accountability.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
ATKM	Autoreit Online Terroristisch en Kinderpornografisch Materiaal (in English = Authority for the prevention of online Terrorist Content and Child Sexual Abuse Material)
CoSEM	Complex Systems Engineering and Management
CSAM	Child Sexual Abuse Material
CSE	Child Sexual Exploitation
C3P	Canadian Centre for Child Protection
DSA	Digital Services Act
EU	European Union
GIFCT	Global Internet Forum Counter Terrorism
IWF	International Watch Foundation
I&C	Information and Communication
ISP	Internet Service Provider
ML	Machine Learning
NCA	National Competent Authority
NetzDG	Network Enforcement Act
NGO	Non-Governmental Organization
NLP	Natural Language Processing
NTD	Notice-and-Takedown
NCMEC	National Center for Missing & Exploited Children
TCO	Terrorist Content Online
UCG	User-Generated Content
VLOP	Very Large Online Platform
VLOSE	Very Large Online Search Engine

1

Introduction

1.1. Background

Digital platforms are central to modern life. They offer spaces for communication, self-expression, and the sharing of information. However, their open and accessible nature also enables the spread of harmful and illegal content (Schneider & Rizoiu, 2023). Among the most serious forms are Child Sexual Abuse Material (CSAM) and Terrorist Content Online (TCO), both of which continue to rise in prevalence and severity (Aiken et al., 2019; Arora et al., 2023; Trivison, 2024; Wolbers et al., 2025).

The National Center for Missing and Exploited Children (NCMEC) documented over 29 million CSAM-related reports in 2021 alone (National Center for Missing & Exploited Children, 2022), with a 300% increase in cases between 2021 and 2023 (National Center for Missing & Exploited Children, 2025). TCO also remains an urgent concern due to its role in radicalization and incitement to violence (Macdonald et al., 2024). Removing such content is both a legal and societal obligation.

Moderating CSAM and TCO presents unique and complex challenges (Seigfried-Spellar et al., 2024). Accurate identification in these areas is critical; however, distinguishing minors from adults, as well as determining the context in potentially dangerous material, often requires nuanced judgment, which increases the risk of errors (Macdonald et al., 2024; Oronowicz-Jaśkowiak et al., 2024). Misclassification, whether failing to detect harmful content or flagging legal material as illegal, undermines the credibility and effectiveness of content moderation systems (Seigfried-Spellar et al., 2024).

One of the most widely used tools in this effort is hash-based detection. Hashes are digital signatures of previously identified illegal content. When a file is verified and hashed, platforms can detect and block it without requiring another manual review (Lee et al., 2020). The effectiveness of hash databases depends on the quality of the verification process that precedes hashing. If a mistake is made during verification, such as adding legal content to a database or failing to include harmful content, the consequences can be severe.

To reduce the chance of such errors, some organisations implement multi-step review processes (Farid, 2021; Google, 2025; National Center for Missing & Exploited Children, 2023). One of these is triple verification, in which three independent reviewers must confirm that the content is illegal before it is hashed (Farid, 2021). This approach is intended to improve accuracy and reduce the risk of false positives. However, multiple reviewers come with their trade-offs. It requires more time, more coordination, and places a greater emotional burden on reviewers who are repeatedly exposed to harmful material (Kloess et al., 2021; Meggyesfalvi, 2024). As the volume of content continues to increase, the long-term feasibility remains uncertain (Spence et al., 2023).

1.2. Research Objective

Although triple verification is applied in practice, it has received little attention in academic research. Most studies focus on automated technologies or the psychological effect of human moderation (Bleakley et al., 2024; Schneider & Rizoiu, 2023; Spence et al., 2023), but few investigate the organisational and procedural decisions of verifying harmful content (Kloess et al., 2019, 2021). As a result, there is a limited understanding of the different verification methods implemented, how stakeholders perceive them, and how more intensive approaches, such as triple verification, add value.

Three knowledge gaps guide this thesis.

First, there is limited academic insight into how multi-step verification procedures are operationalized. Although triple verification is used by several institutions (Google, 2025; National Center for Missing & Exploited Children, 2023), most descriptions remain high-level. It is unclear whether review processes occur sequentially or in parallel, how disagreements are resolved, and how these practices integrate with broader organizational workflows. This research examines those operational structures across multiple organizations involved in CSAM and TCO moderation.

Second, while institutions advocate for more intensive review to improve accuracy, this comes at a cost. Human moderators face increasing workloads and emotional strain due to repeated exposure to traumatic content (Barrett, 2020; Gewirtz-Meydan et al., 2024; Spence et al., 2023). In addition, more content is being generated every year, raising concerns about the long-term scalability of multi-step procedures (Bonagiri et al., 2025; Gillespie, 2020). While foundational works have examined content moderation labor conditions in general (Gillespie, 2020; Roberts, 2017), few studies have explored how professionals perceive multi-step procedures in practice. This study contributes to this gap by conducting interviews with domain experts to explore their perspectives.

Third, no prior empirical work has evaluated the impact of different voting conditions on classification outcomes in CSAM moderation. While other fields (e.g., medical imaging, academic peer review) show that blind and anonymous annotation improves agreement and reduces bias (Du et al., 2024; Rastogi et al., 2024), these insights have not been applied to CSAM. This study addresses this gap by conducting an annotation experiment in which experienced Dutch National Police reviewers assess 2,031 items under both blind and non-blind conditions.

The central research question guiding this study is:

What are the characteristics of verification processes, including triple verification, for hash databases used in moderating online harmful content?

To better understand this overarching question, four sub-questions have been derived to answer the main research question:

- SQ1: Which key concepts, definitions, and principles form the foundation of content moderation and assessment in the context of CSAM and TCO?
- SQ2: What processes do various organizations use to add hashes to databases that contain CSAM or TCO?
- SQ3: How do organizations involved in combating CSAM and/or TCO perceive the triple verification processes of hash databases?
- SQ4: How do different verification approaches affect the final classification of potentially illegal content?

1.3. Linkage with CoSEM Master Program

This thesis aligns with the Complex Systems Engineering and Management (CoSEM) program, especially the Information and Communication (I&C) Track, by addressing a complex socio-technical challenge within a digital information infrastructure. The research bridges the public and private sectors by examining the interaction between technology, humans, and regulatory frameworks. By addressing the systemic complexities and governance of publicly and privately owned systems, the thesis exemplifies the I&C track's goal of improving innovative processes for safety and security in digital environments.

1.4. Academic and Societal Relevance

The academic relevance of this research is highlighted by the examination of verification processes in content moderation systems, particularly those addressing CSAM and TCO. Despite the growing reliance on hash databases, there is a notable lack of empirical research, identified in Chapter 2, into these systems' verification processes.

The societal relevance of this research lies in the need to establish a safe online environment while addressing the issues of CSAM and TCO. Analysts' work, which verifies content, is essential for protecting vulnerable individuals. Additionally, it is important to acknowledge the victims and their families, whose suffering is exacerbated by the continued availability of these materials online.

1.5. Report Structure

This report is structured as follows. Chapter 2 presents related work on the subject and identifies three knowledge gaps. Chapter 3 outlines the methodological framework employed, which is a multiphase mixed-methods approach. Chapter 4 discusses the findings derived from both qualitative and quantitative research. Finally, Chapters 5 and 6 synthesize the insights gained from the research and address the overarching research question.

2
Related Work

This chapter aims to synthesize existing work on online content moderation and assessment, with a specific focus on the detection and verification of CSAM and TCO. It will provide an answer to the following sub-question: *SQ1: Which key concepts, definitions, and principles form the foundation of content moderation and assessment in the context of CSAM and TCO?*

2.1. Method

The literature search was conducted using Google Scholar and Scopus. These databases were selected for their broad interdisciplinary coverage and inclusion of peer-reviewed articles relevant to the research topic. Due to the limited amount of empirical academic studies on verification in moderation systems, grey literature, including governmental publications, non-governmental (NGO) reports, and technical standards, was included.

Sources were selected based on their relevance to the core themes of online content moderation, detection, and verification, with a particular emphasis on CSAM and TCO. Preference was given to sources published within the past ten years, although foundational or policy-defining works predating this window were retained for contextual completeness.

The following will be discussed per section and is aligned with the identified topics in literature:

- Section 2.2 outlines the legal and regulatory frameworks governing online content moderation;
- Section 2.3 describes the multi-stakeholder environment in which content moderation operates;
- Section 2.4 introduces the spectrum of moderation practices;
- Section 2.5 provides the identified knowledge gaps and motivation for this research.

2.2. Definition and Scope of Illegal Content

Legislative frameworks at both national and supranational levels increasingly shape the governance of harmful online content. Within the European Union (EU), a layered and evolving regulatory structure defines the responsibilities of digital platforms, hosting providers, and enforcement bodies. This section examines the legal categorizations of harmful content and the most relevant EU laws that underpin current moderation practices.

2.2.1. Typology of Harmful Content

To contextualize regulation, it is necessary to clarify what constitutes harmful content. Banko et al. (2020) identifies four overarching categories of online harm (See Figure 2.1): hate and harassment, self-inflicted harm, ideological harm, and exploitation. CSAM is categorized under exploitation, involving material that sexually abuses or exploits individuals under 18 years old. TCO falls under ideological harm, encompassing material used to promote, glorify, or coordinate acts of political violence.

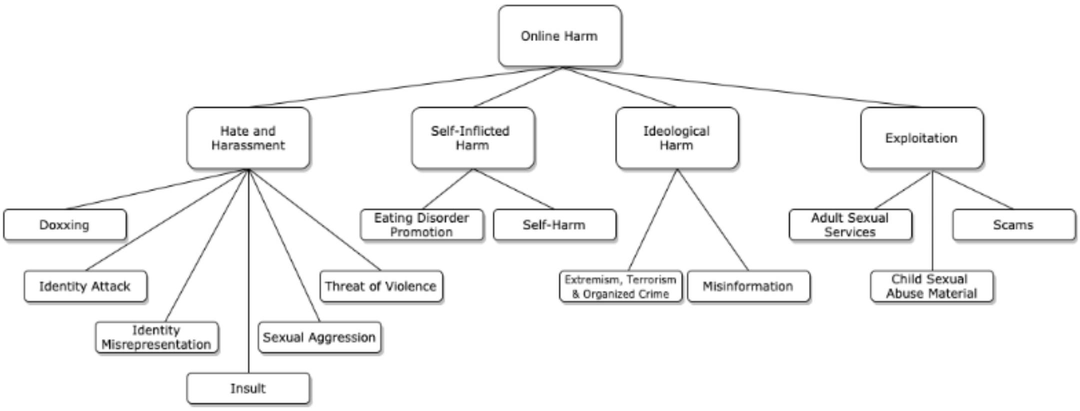

Figure 2.1: Typology of Harmful Content (Banko et al., 2020)

Although various forms of harmful content exist, only four are explicitly prohibited EU-wide: (i) CSAM, (ii) hate speech that is racist or xenophobic, (iii) terrorist content, and (iv) material infringing on intellectual property rights (De Streel et al., 2020). Other forms of harmful content are inconsistently addressed across Member States, often falling into the categories of "legal but harmful" or "unregulated." This fragmentation presents challenges for content moderation that require cross-border coordination.

2.2.2. Overview of EU Regulatory Instruments

Since the early 2000s, the EU has developed a suite of legal instruments aimed at coordinating digital governance. These range from foundational directives on e-commerce to recent regulations like the Digital Services Act (DSA). Table 2.1 provides a chronological summary of noteworthy regulatory milestones relevant to online content moderation.

The foundation of EU regulation, the E-Commerce Directive (2000), exempts intermediary service providers from liability for illegal content, provided they act promptly once notified (European Union,

Table 2.1: Overview of EU Regulations on Content Moderation (Grippo, 2024)

Year	**Regulation**	**Description**
2000	E-Commerce Directive (2000/31/EC)	Sets online services' framework with intermediary liability exemptions.
2010	Audiovisual Media Services Directive (2010/13/EU)	Updates EU broadcast rules for online videos, enhancing minor protection and hate speech restrictions.
2011	Directive on combating the sexual abuse and sexual exploitation of children and child pornography (2011/93/EU)	Mandates actions against child exploitation content.
2018	General Data Protection Regulation (GDPR)	Regulates the processing of personal data, impacting content moderation practices regarding user data and privacy.
2018	EU Code of Practice on Disinformation	Establishes a voluntary framework to combat online disinformation, enhancing transparency.
2019	Copyright and Related Rights Directive (2019/790)	Balances creators' rights with user freedoms on platforms.
2021	Regulation on Dissemination of Terrorist Content Online (EU 2021/784)	Mandates swift removal of terrorist content within one hour of notification.
2022	Digital Services Act (DSA)	Modernises digital space, mandates rapid illegal content removal, and increases transparency and accountability.

2024). Importantly, the Directive forbids imposing general monitoring obligations, meaning providers are not required to pre-emptively scan all content. This liability framework has been supplemented by topic-specific directives, including the 2011 Child Sexual Abuse Directive, which mandates removal of CSAM, and the 2021 Regulation on Terrorist Content, which requires designated platforms to take down flagged terrorist content within one hour (Grippo, 2024).

The most comprehensive regulatory update is the DSA, enacted in 2022. It applies to all intermediary services but places additional obligations on Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). Key requirements relevant to content moderation include (retrieved from European Commission (2025a)):

- **Article 14:** Requires platforms to inform users of content restrictions and explain moderation practices.
- **Article 16:** Mandates effective user-friendly reporting mechanisms for illegal content.
- **Article 17 & 20:** Guarantees transparency in moderation decisions and fair complaint procedures.
- **Article 34–37:** VLOPs and VLOSEs must assess and mitigate systemic risks, publish annual

moderation reports, and undergo audits.

- **Article 22:** Introduces "trusted flagger" status for verified organizations, whose reports must be prioritized.

The DSA marks a shift toward proactive platform accountability while still upholding due process and users' rights. Its implementation provides the legal foundation for moderation workflows.

2.2.3. Trusted Flaggers in Regulatory Context

A key actor introduced by the DSA is the formal recognition of trusted flaggers, detailed in Article 22. These are entities granted special status by national Digital Services Coordinators due to their proven expertise in identifying specific categories of illegal content. Trusted flaggers form a regulatory mechanism to enhance the accuracy and efficiency of the notice-and-takedown (NTD) process, addressing long-standing concerns around overreporting, false positives, and procedural delays (European Commission, 2024).

According to the European Commission, trusted flaggers are organizations that:

> *"are experts at detecting certain types of illegal content online, such as hate speech or terrorist content, and notifying it to online platforms. The notices submitted by them must be treated with priority as they are expected to be more accurate than notices submitted by an average user."* (European Commission, 2024, para. 1)

Organizations must demonstrate subject-matter expertise, operational independence from online platforms, and a commitment to due process and transparency to qualify for this status (European Commission, 2024). Trusted flagger status applies EU-wide and obliges platforms to prioritize reports from trusted flaggers over those from ordinary users. They must also maintain a high level of reporting quality; repeated submission of inaccurate or unsubstantiated notices can result in the revocation of their status.

Furthermore, trusted flaggers are subject to annual transparency obligations. They are required to publish public reports that include: (i) the number and types of notices submitted, (ii) the categories of content flagged, and (iii) actions taken by providers in response to the notices (European Commission, 2024). This reporting framework is intended to safeguard against misuse while strengthening the legitimacy and public accountability of the flagging process (Appelman & Leerssen, 2022).

Trusted flaggers operate primarily through structured notice-and-takedown requests. These notices must meet certain regulatory standards to be legally actionable (European Commission, 2024):

- Clearly state the reason why the content is considered illegal, referencing specific laws or regulations;
- Provide a precise URL or digital location of the infringing content;
- Include the notifier's contact details, with anonymity safeguards for serious offenses;
- Include a good-faith affirmation of the accuracy of the report.

As Van De Kerkhof et al. (2024) argue, trusted flaggers play a dual role: they serve not only as technical moderators but also as policy actors, guiding the interpretation of ambiguous content categories and providing input into platform governance. Their designation reflects a shift toward co-regulation, where civil society and state actors jointly enforce digital norms.

2.3. Multi-Stakeholder Environment

The governance of online content moderation exists within an environment where diverse actors collaborate to identify, evaluate, and remove harmful material. This multi-stakeholder model relies heavily on coordination and information-sharing across sectors. Governments create legal structures; platforms develop and enforce internal policies; hotlines and trusted flaggers serve as intermediaries; and law enforcement acts upon verified intelligence. As noted by Mulligan and Bamberger (2021) and Bleakley et al. (2024), the boundaries between regulation and implementation are increasingly blurred, particularly in co- and self-regulatory arrangements.

This section outlines the specific roles in content moderation of governmental and regulatory bodies, hosting and platform providers, law enforcement agencies, hotlines, and national competent authorities (NCAs).

2.3.1. Governmental and Regulatory Bodies

Governmental and regulatory institutions play a role by defining what constitutes illegal content and designing regulatory mechanisms. These efforts may take the form of either hard law, such as binding legislation, or softer, co-regulatory models. In co-regulation, public authorities and private entities jointly shape enforcement mechanisms. According to Mulligan and Bamberger (2021), the policies enacted by platforms may stem from direct obligations or be shaped indirectly through public policy expectations.

For instance, Germany's Network Enforcement Act (NetzDG), implemented in 2018, mandates that social media platforms promptly remove hate speech or terrorist content deemed illegal under German law, or face fines up to 50 million euros (Claussen, 2018; Wagner et al., 2020). Parallel to such laws, public actors may promote voluntary codes of conduct, as noted by Bellanova and De Goede (2022), encouraging private compliance without formal legislation.

In contrast, self-regulation refers to decisions made autonomously by industry actors to enforce moderation practices that align with internal standards or public expectations. This distinction, between co-regulated and self-regulated approaches, shapes the degree of accountability and transparency in moderation practices (Bleakley et al., 2024; Lanza & Jackson, 2021).

2.3.2. Hosting Providers

Under Article 3(g) of the DSA, hosting providers are defined as entities storing user-generated information (European Commission, 2025a). In essence, they are companies that offer the infrastructure needed to store and serve content on the internet. These providers must fulfill general obligations as intermediaries, such as processing takedown requests and notifying users when their content is restricted or removed. However, the DSA makes exceptions in the case of spammers, where notification may inadvertently help circumvention strategies (Church & Pehlivan, 2023).

Hosting providers are not held liable for illegal content unless they gain actual knowledge of it and fail to act. The DSA reinforces that general monitoring obligations cannot be imposed, maintaining a reactive NTD system (European Commission, 2025a).

2.3.3. Platform Providers

Platform providers, such as social media, search engines, and content-sharing websites, operate as central actors in moderation ecosystems. In comparison to hosting providers, platform providers offer services that include not only hosting but also additional features for content creation, user interaction, and sharing. Unlike hosting providers, they actively manage and curate content.

They implement community guidelines, moderation practices, and algorithms to ensure compliance with regulations. These platforms often rely on a blend of human moderators and automated systems. Their internal frameworks define acceptable content, enforcement mechanisms, and processes for user reporting and appeals (Bleakley et al., 2024; Singhal et al., 2023).

Human moderators may be employees, contractors, or volunteers. For instance, Reddit relies on community volunteers to enforce specific guidelines, while platforms like Twitter and Facebook hire freelance content moderators (Clune & McDaid, 2024; Fasel & Weerts, 2024). Automated detection systems complement these efforts, using technologies such as perceptual hashing and keyword filtering (Farid, 2021). Additionally, partnerships like the GIFCT enable platforms to coordinate against extremist content (Fasel & Weerts, 2024).

Moderation strategies include both hard (removal) and soft (warning, reduced visibility) approaches. During events like the Russian invasion of Ukraine, platforms used labeling and quarantining strategies to control the spread of misinformation (Clune & McDaid, 2024; Fasel & Weerts, 2024).

2.3.4. Law Enforcement Agencies

Law enforcement bodies are required to identify, investigate, and respond to illegal content. Their work often intersects with that of platform moderators and hotlines. As Bleakley et al. (2024) notes, the anonymity of online environments complicates enforcement, requiring proactive engagement.

According to Christensen et al. (2021), six strategies define law enforcement practices: involving communities, assessing information credibility, anticipating crime, evaluating threat levels, forming specialist teams, and building inter-agency collaborations. Studies confirm that combating online child exploitation necessitates transnational and multi-actor cooperation (Baines, 2019; Kloess et al., 2021; Lee et al., 2020).

2.3.5. Hotlines

Hotlines emerged in the late 1990s as public-facing mechanisms for reporting CSAM and other illegal content. These can be operated by government agencies, NGOs, Internet Service Providers (ISPs), or police departments (International Association of Internet Hotline Providers, 2025; Salter & Richardson, 2021). The scale varies significantly: some consist of large analyst teams, while others are staffed by a single individual (Draper, 2022).

The INHOPE network, for instance, coordinates 55 hotlines across 51 countries and focuses on CSAM (International Association of Internet Hotline Providers, 2025). TCO, by contrast, is typically handled through cybersecurity and counterterrorism units of national governments. Verified content is shared with law enforcement and hashed for further detection.

2.3.6. National Competent Authorities

National Competent Authorities (NCAs) enforce EU regulations within their member states. They monitor compliance, issue removal orders, and coordinate with the European Commission. As outlined in the EU regulation on terrorist content, every member state must appoint at least one NCA responsible for implementation (European Commission, 2025b).

For an overview of all NCA's see Table A.1 Appendix A. In the Netherlands, this responsibility lies with the Authority for the prevention of online Terrorist Content and Child Sexual Abuse Material (ATKM). This authority ensures removal orders are executed and maintains oversight over platform practices.

2.4. Moderation Practices

Having outlined the legal definitions of harmful content and the institutional actors involved, this section delves into the practical mechanisms through which online content is identified and evaluated. Moderation practices range from human assessment to automated detection systems. The following subsections explore the different methods.

2.4.1. Human Moderation

Human moderation remains a basic component of content moderation systems. Despite the increasing deployment of automated systems, human moderators continue to play an essential role due to their capacity to interpret context, make nuanced judgments, and assess ambiguous or borderline cases (Gillespie, 2020; Spence et al., 2023).

Human moderators are employed across various sectors, including internal teams at technology companies, specialized moderation firms, nonprofit organizations, and law enforcement agencies. These moderators are tasked with reviewing flagged content, such as text, images, or videos, and determining whether it violates platform guidelines or legal regulations (Barrett, 2020; Spence et al., 2023). The moderation process can be initiated by user reports, algorithmic flagging, or notifications from trusted flaggers (Thakor et al., 2023).

A recurring issue in the literature associated with human moderation is its psychological burden. Prolonged exposure to distressing and graphic content has been shown to cause significant emotional and cognitive strain. Moderators frequently report symptoms of secondary traumatic stress, emotional exhaustion, compassion fatigue, and burnout (Barrett, 2020; Gewirtz-Meydan et al., 2024; Spence et al., 2023). These effects are exacerbated by the repetitive nature of the work (Spence et al., 2023).

Operational challenges also arise from the scale and complexity of moderation tasks. The exponential growth in user-generated content (UGC) has made manual review increasingly unmanageable. Early internet moderation was conducted manually by system operators, but as platforms like YouTube began receiving over 400 hours of video content per minute, manual approaches quickly became infeasible (Langvardt, 2017; Morais Carvalho et al., 2021; Roberts, 2017).

In addition to scalability constraints, consistency in decision-making is a concern. Even among trained professionals, disagreement often arises regarding the classification of harmful content. For instance, Kloess et al. (2019, 2021) found disagreements among experts on whether certain images qualified as indecent CSAM. The authors emphasize the need for regular reliability checks among reviewers to enhance consistency.

Bias is another recurring challenge. Moderators' judgments may be shaped by personal beliefs, cultural perspectives, or institutional norms, which can result in inconsistent enforcement or even discriminatory outcomes. This concern is particularly important in politically or ideologically sensitive areas such as terrorism, where content may be wrongly flagged or suppressed due to differing interpretations (Gorwa et al., 2020; Marsoof et al., 2022).

Given these limitations, several scholars advocate for procedural improvements, including layered verification models. These involve multiple independent reviews of flagged content, which can enhance accuracy, reduce individual bias, and support the legitimacy of moderation decisions (Farid, 2021; Kloess et al., 2021). However, these approaches also introduce additional resource demands and are challenging to scale in environments with high content volumes (Bonagiri et al., 2025; Gillespie, 2020).

2.4.2. Inter-Annotator Agreement

There is still very little research on how annotators (i.e., reviewers or moderators) work together when labeling highly sensitive content, such as CSAM and TCO. This is likely because of ethical concerns and the legal restrictions around these types of data. As a result, most studies on annotation practices come from nearby areas such as hate speech, online abuse, or subjective text classification. These studies can still offer insights into how people label complex or emotionally charged material.

One of the most common challenges in annotation is disagreement between annotators. In many cases, disagreement is treated as a mistake. The common solution is to average out labels or use majority voting to decide what is true. However, studies now show that disagreement can carry valuable information. For example, Sang and Stanton (2022) looked at how people label hate speech and found that personal traits like age and personality affect how people interpret the same content. Their work suggests that instead of ignoring disagreement, it should be studied where it comes from.

This point is supported by Davani et al. (2022), since they argue that forcing all labels into one ground truth can erase the views op people who may see things differently. Their solution is to train models that keep track of each annotator's perspective. This makes it possible to capture patterns in how people disagree, and also gives better estimates of uncertainty. Some of these disagreements are not even mistakes but are linked to the nature of language. Popović (2021) found that people disagree more often when the text is complex or ambiguous. For example, a phrase with unclear meaning or grammar can lead people to mark different errors.

Another factor that affects agreement is the annotators themselves. Yan et al. (2014) shows that not all annotators perform equally well across all cases. Some are more accurate on certain types of content than others. Their method tries to model each annotator's expertise to estimate the true label better. This is especially helpful when no ground truth is available. Similarly, research shows that just because two annotators agree does not mean that they are correct (Jansen et al., 2021). In their study on driver behavior, they found that high agreement sometimes masked a shared mistake.

The way the annotation task is set up also matters. Rimez et al. (2024) introduces a system called Secure and Anonymous Multiparty Annotation System (SAMAS), which lets multiple people annotate healthcare data while staying anonymous. Anonymity can help annotators feel safer and more honest, especially when the content is distressing or controversial. Other work shows that annotators are also influenced by the instructions they are given. Parmar et al. (2022) calls this instruction bias. They found that annotators often copy patterns from the example questions or answers in the instructions, even when they are not meant to. These patterns then show up in the final dataset.

The tools used for annotation also play a role. Research tested different annotation tools for labeling emotion and sentiment (Schmidt et al., 2019). They found that tools with better design helped annotators focus better and be more consistent. When people found the interface hard to use, their performance dropped. For high-risk tasks like CSAM classification, the tools should be easy to use and designed to reduce fatigue and stress.

Finally, the conditions under which annotation is performed, particularly independent decision-making, play a role in reliability. Evidence from other fields supports their importance. In medical imaging, blind independent annotation of retinal scans led to high inter-rater agreement and reduced bias in identifying pathological features (Du et al., 2024). A randomized controlled trial in peer review demonstrated that anonymizing reviewers to each other resulted in more participation and less influence from senior status (Rastogi et al., 2024). Furthermore, according to Larroza et al. (2025), image review protocols showed that structured independence and reviewer anonymity improved honesty and consistency under sensitive conditions.

2.4.3. Multiple Reviewers

The complexities discussed above, such as annotator bias, disagreement, independence, and the influence of tool design, highlight the difficulty of relying on individual human reviewers. In response, institutions and platforms turn to multi-reviewer processes to reduce subjectivity and improve decision reliability (Farid, 2021; Meggyesfalvi, 2024). This section explores the rationale, benefits, and limitations of such processes, particularly in the context of this research subject.

Accuracy is the foremost justification for involving multiple human reviewers. As demonstrated in the field of CSAM moderation, correctly assessing whether a file meets the legal definition of illegal content is a complex task. Kloess et al. (2019, 2021) found variation among expert analysts when evaluating whether images could be classified as indecent CSAM. Their findings highlight the subjectivity involved in such assessments and the necessity for routine reliability testing.

Beyond accuracy, multiple reviewers enhance the accountability and transparency of moderation processes. With growing scrutiny over algorithmic and human moderation alike, stakeholders have expressed concerns about unclear decision-making and the perceived fairness of content removals (Ozanne et al., 2022). Users report uncertainty regarding how moderation decisions are made and have called for clearer mechanisms and more participatory models. Incorporating multiple reviewers can help address this legitimacy gap by introducing checks and balances within the moderation pipeline (Gorwa et al., 2020).

Furthermore, this approach reduces the risk of overreach, particularly in content categories where boundaries are frequently contested, such as hate speech, terrorism, and politically charged material. Research by Parker (2024) and Saleem and Kamande (2024) indicates that moderation systems must avoid suppressing legitimate discourse, including political dissent or critical commentary, which can sometimes be mistaken for extremist or harmful content. Engaging multiple reviewers can ensure a broader perspective is considered before final decisions are made.

Despite these advantages, employing multiple reviewers is not without its challenges. The most pressing issue is scalability. As the volume of flagged content continues to grow, reviewing each item multiple times becomes resource-intensive and potentially unsustainable (Bonagiri et al., 2025; Gillespie, 2020). Moreover, while the involvement of more reviewers might reduce the risk of individual error, it does not fully eliminate subjectivity or bias. The background, training, and institutional setting of each reviewer continue to influence their judgment (Gorwa et al., 2020; Marsoof et al., 2022).

The implementation of triple verification is an example of this approach. As noted by Farid (2021), this method involves three independent human reviewers assessing the same content before its classification, such as the decision to hash it for inclusion in a hash database, is finalized. Although this procedure is already in use by certain institutions, such as NCMEC and IWF (Google, 2025; National Center for Missing & Exploited Children, 2023), this and other approaches remain underexplored in academic literature.

2.4.4. Automated Detection Tools

The increasing volume of harmful online content has necessitated the development and integration of automated detection tools. Based on the literature, three categories of automated detection systems can be identified: hash databases, web crawlers, and deep learning techniques (Lee et al., 2020). While each serves a different purpose, they often operate in combination with human review and verification processes (Gillespie, 2020; Macdonald et al., 2024).

Hash Databases

Hash databases are among the most prominent tools for detecting previously identified illegal content, particularly CSAM (Westlake et al., 2012) and TCO (Macdonald et al., 2024). Hashing involves generating a digital fingerprint, a hash, from an image, video, or audio file. This value is then compared to a database of pre-existing hashes. If a match is found, the uploaded content is flagged for removal or further evaluation (Guerra & Westlake, 2021).

Hash values such as MD5 or SHA-1 are fixed-length alphanumeric representations created using cryptographic algorithms that uniquely identify digital data (Dos Santos et al., 2024). However, standard cryptographic hashes are sensitive to even minor modifications. As a result, altered versions of harmful content, such as resized or cropped images, may evade detection due to hash mismatches (McGarvie, 2023).

Perceptual hashing has been introduced to address this limitation. Unlike cryptographic hashing, perceptual hashing is based on visual characteristics of content, allowing it to detect files that have been modified but are still visually similar to the original (Dos Santos et al., 2024; Farid, 2021; McGarvie, 2023). This is especially relevant for identifying altered content that may otherwise bypass detection systems (Guerra & Westlake, 2021).

International hash databases include INTERPOl's International Child Sexual Exploitation database, the Internet Watch Foundation (IWF) Hash List, and the NCMEC Hash List for CSAM (Internet Watch Foundation, 2025; Interpol, 2025; National Center for Missing & Exploited Children, 2023). For TCO, the Global Internet Forum to Counter Terrorism (GIFCT) maintains a hash database (Global Internet Forum to Counter Terrorism, 2023). Microsoft's PhotoDNA, introduced in 2010, is one widely adopted tool that enables platforms to automatically scan and compare media content against a known database (Lee et al., 2020).

Despite the technical advancements, hash databases remain dependent on human moderation. Most hashes originate from content already identified and verified by analysts (Gillespie, 2020). As such, the effectiveness of hashing relies not only on automation but also on initial human classification and curation (Macdonald et al., 2024).

Web Crawlers

Web crawlers, also referred to as spiders or bots, are automated tools designed to systematically browse and analyze online content. They play a role in detecting illegal material by navigating websites, extracting content, and flagging suspected violations. These tools enable the scanning of digital spaces with minimal human intervention, and they can be enhanced through Natural Language Processing (NLP), sentiment analysis, and classification algorithms (Lee et al., 2020; Westlake et al., 2017).

In the domain of CSAM detection, established implementations include the IWF smart crawler and the Canadian Centre for Child Protection's (C3P) Project Arachnid. These systems combine technologies like PhotoDNA with intelligent scanning rules. For example, the IWF crawler terminates its search after evaluating two non-CSAM URLs before restarting at a new node (Internet Watch Foundation, 2025), while Project Arachnid processes over 100,000 images per month to identify potentially harmful material (Canadian Centre for Child Protection, 2023).

Recent developments have also introduced more web crawlers specifically aimed at countering extremist and terrorist content. As highlighted by Scrivens et al. (2019), custom-built web crawlers are now capable of classifying and interpreting large-scale extremist datasets. Mei and Frank (2015) developed a sentiment-guided web crawler capable of discerning between four types of websites: pro-extremist, anti-extremist, neutral news sources, and irrelevant content. In parallel, Reid et al. (2005)

proposed a semi-automated methodology for capturing and analyzing jihadist websites, addressing the challenges posed by their ephemeral nature.

Deep Learning Techniques
The use of deep learning represents the most latest stage of automated detection. These systems rely on artificial neural networks to learn patterns in large datasets and classify new content accordingly. Two major applications can be distinguished: the first matches new content to existing patterns in databases, and the second trains models to label novel content independently (Udupa et al., 2023).

For CSAM, deep learning is used to detect explicit content stored on user devices (Al-Nabki et al., 2023; Guerra & Westlake, 2021; Pereira et al., 2023; Wolbers et al., 2025). More refined models incorporate multiple classifiers, for instance, detecting pornography, estimating the subject's age, or identifying nudity and skin tone (Anda et al., 2020; de Macedo Neto et al., 2019; Gutfeter et al., 2023; Laranjeira da Silva et al., 2022; Ngo et al., 2024; Oronowicz-Jaśkowiak et al., 2024; Sanchez et al., 2019).

Such tools also support investigative efforts. They help law enforcement categorize material and even link victims to offenders using facial or voice recognition software (Laranjeira da Silva et al., 2022; Ramesh Babu et al., 2024; Rondeau, 2019; Westlake et al., 2022; Woodie, 2016). In the domain of TCO, deep learning and NLP models assist in classifying online texts related to extremist content. These models are trained using manually labeled datasets curated by human analysts (Macdonald et al., 2024).

Despite their scalability and speed, deep learning tools face challenges. They often rely on large datasets for training, which, if inadequately constructed, may introduce biases into the model (Gillespie, 2020; Gorwa et al., 2020; Udupa et al., 2023). Moreover, nuanced language such as satire, slang, or regional dialects can be misinterpreted, leading to false positives or negatives (Parker, 2024; Saleem & Kamande, 2024).

Many platforms promote deep learning and AI as cost-effective and scalable solutions for content moderation (Thakor et al., 2023). However, in practice, most systems still operate based on pattern matching with known examples, rather than true autonomous understanding (Gillespie, 2020).

2.4.5. Types of Workflows

As shown in Table 2.2, Link et al. (2016) identifies three types of content moderation workflows: manual, fully automated, and hybrid. Each distinct involves a trade-off in terms of scalability, accuracy, and flexibility.

Manual workflows involve human moderators evaluating and managing content based on contextual judgment (Link et al., 2016). While highly accurate, they are labor-intensive and difficult to scale. These are often observed in smaller communities or specialized domains, where expert oversight is important (Link et al., 2016).

Fully automated workflows rely entirely on AI systems, such as deep learning and NLP, to detect and act on content without human involvement (Link et al., 2016). These systems offer scalability, but often lack nuance. Platforms like YouTube and Facebook have deployed such systems to remove

Table 2.2: Overview of Content Moderation Workflows (based on Link et al. (2016)

Workflow Type	Description	Advantages	Limitations
Manual	Human experts perform all moderation tasks	High contextual accuracy; handles ambiguity well	Low scalability; prone to overload
Fully Automated	Algorithms execute all moderation steps without human input	High scalability and speed	Limited adaptability; poor with novel cases
Hybrid	Combines machine learning with human oversight	Balances efficiency and reliability; reduces workload	Requires careful system design; dependent on classifier confidence

spam, hate speech, and violent content. For example, Facebook reports that over 95% of hate speech is removed automatically (Galli et al., 2022).

Hybrid workflows combine automated tools with human intervention, aiming to balance scalability with accuracy (Link et al., 2016). Typically, AI filters content and flags uncertain cases for human review. This model is prevalent among major platforms. Facebook employs over 15,000 human moderators to complement algorithmic filters; YouTube uses hybrid workflows to proactively remove content, often before it is viewed (Galli et al., 2022). Reddit also follows a decentralized hybrid model, empowering community moderators supported by AI-assisted tools (He et al., 2024)5.

However, the implementation of hybrid systems in other sectors or platforms, such as hotlines and law enforcement, is less transparent. While automated detection tools are known to assist in identifying illegal content such as CSAM (Laranjeira da Silva et al., 2022; Ramesh Babu et al., 2024; Rondeau, 2019; Westlake et al., 2012; Woodie, 2016), their integration into workflows often requires human verification due to procedural and constitutional safeguards. As such, there is limited publicly available information or peer-reviewed research detailing these workflows in regulated or confidential domains.

2.5. Chapter Summary and Gaps

This chapter has provided an overview of the online content moderation ecosystem, with a focus on CSAM and TCO. The landscape is shaped by a combination of human, technological, and institutional actors, operating under evolving regulatory frameworks and across diverse stakeholder groups.

2.5.1. Context

The field of content moderation has evolved significantly since the early internet era, transitioning from community-based manual review practices (Morais Carvalho et al., 2021; Roberts, 2017) to increasingly automated systems capable of processing vast amounts of data. Despite these technological advances, including perceptual hashing (Farid, 2021), web crawlers (Scrivens et al., 2019; Westlake et al., 2017), and deep learning tools (Laranjeira da Silva et al., 2022; Macdonald et al., 2024), human moderation remains essential. Human reviewers are tasked with interpreting complex and sensitive material, yet often operate under psychological pressure (Barrett, 2020; Gewirtz-Meydan et al., 2024; Spence et al., 2023). Recent research suggests that annotation outcomes are influenced not only by content complexity but also by how review tasks are structured, such as reviewer independence or specific instructional formats (Du et al., 2024; Parmar et al., 2022; Rastogi et al., 2024; Rimez et al., 2024).

Technologies such as hash databases are widely used to identify previously detected material, offering scalability and automation. However, they rely heavily on accurate initial human categorization and fail to detect new or slightly altered content (Gillespie, 2020; Guerra & Westlake, 2021). In practice, major databases such as those maintained by INTERPOL, the Internet Watch Foundation (IWF), and the National Center for Missing and Exploited Children (NCMEC) use hashes derived from previously verified material (Internet Watch Foundation, 2025; National Center for Missing & Exploited Children, 2023).

To mitigate subjectivity and improve accuracy, some organizations have implemented layered verification mechanisms. The most notable of these is triple verification, whereby three independent human analysts must agree before content is hashed and added to a database (Farid, 2021). While adopted in practice, especially for CSAM, the academic literature offers limited empirical evaluation of this procedure.

Furthermore, legal frameworks such as the EU DSA and national regulations mandate transparency, accountability, and trusted flagger systems to support detection and removal processes (Appelman & Leerssen, 2022; European Commission, 2025a). These developments highlight the growing institutional complexity surrounding content moderation, involving actors from civil society, law enforcement, hosting platforms, and regulatory bodies.

2.5.2. Identified Limitations in Existing Research

Taken together, this review has identified several gaps in the academic understanding of multi-step verification processes in online content moderation.

(1) Operationzalization of Verification Processes

While triple verification is acknowledged (Farid, 2021), and implemented as a practice for CSAM (Google, 2025; National Center for Missing & Exploited Children, 2023), there is little academic research into how these verification processes are structured and applied in real-world practice. Much of the current understanding relies on high-level descriptions, with limited insights into whether classification workflows are performed sequentially or in parallel, how disagreement is handled, or how verification integrates with broader organizational procedures. As Gillespie (2020) notes, the content moderation system is often not transparent, making it difficult to analyze how core processes such as classification are structured across different organizations.

While scholars have explored the technological logic behind content detection, such as hash databases or web crawlers (Guerra & Westlake, 2021; McGarvie, 2023; Westlake et al., 2017), the human mechanisms for validating illegal material remain underexamined. Although some annotation systems in machine learning, hate speech, online abuse, and medical contexts describe layered workflows (Du et al., 2024; Marchal et al., 2022; Rastogi et al., 2024; Yan et al., 2014), these insights have not been applied to CSAM or TCO moderation. This study contributes by first investigating how organizations working with CSAM and TCO hash databases structure their verification workflows.

(2) Perceptions on Triple Verification

While triple verification is widely promoted as a strategy to enhance accuracy and reduce subjectivity in classification, it is not without operational and ethical challenges. As identified in both literature and practice, the scalability of such procedures is a growing concern. Reviewing every piece of flagged content three times requires significant human resources, and as the volume of content increases, many institutions question the feasibility of sustaining this model long term (Bonagiri et al., 2025; Gillespie, 2020). In addition, human moderation is known to carry a high psychological burden. Reviewers of CSAM and TCO materials often experience secondary trauma and fatigue, which can be compounded by repeated exposure under verification regimes (Barrett, 2020; Gewirtz-Meydan et al., 2024; Spence et al., 2023).

Despite the institutional push for verification as a quality safeguard, little is known about how frontline professionals actually perceive this process. Do they experience it as supportive and confident, or as an added layer of emotional strain and inefficiency? Existing literature offers limited insight into these perceptions. While Roberts (2017) and Gillespie (2020) have examined the labor conditions of content moderation more broadly, they do not specifically address how human verifiers experience multi-step review processes. This study will explore the gap by examining how professionals experience triple verification, providing insights into both its perceived strengths and limitations.

(3) Empirical Evaluation

Although multi-rater review systems aim to produce more nuanced outcomes, the effects of annotation conditions are rarely assed in the context of CSAM. While annotation conditions are relevant to both CSAM and TCO, this research concentrates on CSAM due to the earlier availability of qualified annotators and access to relevant materials.

This study builds on prior research into the reliability of classifying CSAM, particularly the work of Kloess et al. (2019, 2021), but introduces several methodological distinctions. Unlike Kloess et al. (2019), which involved coders without day-to-day responsibilities for CSAM classification, and Kloess et al. (2021), which conducted a small-scale (N = 100) pilot with five digital forensic raters, this study engages two experienced raters from the Dutch National Police who routinely perform such tasks as part of their official duties with a much larger dataset (N = 2,031). Furthermore, it employs the classification categories currently used in Dutch law enforcement practice (CSAM, Animal Pornography, and Other) rather than the UK's A/B/C offense categorization system. By using a larger dataset and embedding the analysis within a national framework, this study offers increased validity and provides insight into jurisdiction-specific classification dynamics.

Beyond sample size and legal relevance, the current study also addresses a major omission in the prior literature: the impact of annotation conditions. Studies in medical imaging and academic peer review show that blind and anonymous conditions improve inter-rater agreement and reduce cognitive or social bias (Du et al., 2024; Rastogi et al., 2024). Secure annotation systems such as SAMAS have been developed to protect reviewer identity and encourage honest labeling in emotionally sensitive domains (Rimez et al., 2024). Annotation research in NLP has further shown that task framing, example prompts, and interface design can significantly bias labeling behavior (Parmar et al., 2022; Schmidt et al., 2019). However, none of these insights have yet been applied to CSAM classification, despite the high subjectivity and ethical sensitivity of such work.

To fill this gap, the experiment compares classification outcomes under blind and non-blind voting conditions, examining effects on agreement levels. By focusing on experienced professionals, real case material, and jurisdictional labels, this study provides one of the first empirical analyses of how annotation setup influences outcomes in CSAM moderation and contributes to a more evidence-based foundation for future verification systems.

Based on this literature review, the following main research question is formulated:

What are the characteristics of verification processes, including triple verification, for hash databases used in moderating online harmful content?

3 Methodology

The previous chapter identified a lack of research on multi-layered verification processes for hash databases. This chapter will describe the method used in this study.

3.1. Research Design

The research employed a multiphase mixed-methods approach. This approach is used to facilitate a holistic understanding of the research problem, as the strengths of qualitative and quantitative approaches complement each other (Creswell et al., 2011). One advantage of this approach is its capacity to provide complementary insights. Qualitative data can reveal motivations and contextual factors, while quantitative data provides generalizability and measurable relationships (Malina et al., 2011).

The design consisted of three phases (See Figure 3.1). Phase 1 was a qualitative exploration aimed at identifying key themes from interviews and informing the experimental design. Phase 2 was divided into two sub-phases: Phase 2a involved quantitative data collection through an experiment, while Phase 2b consisted of a qualitative follow-up with participants to deepen the understanding of the quantitative results. Phase 3 involved the integration of all data to interpret and synthesize insights across the phases. Each phase was aligned with one or more research questions.

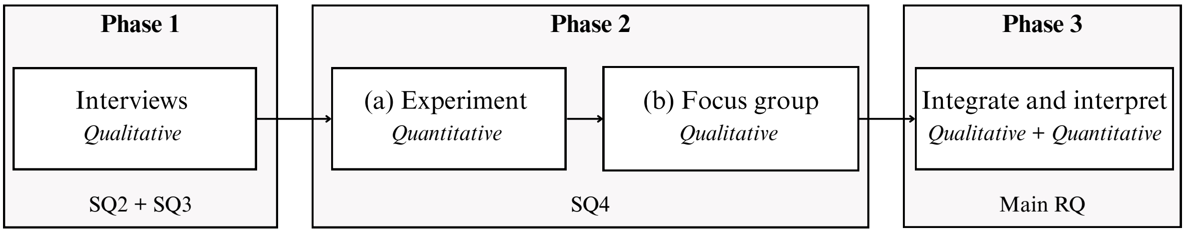

Figure 3.1: Multiphase Mixed-Method Design for this Study

In order to answer the main research question: *What are the characteristics of verification processes, including triple verification, for hash databases used in moderating online harm- ful content?*, sub-questions have been formulated. Each question focuses on specific aspects that contribute to an understanding of these verification processes, and especially triple verification:

- SQ1: Which key concepts, definitions, and principles form the foundation of content moderation and assessment in the context of CSAM and TCO?
- SQ2: What processes do various organizations use to add hashes to databases that contain CSAM or TCO?
- SQ3: How do organizations involved in combating CSAM and/or TCO perceive the triple verification processes of hash databases?
- SQ4: How do different verification approaches affect the final classification of potentially illegal content?

This study employed a combination of qualitative and experimental methods, selected to suit the nature of each sub-question. Semi-structured interviews were chosen to explore organizational workflows and practitioner perspectives in greater depth. This method allowed for flexibility during the conversations, enabling the researcher to ask follow-up questions and clarify complex or unclear points. The combination of interviews and experimental testing ensures both practical insight and empirical grounding.

Sections 3.2 to 3.4 will describe the different research phases in terms of collection and analysis. The first two phases will be explained in detail, as they involve the collection and analysis of data. In contrast, phase 3 will focus on the integration and interpretation of the data (i.e., Chapter 5 Discussion) from phases 1 and 2, without any new data collection or analyses, and will not be described further.

3.2. Phase 1 - Interviews

Data is collected through interviews to understand the processes and perspectives of stakeholders involved in combating CSAM and/or TCO. A semi-structured interview style was adopted, allowing for prepared questions while remaining flexible to stimulate informal discussions (Adams, 2015).

3.2.1. Data Collection

The selection of participants focused on organizations actively working to combat CSAM and/or TCO. Central to the selection process was the inclusion of individuals with operational expertise in the verification and classification of such content. Additionally, special attention was given to those representing organizations that not only perform verification or moderation tasks but also maintain and operate their own hash databases. Participants were purposefully chosen based on their relevant knowledge and experience, employing a method known as purposive sampling (Campbell et al., 2020).

In total, 14 participants were recruited for the research. The study's timeframe limited the total number of interviews conducted. Ten out of fourteen participants were recruited via ATKM's connections, who helped identify organizations with relevant expertise, particularly those knowledgeable about hash databases. These organizations were specifically reached out to as they were expected to provide valuable insights for the study. However, organizations such as NCMEC and Tech Against Terrorism were not included in the sample, as they did not respond to inquiries. The remaining four participants were identified during the internship. To support clarity in the analysis, participants' roles have been grouped into two categories, as shown in Table 3.1.

- **Verification roles** include participants whose responsibility lies in classifying content, but who do not maintain their own hash database.
- **Hash database management roles** refer to participants whose organizations are responsible for content classification and the maintenance of a hash database.

Table 3.1: Overview of Interview Participants

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

It is important to note that these descriptions do not fully capture the range of responsibilities or the complete mission of the organizations involved. However, for this study, participants were grouped in this way to ensure consistent and anonymous presentation of roles and responsibilities while maintaining analytical relevance.

Before the interviews started, participants were verbally informed about the research objectives and the intended use of the interview data, emphasizing their voluntary participation and their right to withdraw from the study at any time without providing a reason. A detailed copy of the verbal consent form is available in Appendix B. Interviews were conducted either in person or online, depending on the participant's preference and scheduling needs. Ten out of fourteen interviews were conducted individually. Due to rescheduling and the organization's choice, one group interview was conducted online with four people at the same time.

Interviews were conducted using a semi-structured format, with specific protocols tailored for law enforcement agencies and other organizations to ensure consistency in questions, as detailed in Appendix C. Overall, the nature of the semi-structured approach allowed for tailored inquiries based on the individual contexts of each organization.

3.2.2. Data Analysis

All data pertaining to the interviews has been securely stored on the researchers' TU Delft OneDrive. With participants' consent, audio recordings of the interviews were obtained using the dictaphone application on the researcher's mobile device. These recordings were transcribed employing the

Whisper model (Spiller et al., 2023). Some interviews were conducted in English, while others were conducted in Dutch. However, all quotes were translated into English.

The audio was re-listened to, and the transcriptions were adjusted to ensure accuracy with the original recordings. Any nonessential texts or discussions unrelated to the interview objectives were removed to enhance data integrity for analysis. The transcribed data underwent thorough reading to ensure familiarity with the content and context.

The interviews lasted 45 to 60 minutes, resulting in lengthy transcripts. The transcriptions were analyzed through coding using ATLAS.ti. Coding is a technique that helps simplify large amounts of text, allowing researchers to understand the data better. According to Chametzky et al. (2016), coding facilitates the conceptualization of data. This analysis facilitated a more efficient process, helped uncover patterns and relationships, and prepared the data to address the first and second sub-questions.

A thematic analysis was conducted. Thematic analysis is useful for examining interview data in qualitative research, as it reveals, assesses, and presents themes in intricate datasets (Joffe, 2011). Inductive coding was used. This refers to the process of generating codes from the data itself and is described as exploratory and data-driven (Jones, 2022).

After generating the initial codes, a review process was undertaken to refine the findings further. This involved evaluating each code to identify redundancies and merging similar codes to enhance clarity and conciseness. The process of reorganization allowed for a more streamlined set of initial codes, which were then analyzed for their relationships and relevance.

Next, these refined codes were grouped into broader main themes that encapsulated related concepts and patterns, providing a comprehensive overview of the data. Following this, sub-themes were articulated from the main themes, ensuring that they accurately captured the nuances present in the interview data. This iterative process of review and refinement was essential in achieving a deeper understanding of the participants' insights and experiences.

The thematic analysis led to 47 codes. Main themes and sub-themes have been identified based on the codes to provide an overview of recurring themes. The main themes are benefits of triple verification, challenges of triple verification, necessity of triple verification, future perspectives and opportunities, and comparison to TCO. The codebook, which contains the main themes along with their sub-themes, codes, and example quotes, is presented in Table D.6 of Appendix D.

In the results section, the reporting and visualization of the qualitative findings are structured to provide clarity and coherence. Participants are referred to from P1 to P14. No additional information will be provided due to the need for anonymization and to maintain the confidentiality of organizational workflows. The results are divided into two main parts, both presented in a narrative format.

To answer the second sub-question, the findings include a comparative table outlining various organizations' verification processes. This table serves as a reference point, summarizing key differences and similarities among the approaches. Additionally, these processes are narratively discussed in a randomized order to illustrate the diversity of practices.

To answer the third sub-question, the findings focus on the themes derived from the qualitative data, wherein each theme is presented alongside its corresponding sub-themes. The organization of the sub-themes is intentional, as it reflects the frequency of discussion during the interviews; topics that emerged most often appear first, offering insights into what participants deemed most significant. Tables D.1 to D.5 in Appendix D display the frequency of topics among participants for each main theme.

3.3. Phase 2a - Experiment

The experimental design of this study was shaped by a combination of insights from the literature and findings from the qualitative phase. Across academic domains, annotation conditions, such as the number of reviewers and whether they work independently or have access to prior decisions, have been shown to affect classification outcomes.

However, these findings have not yet been extended to CSAM moderation, despite its high sensitivity and reliance on human classification. At the same time, qualitative interviews conducted during Phase 1 of this study revealed significant variation in how organizations implement verification processes for CSAM and TCO (see Table 4.1 in Chapter 4). Some organizations mandate triple-blind voting, while others allow for non-blind sequential review, or apply no fixed process at all. The number of reviewers involved also varies widely, raising questions about how consistency and accuracy are achieved.

Taken together, the literature and interview data point to two key variables that deserve empirical evaluation:

- **Anonymity Conditions:** Examining the difference between blind and non-blind voting processes.
- **Verification Levels:** Comparing the times when there is a disagreement between the two or three votes.

3.3.1. Data Collection

The two experimental conditions implemented in this study reflect verification procedures that emerged from the qualitative phase. Interviews revealed that organizations working with CSAM and/or TCO employ different verification setups in practice. Some described workflows in which reviewers assess content independently, without knowledge of prior votes, which are referred to as blind verification. Others reported sequential processes where reviewers have access to earlier classifications before casting their own, which is referred to in this study as non-blind verification.

Multiple organizations mentioned these two approaches during the interviews, indicating that both are actively used in the field. The experiment compares these two conditions to explore the possible effects of these verification structures. In the blind setting, reviewers worked independently and were unaware of each other's input. In the non-blind setting, voting occurred sequentially, and the final reviewer had access to the earlier vote(s).

The classification categories used in the experiment (CSAM, Animal Pornography, and Other) were not devised by the researcher, but reflect the categories used in Dutch Law enforcement practice. These categories correspond to Article 252 (CSAM) and Article 254 (Animal Pornography) of the Dutch Criminal Code [1], which define the legal boundaries for criminal content in this domain. The *Other* category is used operationally when material is deemed not to fall under either of these offense classifications.

In this experiment, ground truth was defined as the outcome on which all three reviewers agreed on one category in the workflow. This reflects how content is finalized for inclusion in the Dutch National Police's hash database: when three professionals assign the same classification, it is considered sufficiently verified for inclusion. However, this definition should be seen as procedural rather than absolute.

The department selected two raters from the Dutch National Police, both with similar levels of experience, to participate in the experiment. The police selection ensures that the raters' expertise is representative of typical investigators in the field. Classifying potential CSAM content is part of their daily work.

The data used for this experiment comprises 2,031 images and videos formed by the Dutch National Police. The number of items is chosen based on the limitations and future directions of Kloess et al. (2021), who suggested using a larger dataset than their research (N = 100). The dataset by the Dutch National Police is created to represent a diverse range of content from different age ranges and levels of severity, also suggested by the research of Kloess et al. (2021). The selection is designed to keep series content very low, to ensure variability and independence in the dataset's content. Fully omitting series content would not reflect real-life datasets and was therefore not advised by the Dutch National Police.

The content in the dataset is sourced from the Police's hash database. The content has already received one vote. This is for three reasons. First, in various interviews, the difference between double and triple voting has emerged as a theme. This led to the question of how often three individuals disagree. Second, it prevents the need for a third analyst, which saves time and resources. Third, data only enters the Dutch National Police hash database if it has already received one vote from a connected system that investigators use in their day-to-day investigations.

The created dataset is split into two subsets since the experiment will be conducted in two phases. Each phase corresponds to one of the following scenarios:

Phase 1 - Blind Conditions: Each rater independently assessed a set of 1,000 items using their standard local environment and tools. Every rater was assigned one unique case consisting of 1,000 items. Voting was conducted under blind conditions: raters were aware that one prior vote had been given but were not informed whether they were acting as the second or third verifier. Raters classified each item according to standard organizational categories: CSAM, Animal Pornography, or Other. No visibility into other raters' votes was available, mirroring their typical working environment.

[1] https://wetten.overheid.nl/BWBR0001854/2024-07-01/0BoekTweede_TiteldeelXIV_Artikel252

Phase 2 - Non-Blind Conditions: For the non-blind voting condition, a new dataset was prepared consisting of four sets, each containing 250 items. The evaluation procedure was organized as follows:

- Rater A assessed the first 250-item case as the second voter and concluded their review after this case.
- Rater B then performed two tasks:
  - Reviewed the same 250-item case as the third voter, with access to the prior vote.
  - Reviewed a second, distinct 250-item case as the second voter, without knowledge of future votes.
- Rater A then returned to assess the second case reviewed by Rater B, now acting as the third voter.

This cycle was repeated to complete all sets (see Figure 3.2), ensuring that each rater served equally in the roles of second and third voter. It was not possible to do this randomly due to the system used by the police and the timeframe of this study.

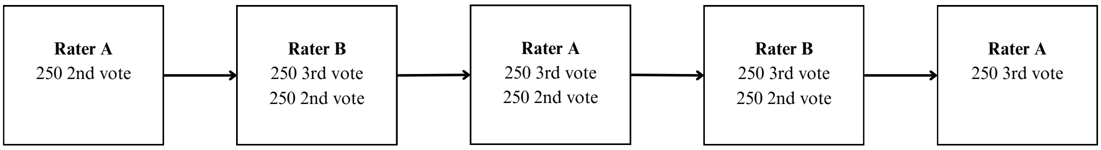

Figure 3.2: Voting sequence in the non-blind condition

The raters were instructed to open a designated tab within their system that provided visibility into prior votes. This normally restricted functionality was explicitly enabled for this experiment to facilitate the non-blind condition. Raters coded the materials again using the same categories as in the blind phase.

From both conditions, a separate output file was created. For the first condition (See Table 3.2 for example blind condition), this file contained the following data: Rater A (CSAM, Animal Pornography, Other), Rater B (CSAM, Animal Pornography, Other), File extension, FileSize (bytes), Height (Pixel), Width (Pixel), *Only for video:* Length (Seconds), Mimetype (Image or Video), Visualgroup (Grouped by perceptual hashing), Number (1-2000 replaced for hash numbers).

Table 3.2: Example of output file structure for the blind condition

Number	Rater A	Rater B	Mismatch	Extension	Filesize (B)	Height	Width	Length (s)	Mimetype	Visualgroup
001	CSAM	CSAM	False	.jpg	152487	720	1280	–	Image	VG103
002	Other	Animal	True	.mp4	3290471	1080	1920	12	Video	VG224
003	Animal	Animal	False	.jpg	89045	600	800	–	Image	VG145
...	...	...	...	...	...	...	...	...	...	...

The second condition (See Table 3.3 for example non-blind condition) contained the same metadata, but also the Set (1,2,3,4) to identify which Rater was which vote, and Master Category (CSAM, Animal

Pornography, Other, or Not classified), which provides information whether the three votes given in total were similar or not. This data was not provided in the first condition because the classification process was done in a different environment.

Table 3.3: Example of output file structure for the non-blind condition

Number	Set	Vote A	Vote B	Mismatch	Master Cat.	Extension	Filesize	Height	Width	Length	Mimetype	Visualgroup
001	1	CSAM	CSAM	False	CSAM	.jpg	142000	720	1280	–	Image	VG103
002	2	Animal	Animal	False	Not classified	.mp4	2900000	1080	1920	14	Video	VG224
003	3	CSAM	Other	True	Not classified	.jpg	134500	800	1280	–	Image	VG145
...	...	...	...	...	...	...	...	...	...	...	...	...

All assessments were conducted within the secure office environment of the Dutch National Police precincts, using the raters' standard software and systems. This ensured ecological validity by maintaining familiarity with operational conditions. The assessment took place over a two-week timeframe in May 2025, not to pressure the raters but to provide them with time besides their normal workload.

3.3.2. Data Analysis

All analyses were conducted using IBM SPSS Statistics. SPSS was selected due to its user-friendly interface and built-in support for categorical data analysis (Jain & Sengar, 2024).

Inter-annotator agreement was evaluated using Cohen's Kappa statistic to assess the annotation process's reliability. Unlike simple percent agreement, Kappa provides a more robust estimate of agreement. Kappa values range from -1 to 1. Cohen's Kappa (κ) values were interpreted following the widely used benchmark by Landis and Koch (1977), which classifies agreement as listed in Table 3.4.

Table 3.4: Interpretation of Cohen's Kappa Values (Landis & Koch, 1977)

Kappa Range	Interpretation
< 0.00	Poor agreement
0.00–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

Crosstabulations of Rater A and Rater B's annotations were generated, and Kappa values were computed both for the whole dataset and for subsets of the data split by file type (image vs video). This allowed for a more granular comparison of agreement across different media formats. In total, analyses were conducted separately for both conditions and across file types within each condition.

In the non-blind condition, it was possible to investigate whether the order in which raters voted influenced their level of agreement. For this purpose, two variables were created, indicating whether each rater voted second or third. Because the order of voting was systematically alternated across four dataset subsets, it was possible to compare outcomes based on voting position. A Chi-square test of independence was used to assess whether agreement rates differed significantly based on vote order. Results for Rater B mirrored those of Rater A and were therefore not reported separately.

The non-blind condition also enabled an analysis of triple verification, as the system could track when all three raters independently chose the same category. This was recorded in the Mastercategory variable. To build on this, an additional variable was constructed:

- **Full agreement:** Raters A and B assigned the same label, and the third rater agreed as well.
- **A & B agreed, third disagreed:** Rater A and B agreed, indicating that the third rater voted differently, and no label received triple votes.
- **Disagreement, unresolved:** Rater A and B disagreed with each other, and no consensus was reached (Mastercategory = 'Not classified').

Table 3.3 shows an example of how these variables were computed. For instance, in row 001, all three raters selected *CSAM*, resulting in *Mastercategory = CSAM*, representing full agreement. In row 002, A and B agreed on *Animal Pornography*, but the vote which was already given assigned a different category, since *Mastercategory = Not Classified*. In row 003, A and B disagreed, and no consensus followed, resulting in an unresolved disagreement. This classification was only possible in the non-blind setting, where voting was sequential and the platform tracked when consensus was achieved. Frequencies were calculated to determine how often each type of agreement occurred.

3.4. Phase 2b - Focus Group

After analyzing the data from Phase 2a, a post-assessment focus group was conducted with both raters. They were asked to explain their classification choices for disagreements between votes, which led to some discussion. These insights supplement the quantitative data and are used to contextualize patterns of consensus and disagreement.

During the two-hour rater discussion session, a total of 49 items were reviewed. The session was conducted in a hybrid format, with one rater present in the office and the other joining remotely. This setup, alongside the time required to retrieve and assess the files, constrained the overall number of items that could be discussed. The reviewed sample encompassed items from both phases of the dataset. In terms of content type, the raters reviewed a total of 45 images and 4 videos.

The researcher wrote their responses down while recording and prompting them to explain the rationale behind their voting choices. The recordings were relistened to, and the findings were reviewed and discussed narratively. The findings include the main insights of these discussions.

3.5. Research Ethics

The utmost care was taken to ensure that all data collection activities were conducted in an ethical, responsible, and respectful manner. Before doing the interviews and experimental phase of the research, the study underwent an ethical review process managed by the Human Research Ethics Committee (HREC) of TU Delft. This involved submitting an HREC form, which incorporated a risk assessment to ensure the research design adhered to ethical standards. Additionally, a Data Management Plan (DMP) was developed. This outlined the storage and handling of all research data. A consent form was also prepared for the interviews to inform participants about the study's nature, purpose, and their rights. Initially, the HREC requested revisions. Specifically, it required the inclusion of signed agreements and the improvement of the DMP supported by a data steward. After these adjustments, the revised application received approval. This allowed the research to proceed with ethical clearance.

4

Findings

This chapter presents the findings from the interviews and the experiment. This chapter is structured as follows. Section 4.1 discusses the different verification processes per organization and answers SQ2. Section 4.2 discusses the thematic analyses and answers SQ3. Section 4.3 and 4.4 present the quantitative and qualitative results of the experiment conducted with the Dutch National Police and answer SQ4.

Before diving into the findings, it is important to clarify a term that will be frequently used throughout this chapter: baseline. In the context of this research, 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 (INHOPE, 2025; Interpol, 2025). This content is considered illegal in all 196 member countries of INTERPOL (Interpol, 2025).

4.1. Results Phase 1: Interviews - Organizational Workflows

This section will provide an answer to the following sub-question: *SQ2: What processes do various organizations use to add hashes to databases that contain CSAM or TCO?*

Table 4.1 provides an overview of the different workflows across different organizations. Organizations are labelled as A to E to anonymize the interviewed organizations. It outlines key factors such as the type of verification mandated, the number of reviewers involved, classification criteria for content, voting types, and the size of the databases each organization manages. For the number of reviewers:

- Small: fewer than 20 reviewers
- Medium: between 20 and 50 reviewers.
- High: more than 50 reviewers.
- Unknown: Information was unavailable or not provided.

Given the variation in database sizes among the organizations studied, ranging from under 500,000 to over 20 million hashes, categorical descriptors ('small,' 'medium,' 'large,' and 'very large') were adopted to illustrate differences while maintaining confidentiality of exact operational data.

- Small: fewer than 500,000 hashes.
- Medium: between 500,000 and 2,000,000 hashes.
- Large: between 2,000,000 and 10,000,000 hashes.
- Very large: more than 10,000,000 hashes.

Table 4.1: Overview of processes across different organizations

	Org. A	**Org. B**	**Org. C**	**Org. D**	**Org. E**
Verification Process	Triple verification mandated	No prescribed verification process	Triple verification mandated	Variable, based on context (single, double, triple, or more)	Switched from triple to double
Number of Reviewers	High	Unknown	Small	Medium	High
Classification Criteria	3 categories, including child pornography, animal pornography, and 'or else'	Transparent taxonomy based on human rights and specific criteria for membership	Severe content only, real images of children under 13	3 categories with extensive metadata, 21 tags including age, gender, specific acts, etc.	CSAM and harmful to children, multiple categories within both buckets
Voting Type	Blind	Unknown	Non-blind	Blind/Non-blind	Non-blind
Database Size	Large	Large	Small	Large	Very Large

Sections 4.1.1 to 4.1.5 will delve deeper into each organization's practices. Since some organizations are, for example, law enforcement and some are hotlines, a person who classifies content is called *reviewer* and not *employee* or *officer*, to ensure anonymity.

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 position in the voting sequence, though they are not influenced by prior votes.

Images that receive three identical classifications 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.

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 inject 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 extremist or otherwise harmful material. The taxonomy ensures consistency, even in the absence of a uniform verification protocol.

While the number of reviewers involved and the voting structure (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. The voting process is non-blind, as 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

Organization D 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 workstream. 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. Instead, decisions about inclusion into the database are made once the threshold is met, regardless of sequence.

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 moderator. 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 moderator exposure while maintaining classification reliability and ethical standards.

Key Takeaways from the Organizational Comparison

- **Verification models vary widely, but triple voting remains a strong norm.** Three of five organizations used or formerly used triple verification, though some have shifted toward more flexible or efficiency-driven alternatives.
- **Voting context (blind vs. non-blind) differs by system design.** Most organizations use non-blind or mixed systems, where prior votes are visible or inferred, indicating a balance between independent assessment and workflow optimization.
- **Classification frameworks range from minimal to highly granular.** While some use simple three-label systems, others apply over 20 metadata tags, reflecting divergent goals in database use.

4.2. Results Phase 1: Interviews - Thematic Analysis

This section will provide the results of the thematic analysis of the interviews. In total, five main themes have emerged from the coding: benefits, challenges, necessity, future perspectives and opportunities, and the comparison to TCO. Each main theme contains approximately 3-5 sub-themes. This part will also answer the following sub-question: *SQ3: How do organizations involved in combating CSAM and/or TCO perceive the triple verification processes of hash databases?*

Sections 4.2.1 to 4.2.5 will discuss these themes. Section 4.2.6 will provide additional points that participants wanted to emphasize when closing the interviews.

4.2.1. Main Theme 1: Benefits of Triple Verification

This section delves into the participants' perspectives on the benefits of implementing a triple verification process in the assessment of online harmful content, focusing on five key sub-themes: legal considerations, content assessment, inconsistent assessment, quality, and operations.

Legal Considerations. The legal implications of CSAM are discussed by all participants. CSAM content necessitates stringent verification processes to comply with diverse international laws. *P2, P5, P6, P12, P13* support this by noting that one of the main reasons for implementing multiple review verification processes is to determine whether an image fulfills the legal criteria to be classified as criminal. The participants, who came from different organizations in different countries, highlighted that baseline material is illegal all around the world and has the same classification criteria, but national legislation differs per country.

However, baseline is a classification for CSAM set by Interpol, which is considered illegal in any country. The difficulty lies in the national variation, as discussed by *P2*, for example, is that in the Netherlands, reviewers need to check whether it is CSAM, as referred to in Article 252. This view is reinforced by *P5, P6, P8, P9, P10, P11.* As voiced by *P10 '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. Another benefit triple verification adds to is the difficulties that arise in content assessment. Ten out of fourteen participants noted that there are certain challenges in assessing CSAM content; triple verification can confirm or deny whether three people agree that a video or image is CSAM in difficult cases. Specifically regarding the identification or estimation of age, sexual act, and/or sexual pose.

P12 said *'If I have like a picture and I show that picture to my colleagues and say: can you estimate the age of this child? The result might be different. What I say is that it is just what you see in the picture. Verification is much, it's so hard...'.* As emphasized by *P7, P8, P9, P10, P13, P14*, context can be sensitive, making it difficult to interpret accurately. As said by *P14*, when evaluating content, you should consider various factors, including *'cultural, situational, and emotional elements'.*

Inconsistent assessment. Inconsistent assessment was mentioned in nine interviews. *P2, P4, P5, P6, P8, P9, P10, P13, P14* highlighted that even with good training and education, people still make mistakes. In the interviews, *P5, P13, P14* emphasized the challenges of achieving consistency among team members, even when all have undergone similar training and education. They noted that, despite a *'solid foundation of knowledge'*, as stated by *P13*, individual perspectives, experiences, and interpretations can lead to discrepancies in how content is classified. Triple verification serves as a safeguard against such mistakes.

In addition to the previously mentioned factors, other elements also influence the assessment process. One particular statement from *P6* emphasized, *'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 3 pairs of eyes to take out those rotten ones'*. This view was reinforced by *P5*, who mentioned that sometimes a person clicks page down and is distracted, and they sometimes give the wrong assessment unknowingly.

Quality. *P1, P2, P4, P5, P6, P9, P10* stated that when content is checked and assessed by three independent human layers, it adds to the accuracy and quality of hash databases. A perspective shared by *P1* and *P4*, they emphasized that multiple assessments minimize false positives and uphold the integrity of hash databases. *P2* highlighted, *'I think just for the purity of the database. The moment three people look at it, yes, that's just then the purest. The chances of there still being a false positive in it then, yes, that is of course minimal'*.

Operations. According to four participants, assessing illegal content does not take excessive time. *P5* stated that assessing illegal content is not an *'exact science'*. Moreover, as mentioned by four other participants, training and education are important. This ongoing education process can further influence the overall efficiency and consistency, as reviewers strive to make informed and capable assessments.

Ultimately, while the initial assessment may not be time-consuming, the intricacies involved in making informed, legally compliant decisions can add layers of complexity to the operational process. *P13*, who works for an organization where reviewers from other hotlines worldwide help assess content and add hashes, commented *'It's actually a lot of work. So reviewers in other countries will have had other training, right? So, like some of the base training related to sexual maturation rates is pretty universal. But, you know, the law within your own country, you might start to learn with that. That is the reason why we have extensive trainings with them to start to explain the different categories'*.

4.2.2. Main Theme 2: Challenges of Triple Verification

This section explores the participants' perspectives on the challenges associated with the triple verification process in the assessment of online harmful content, focusing on five sub-themes: human impact, content volume, resource and costs, technological limitations, and accuracy.

Human Impact. The human aspect of assessing CSAM is a recurring challenge of triple verification and was mentioned by all participants. All participants agreed that reviewing such disturbing content can take an emotional toll. The requirement for three separate verifications can lead to unnecessary burdens in clear-cut cases, while not providing enough reviewers to manage the volume effectively. *P1* emphasizes the logistical impossibility of matching human resources against the volume of content, while *P2* questions the practicality and frequency of disagreements among the three reviewers.

In the discussions of human impact regarding verification processes, *P14* highlighted this as a reason for their organization's transition from triple to double verification. *P13* observed that for cases involving individuals under 14 years old, they decided that three rounds of verification might be excessive. Instead, they argued that another approach, using just one or two reviews, can often suffice.

Content Volume. A challenge emphasized by thirteen participants is the overwhelming volume of content requiring (triple) verification. This increasing burden not only strains resources but also complicates the management of verification processes. As *P6* metaphorically describes, it is like '*rolling the stone up the hill*', highlighting the continuous and growing nature of this challenge. Similarly, P10 notes that the verification processes are '*overloaded*'.

However, the content volume does differs across organizations due to variations in their sources of origin. For instance, from a law enforcement or hotline perspective, the approach to handling cases varies—some organizations proactively seek out cases for investigation, while others focus on those that are reported. For *P5* and *P6*, a single case can contain hundreds of thousands of images and videos that require triple verification. In contrast, *P13* faces challenges with numerous sources leading to an inundation of content. Additionally, *P14*, which utilizes a web crawler, reported that the volume became too much, prompting a shift from triple to double verification to manage the influx more effectively.

Resource and Cost. The resource-intensive nature of triple verification is emphasized by *P5, P6, P7, P8, P9, P10, P11, P12, P13, P14*. Acknowledged by *P5, P8. P9, P10, P11*, triple verifying is expensive. In their view, the benefits outweigh the costs and resources required, viewing the investment as worthwhile for the victims helped with the database. *P7* points out the lack of human resources available for such intensive processes, indicating that only the largest companies can afford to allocate sufficient personnel toward these efforts.

P6 commented that *'Because you have to imagine that in our database, some or most have only been seen once, so you have to do two or more anyway. So, in total, it's not 2.5 million images and videos, but 5 million to get from 1 to 3 views and classifications. So, in ideal situations then you're working 50 - 100 days anyway with 6 people. That is 300 - 600 working days of constant classification, which is impossible to ask of staff. That's pretty expensive, tedious and exhausting and you have to do something about it'*. In addition, *P14* explained that their organization switched from triple to double verification because *'it is a massive waste of time and reviewer'*.

Technological Limitations. According to *P1, P5, P6, P12, P13, P14* the subtleties of digital content pose technological challenges, especially concerning the creation of hash codes for images. Minor alterations in image data can lead to new hash codes, complicating the identification of duplicates or nearly identical content. *P1* provides insights into this issue, explaining how slight differences, such as a single pixel change or a Snapchat filter, can result in entirely new hashes, exacerbating the challenge of an already overfull backlog. On its own, this is not a problem.

Out of these six, *P5, P6, P12, P14* point out that the actual growth of the database remains minimal due to the process of double-checking for duplicates or nearly identical content. *P5* said *'Because of* *the three eye principle and the numerous duplicates, it hinders the growth of the database'.*

Accuracy. Lastly, something that recurred as a sub-theme as a benefit also came forward as a challenge. Even though something is checked three times, that does not mean you achieve absolute accuracy. *P6* reflects on the complexity and inherent subjectivity of striving for 100% accuracy, which, while ideal, is rarely attainable in practice, also based on the fact that different countries have different legislation and interpretation differences will arise.

Additionally, *P1 and P2* mentioned that having three people assess the content does not guarantee its correctness. They echoed that even with multiple checks, *'mistakes can still occur'.* As discussed by *P5*, who wants the database to be *'as good as possible'*, stated that even though three individuals have checked the content, there is no guarantee that *'no mistakes are made'.*

4.2.3. Main theme 3: Necessity of Triple Verification

This section examines the participants' perspectives on the necessity of triple verification processes. It addresses four key sub-themes: process, utility and impact, human factors, and ethical considerations.

Process. Eight out of fourteen expressed a clear understanding and support for the triple verification process. Especially in scenarios where the stakes involve human and legal consequences. According to *P3*, the triple verification process could be important for maintaining high standards of accuracy in hash databases, which are pivotal in legal contexts where the stakes include human rights and judicial integrity. *P8* stated that *'If it so invasive that you are going to use censorship or intervene on human lives, I can imagine triple verifying is a good idea'.*

Conversely, six participants advocated for a more eyes principle, suggesting a flexible approach to the number of verifications based on the case's complexity and the content's clarity. *P2* was one supporter of the more eyes principle, but questions how many eyes are *'too much'.* The discussion on necessity also brought up how different types of content might require different levels of assessment. While baseline might not necessitate three reviews, more ambiguous cases clearly benefit from multiple evaluations. As noted by *P13*, there is no requirement for three checks on an image when it is evident that *'that is clearly a child'.* However, if content requires additional checks beyond three, the company associated with *P13* will conduct those as needed.

Utility and Impact. The need for triple verification stems from the potential consequences that inaccuracies in hash databases can cause, particularly regarding the organization's objectives. Many participants acknowledged the serious implications that such errors could have, especially since these databases are utilized by different parties to make legal decisions. *P6* illustrates a scenario that stresses the risks of depending on these databases, which might result in convictions based on matches that are deemed *'not good enough, but that we also have to be realistic'.* As said by *P6 'It is highly unlikely that an offender is sentenced based on just one image that will be in a series or set, and may be questionable. The proof in court will always be based on solid classifications and images or videos that pose no questions as to their illegal nature and have to be described in the statement of the sworn officer'.* Additionally, *P5, P10, P11, P12, P13, P14* emphasized the importance of distinguishing the purpose of hash databases, whether they are intended for law enforcement or hotlines.

P5 further emphasizes the preference for *'a smaller, more accurate database over a larger, less reliable one'*, pointing out the repercussions of database errors that lead to disuse and distrust. In contrast, *P14* commented that *'What are we actually doing this (triple verification) for? For this, this tiny fraction of maybe we'll catch some errors? If they see something that doesn't look right, they will get back to us, and we can correct the error.'*

Human Factors. A sub-theme that emerged as a necessity is human factors. Reviewers bring their own experiences and perspectives to the table, which can influence how much they have seen and how they assess various situations. For instance, a reviewer with extensive field experience may interpret data differently than someone who primarily relies on less knowledge. This variance in experience can shape their understanding of nuances in information. This aspect was discussed by *P3, P5, P13*, highlighting the importance of recognizing this and necessitating multiple reviewers.

The influence of personal biases and perspectives in assessing content was mentioned during five interviews. A comment by *P5* highlighted the importance of triple verification to mitigate these subjective interpretations *'You have colleagues, of course, who just started. I wouldn't be skeptical about that then, not at all, but then I would pay a little more attention to that. Because then, where does your assessment come from? Is it education or background? For example, there are people who say; I don't take pictures of my children in the bathtub. Because there are also people who do. And then you do have to see if there is an erotic character in it of that child in the bath. If he's in the bath with his pacifier, I don't think that's erotic. No. But others do, and so that's where the difficulty lies. And if you don't triple verify that, it goes into the database'.*

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. Main Theme 4: Future Perspectives and Opportunities

This section examines the perspectives of the participants on the future perspectives and opportunities within the verification processes related to CSAM, focusing on four sub-themes: AI and ML, oversight, automation, and technological diversification.

AI and ML. Twelve participants brought up the potential role of AI and/or ML in enhancing the verification processes. *P1, P2, P3, P4, P6* see an opportunity for these technologies to assist in the first or second vote. *P4* sees potential in AI and ML technologies to support the initial stages of content verification, thereby reducing the workload on human reviewers and enhancing the speed and efficiency of the process.

P6 suggested a layered approach where AI could handle the first or second assessment, potentially reducing the workload for human reviewers and improving the speed and scalability of the process. *P9, P10, P11, P12, P13, P14* disagreed and see a role for AI/ML only in prioritizing content for assessment, but not automatically classifying or voting.

Oversight. Ten participants were motivated by the necessity of maintaining human oversight, especially due to the sensitive nature of the content and the implications of errors. Participants expressed concern about AI systems making autonomous decisions without human intervention. *P6* commented *'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'.*

P12 reinforces this viewpoint, arguing that while AI can assist, humans must ultimately control and oversee the process and ensure the ethical handling of cases. Lastly, double verification, or the multiple-eye principle, has been discussed in all interviews. Seven participants see a future for double verification, and seven participants support triple verification.

Automation. Besides AI and ML learning for the voting of verification processes, ten participants commented that automation in the process could and maybe should happen. This automation should especially happen, according to six participants, in the filtering and prioritizing of the content before the verification. Reflecting on the automation, *P12* expressed skepticism about completely replacing human oversight with automated systems, despite acknowledging the potential of these technologies to enhance the verification process. *P2's* perspective suggests a cautious approach to automation, advocating for a hybrid model where technology supports, but does not replace, human expertise.

Technological Diversification. Five participants highlighted the matter of technological diversification in the verification process. Among them, four participants specifically advocated for the use of perceptual hashing as a method. *P13* noted the strategic advantage of collaborating with companies that train their classifiers, explaining that by running their software against a dataset of tagged imagery, they could automatically identify specific content, such as images of children within a certain age range.

4.2.5. Main Theme 5: Comparison to TCO

This section examines the participants' perspectives on the verification processes related to CSAM in comparison to TCO. It focuses on three key sub-themes: classification criteria, verification, and knowledge gap.

Classification Criteria. The matter of why there is a difference in verification processes of adding hashes to TCO or CSAM databases is perceived differently among the participants. Some reasoned that TCO is easier to assess and verify than CSAM, while others argued it may be even more difficult. *P8* explained that terrorist content is easier to flag than CSAM, because there are *'more nuances in the CSAM world'.*

P4, on the other hand, argued *'TCO is often multi-interpretable. For example, you notice that the concept of glorification of terrorist crimes leads to a lot of discussions. But you have, say, five steps there after that to get there for glorification, say there. And the justices, for example, did have material. Anders Breivik committed those attacks in Norway. Well, there are people who adore him. Who, for example, on his birthday, just posted online messages of happy birthday Anders Breivik. Yes, is that then the glorification* *of terrorist crime? That you wish a pre-divided assassin a happy birthday? Yes, I imagine that is also a bit subjective for per person difference'.*

Verification. The verification of TCO involves several variables that influence the decision on its legality. Unlike CSAM, where any depiction is illegal, TCO may not directly lead to legal consequences unless specific criteria are met. The process of adding TCO hashes to databases typically involves fewer stricter guidelines. This distinction, mentioned by *P3, P4, P7* highlights the nature of legal frameworks surrounding these offenses.

Moreover, the handling of TCO is marked by *P9*, as a process of adding hashes to databases typically involves *'fewer stricter guidelines'*. The implications of this difference are significant. While CSAM is treated as *'an absolute violation'*, as stated by *P4*, the participant acknowledges that TCO cases require a more discerning approach.

Knowledge Gap. The understanding of seven participants regarding the verification processes for TCO revealed a knowledge gap. Four participants indicated that they did not comprehend how these processes work, expressing confusion about the criteria and factors involved in assessing TCO. They noted a lack of familiarity with the verification methods, which resulted in uncertainty when discussing the subject. In contrast, three participants acknowledged they had some awareness but felt uncertain regarding the specifics of TCO verification processes. They recognized the existence of these processes but were unsure how to navigate the complexities inherent in them.

4.2.6. Consensus on Interview Coverage

This section discusses the insights that participants specifically requested to highlight as the interviews concluded. These insights did not fit into the main themes previously identified, but are nonetheless important to provide. *P1, P2, P3, P4, P6, P8, P9, P11* said that everything was discussed and had nothing further to mention.

P5 emphasized the significance of the database's purpose, which dictates its use and operational focus. For *P5*, the purity of the database is important as it can significantly impact every victim. The primary concern for their organization is to ensure that their database serves its intended purpose effectively and ethically.

P7 discussed the challenges of monitoring companies in the trust and safety sector. *P7* noted that the focus often remains on a few large companies presumed to have substantial resources. However, they pointed out the necessity of considering smaller and medium-sized companies that might lack extensive human and technical resources. *P7* stressed the importance of adapting trust and safety tactics to suit a diverse array of platforms, emphasizing that content hashing is just one aspect of a multi-level strategy that includes network and behavioral analysis to identify harmful actors and content.

P10 advocated strongly for triple verification processes and expressed a desire to eventually eliminate the need to view harmful material altogether, highlighting the significant psychological impact handling such content can have on reviewers.

P12 highlighted the ease and effectiveness of the triple verification system despite its seemingly complex setup. *P12* views triple verification as essential, connecting directly to their operational criteria and improving the accuracy and reliability of their database.

P13 reflected on the evolution of verification processes, suggesting that a shift from a three-vote to a two-vote system could be beneficial given the current prevalence of CSAM. *P13* believes such a move could maintain or even enhance data quality with minimal risk. However, they cautioned that this shift should only be considered once there is absolute confidence in the internal grading standards and data quality of the organization.

P14 shared insights into the unique nature of their verification processes, indicating that their experiences might not be directly transferable to other contexts. In their view, the specific practices developed in their organization are well-suited to their unique operational environment and legal framework.

4.2.7. Relation between the Themes

This section explores the interrelation between the main themes emerged from the interviews, which are highlighted in the accompanying figure and table. Figure 4.1 visually represents the relationships among these themes, based on the findings of the interviews. Each arrow in the figure delineates a specific relationship.

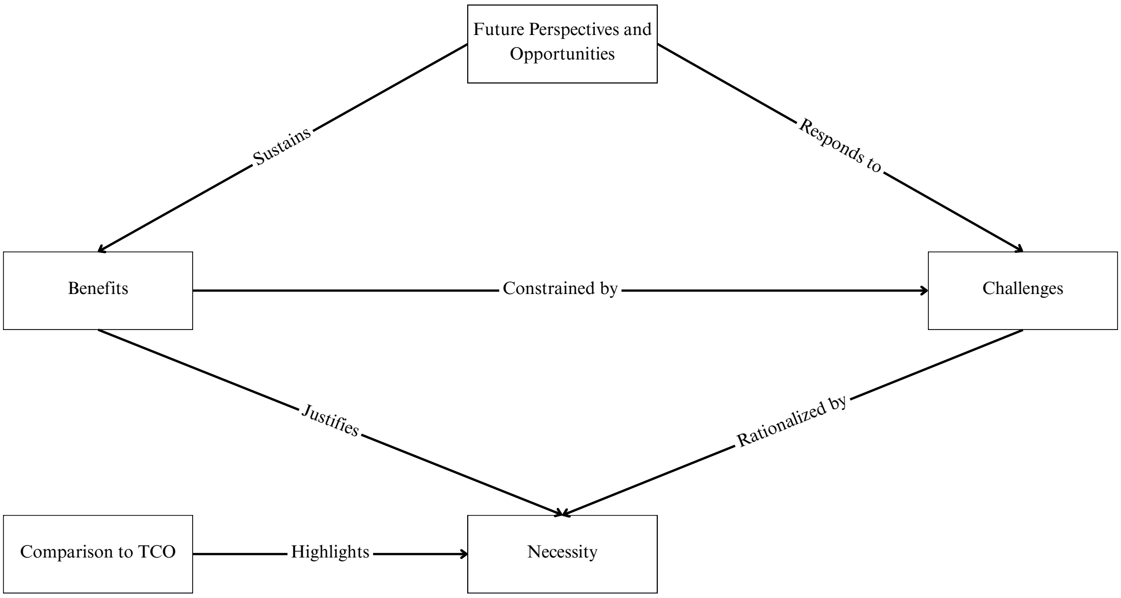

Figure 4.1: Interrelations between Themes

Table 4.2 outlines the specific interactions between themes, categorizing their connections. In the sections that follow, the nuances of these interrelations are described.

Table 4.2: Interrelations Between Themes

From	To	Relation	Explanation
Benefits	Necessity	Justifies	Benefits explain the need for verification.
Benefits	Challenges	Limited by	Challenges constrain achievable benefits.
Challenges	Necessity	Rationalized by	The need outweighs practical challenges.
Future Persp.	Challenges	Addresses	Innovation could resolve challenges.
Future Persp.	Benefits	Enhances	Tools could maintain and grow benefits.
Comp. to TCO	Necessity	Reinforces	Differences with TCO stress CSAM's need for strict checks.

An interrelation exists between the perceived benefits and challenges of triple verification. While participants described triple verification as contributing positively to the accuracy, legal robustness, and quality of CSAM databases, they simultaneously acknowledged the practical constraints associated with its implementation. Specifically, accuracy was mentioned both as an outcome enhanced by multiple independent assessments and as a construct that remains difficult to guarantee even under triple verification. Human judgment, often seen as beneficial for its ability to interpret contextual nuances, was also identified as a source of inconsistency and cognitive burden.

The theme of necessity is linked to the identified benefits, particularly in high-risk legal and ethical contexts. Participants referred to the value of triple verification in maintaining database integrity, minimizing false positives, and adhering to legal standards as central justifications for its continued use. These perceived benefits form the rationale for considering triple verification not merely as a preferred practice, but as a required measure in situations where the implications of error are severe. Legal considerations, in particular, were cited as making triple verification indispensable, especially in cases involving ambiguity or jurisdictional assessments.

Although participants identified various logistical and emotional challenges associated with triple verification, including high content volume, limited resources, and emotional fatigue, they emphasized that these burdens are often outweighed by the necessity of maintaining a high standard of verification. The need to ensure legal and ethical accountability was frequently invoked as a justification for maintaining the process despite its limitations. In this context, necessity functions as a mediating concept, enabling participants to rationalize continued adherence to a resource-intensive process by foregrounding its role in error prevention and judicial reliability.

The future-oriented perspectives shared by participants were primarily positioned as responses to the challenges identified in the current verification process. Automation, AI, and perceptual hashing were cited as potential solutions to reduce human workload, address the scale of content volume, and enhance procedural efficiency. However, these solutions were not presented as replacements for human oversight, but rather as supportive mechanisms intended to mitigate resource strain. Several participants advocated for a hybrid verification model, where AI could assist in filtering or prioritizing content, while final decisions would remain under human control.

Future developments in the verification process were also linked to the continuation and sustainability of its perceived benefits. Participants suggested that integrating AI and other technological methods may enable faster and more scalable assessments while preserving key advantages, such as contextual sensitivity and legal compliance. Technologies such as perceptual hashing were mentioned as having the potential to improve the efficiency and consistency of identifying duplicate or near-identical content, thereby strengthening database quality. These perspectives indicate that the adoption of new technologies is not intended to diminish the benefits of triple verification, but rather to preserve and extend them under conditions of growing content volume and limited human capacity.

The comparison between CSAM and TCO emerged as a reflective mechanism through which participants assessed the specificity and necessity of the triple verification process. Several participants noted that the legal and ethical thresholds for CSAM are more clearly defined and universally recognized than for TCO, which often involves greater interpretive ambiguity. Consequently, the rigorous standards applied to CSAM were viewed as more justified relative to TCO. Furthermore, the identified knowledge gap regarding TCO verification processes reinforced the perception that CSAM is governed by a more institutionalized and stringent verification framework.

4.2.8. Reflective Summary on Interview Perspectives

Across all interviews, a shared commitment to protecting children and ensuring a safer digital environment emerged as a common goal, despite considerable variation in verification practices and organizational roles. However, the rationales and perspectives expressed by participants often reflected the institutional context in which they operate. For instance, law enforcement agencies, whose core responsibility includes evidence collection for prosecution, tended to emphasize the legal rigor and necessity of triple verification. Their focus was strongly aligned with judicial reliability and the minimization of legal risk. Conversely, hotline employees, though equally dedicated to safety, were more concerned with efficient content removal and were often more vocal about operational constraints and resource limitations.

This divergence highlights a pattern. Participants working in similar types of organizations frequently held comparable views, influenced by shared mandates and workflows. Those in law enforcement roles were more likely to defend the status quo, whereas those in NGOs or hotlines displayed greater openness to innovation and resource-sensitive approaches. In some cases, interviewees openly reflected on the practical burdens of triple verification, citing costs, personnel strain, and emotional fatigue. For some, this led to procedural changes (e.g., shifting to double verification), while for others, such trade-offs were dismissed as secondary to mission integrity.

Notably, the necessity of triple verification was not consistently subjected to critical reflection. Participants from well-resourced organizations rarely questioned the cost-effectiveness of their procedures, implying that the availability of resources may limit the perceived need for alternative approaches. In contrast, organizations facing tighter constraints demonstrated greater awareness of operational burdens and were more inclined to explore future-oriented solutions.

When tying responses together into overarching research themes, it became evident that many participants focused more heavily on the benefits and necessity of triple verification, often emphasizing legal safeguards, while paying less attention to the feasibility challenges or alternative strategies. In particular, the comparison with TCO revealed both knowledge gaps and differing attitudes about ambiguity and subjectivity in classification, further reinforcing the exceptionalism surrounding CSAM moderation.

In sum, while there was broad alignment on the importance of verification, the way in which this importance was justified, operationalized, and evaluated varied considerably. This variation was shaped not only by professional roles and organizational mandates but also by the degree of resource availability, willingness to innovate, and cultural openness to self-criticism.

Key Takeaways from the Thematic Analysis

- **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. Results Phase 2a: Experiment

This and Section 4.4 will provide an answer to the following sub-question: *SQ4: How do different verification approaches affect the final classification of potentially illegal content?*

In total, the raters classified 2,031 items across both conditions. However, the output files contained a higher number of rows. This is because the system uses perceptual hashing to visually group near-duplicate files. As a result, items that are visually similar may appear as separate entries in the output. This explains why the blind condition contains 1,140 items and the non-blind condition 1,373, even though the number of classified items was fixed at 2,031.

4.3.1. First Phase: Blind Conditions

In the first condition, inter-annotator agreement was assessed using Cohen's Kappa to evaluate consistency between two raters. The dataset consisted of 1,140 items classified into three categories: *CSAM*, *Animal Pornography*, and *Other*. The raters agreed on 1,019 of 1140 items (89.4%) and Cohen's Kappa of 0.670, indicating a *substantial* level of agreement (Landis & Koch, 1977).

As can be seen in Table 4.3, the majority of the agreements occurred along the diagonal of the matrix, particularly in the *CSAM* category, where both raters agreed in 863 out of 919 cases (93.9%). However, most disagreements arose between the *CSAM* and *Other* categories. Specifically, 54 files labeled as *CSAM* by Rater A were labeled as *Other* by Rater B, while 63 files labeled as *Other* by Rater A were labeled as *CSAM* by Rater B. Rater A and Rater B both labeled 52 items as *Animal Pornography*, but they agreed on 50 of those cases (96.2%).

Table 4.3: Crosstabulation of Rater A and Rater B classifications in the blind condition

		Rater B			
		Animal Porn.	CSAM	Other	**Total**
Rater A	Animal Porn.	50	1	1	52
	CSAM	2	863	54	919
	Other	0	63	106	169
	Total	52	927	161	1140

Inter-annotator agreement was assessed separately for image and video files using Cohen's Kappa. For image files (n = 938), agreement was *moderate to substantial* with a Kappa of 0.601. Disagreements were most frequent between *CSAM* and *Other* categories. For video files (n = 202), agreement was higher, with a Kappa of 0.812 indicating *almost perfect* agreement. This suggests that reviewers found video material easier to classify consistently.

Table 4.3 presents the inter-annotator cross-tabulation for image and video files. Among image files (n = 938), reviewers agreed on 835 files (89%), with most disagreements occurring between the *CSAM* and *Other* categories (58 and 42 cases respectively). In contrast, video files (n = 202) showed higher agreement with only one mismatch in the *Animal Pornography* category and fewer *CSAM-Other* mismatches (17 in total). This indicates that video content may provide clearer cues for classification.

Table 4.4: Crosstabulation of Rater A and Rater B classifications by file type in the blind condition

			Rater B			
			Animal Porn.	CSAM	Other	**Total**
Image	**Rater A**	Animal Porn.	4	1	1	6
		CSAM	1	734	42	777
		Other	0	58	97	155
		Total	5	793	140	938
Video	**Rater A**	Animal Porn.	46	0	0	46
		CSAM	1	129	12	142
		Other	0	5	9	14
		Total	47	134	21	202

4.3.2. Second Phase: Non-blind Conditions

In the second condition, inter-annotator agreement was again assessed using Cohen's Kappa to evaluate the consistency of independent categorization by two raters. The dataset consisted of 1,373 items classified into three categories: *CSAM*, *Animal Pornography*, and *Other*. The raters agreed on 1,333 out of 1,373 items (97.1%), and Cohen's Kappa was calculated as 0.893, indicating *almost perfect* agreement according to the interpretation guidelines proposed by Landis and Koch (1977).

As shown in Table 4.5, nearly all classifications were aligned between the two raters. The highest agreement was found in the *CSAM* category, with both raters assigning this label to 1,143 out of 1,164 items labeled as such by Rater A and B (98.2%). Agreement was almost perfect in the *Animal Pornography* category, where 1 disagreement was labeled.

Only a small number of disagreements occurred: 21 files labeled as *CSAM* by Rater A were classified as *Other* by Rater B, and 18 files labeled as *Other* by Rater A were labeled as *CSAM* by Rater B. One additional discrepancy occurred between *Animal Pornography* and *Other*. These mismatches again point to some ambiguity between *CSAM* and *Other*, though far less frequent than in the first condition.

Table 4.5: Crosstabulation of Rater A and Rater B classifications in the non-blind condition

		Rater B			
		Animal Porn.	CSAM	Other	**Total**
Rater A	Animal Porn.	114	0	0	114
	CSAM	0	1143	21	1164
	Other	1	18	76	95
	Total	115	1161	97	1373

To examine whether inter-annotator agreement differed between file types in the second condition, Cohen's Kappa was calculated separately for image and video files. For image files (n = 1,138), Cohen's Kappa was 0.893, indicating *almost perfect* agreement. For video files (n = 235), Cohen's Kappa was slightly lower at 0.829, also falling within the range of *almost perfect* agreement (Landis & Koch, 1977).

Table 4.6 presents the inter-annotator cross-tabulation for image and video files in the non-blind condition. Among image files (n = 1,138), raters agreed on the majority of cases, with high agreement in the *CSAM* category (97.9%). Most disagreements occurred between the *CSAM* and *Other* categories, with 20 files labeled as *CSAM* by Rater A and *Other* by Rater B, and 17 files showing the reverse pattern.

In contrast, video files (*n* = 235) showed near-perfect agreement across all categories. Of the 229 files labeled as *CSAM* by Rater A, 228 (99.6%) were labeled the same by Rater B. The remaining video files labeled as *Other* by Rater A also showed strong agreement, with 5 out of 6 matches. No video files were labeled as *Animal Pornography* by Rater A in this set. These results indicate highly consistent classifications in the non-blind setting, with video files again demonstrating minimal ambiguity.

Table 4.6: Crosstabulation of Rater A and Rater B classifications by file type in the non-blind condition

			Rater B			
			Animal Porn.	CSAM	Other	**Total**
Image	**Rater A**	Animal Porn.	114	0	0	114
		CSAM	0	915	20	935
		Other	1	17	71	89
		Total	115	932	91	1138
Video	**Rater A**	CSAM	0	228	1	229
		Other	1	5	6	6
		Total	1	229	6	235

To investigate whether the position of the raters in the voting sequence influenced inter-annotator agreement, a chi-square test of independence was conducted. The variable Agreement indicated whether Rater A and Rater B assigned the same label (1 = agreement, 0 = disagreement). The variable RaterA_Voteposition denoted whether Rater A was the second or third voter in the classification process. The same goes for the variable RaterB_VotePosition.

As shown in Table 4.7, agreement occurred in 682 of 709 cases (96.2%) of the cases when Rater A voted second. In the cases where Rater A voted third, 651 out of 664 (98.0%) agreement occurred. Despite the overall high agreement, the chi-square test revealed a statistical difference between these two conditions, $\chi^2(1, N = 1373) = 4.151, p = 0.042$.

Since the voting order was systematically reversed between Rater A and Rater B across the datasets, the corresponding analysis for Rater B yielded identical counts and significance values. Specifically, when Rater B voted second, agreement occurred in 651 out of 664 cases (98.0%), and when Rater B voted third, in 682 out of 709 cases (96.2%). For this reason, only the cross-tabulation for Rater A is presented.

Table 4.7: Crosstabulation of agreement by Rater A's voting position

		Agreement		
		No Agreement	Agreement	**Total**
Rater A Position	Second voter	27	682	709
	Third voter	13	651	664
	Total	40	1333	1373

In this condition, the classification process was in another environment, enabling the system to log a Mastercategory once three votes were assigned to a file. This made it possible to infer not just agreement between Rater A and Rater B, but also whether consensus had been reached with the third vote (which was already given before the start of this experiment).

As shown in Table 4.8, full agreement was observed in 1257 of the 1,373 files (91.6%). In 76 of the 1373 cases (5.5%), Rater A and B agreed with each other but not with the third rater, resulting in no accepted label. In the remaining 2.9% of the cases, raters A and B disagreed, and no Mastercategory was recorded. These results indicate a high level of convergence in the non-blind setting, though the presence of unresolved or conflicting votes in roughly 8.4% of the files underscores the inherent difficulty of the classification task.

Table 4.8: Distribution of agreement levels across three raters in the non-blind condition

Agreement Level (3 Raters)	**Count**	**Percent**
A & B agreed, third disagreed	76	5.5%
Disagreement, unresolved	40	2.9%
Full agreement	1257	91.6%
Total	1373	100.0%

4.3.3. Comparison between both conditions

A direct comparison between the blind and non-blind conditions reveals several differences in annotation outcomes (See Table 4.9).

Table 4.9: Comparison of annotation outcomes between blind and non-blind conditions across multiple dimensions.

Metric	**Blind Condition**	**Non-Blind Condition**
Number of items	1,140	1,373
Overall agreement two-rater	89.4%	97.1%
Cohen’s Kappa (overall)	0.670	0.893
By file type		
Kappa – Image files	0.601	0.893
Kappa – Video files	0.812	0.829
Main disagreement pattern	CSAM–Other (117 files)	CSAM–Other (37 files)
Voting order effect (Rater A)	–	$\chi^2 = 4.151$, $p = .042$
Agreement (Rater A = 2nd)	–	96.2%
Agreement (Rater A = 3rd)	–	98.0%
Agreement (Rater B = 2nd)	–	98.0 %
Agreement (Rater B = 3rd)	–	96.2 %
Three-rater agreement	-	
Full agreement	–	91.6%
A & B agreed, third disagreed	–	5.5%
All disagreed (unresolved)	–	2.9%

First, overall inter-annotator agreement improved substantially in the non-blind condition. Cohen’s Kappa increased from 0.670 in the blind setting to 0.893 in the non-blind setting, reflecting a shift from *substantial* to *almost perfect* agreement (Landis & Koch, 1977). This increase was accompanied by a higher raw agreement percentage, rising from 89.4% to 97.1%, suggesting that the availability of contextual or prior input in the non-blind condition contributed to greater consistency between raters.

The improvement was observed across both image and video files. For image files, Kappa increased from 0.601 to 0.893; for video files, it rose slightly from 0.812 to 0.829. While agreement in the video category was already high in the blind setting, the non-blind setup appears to have stabilized classifications further, especially for images.

Beyond agreement scores, the nature of disagreements also shifted. In the blind condition, most disagreements occurred between the *CSAM* and *Other* categories, with 117 total mismatches across those two labels. In the non-blind condition, these mismatches were far fewer: only 37 files were misaligned between those categories. This reduction suggests that the collaborative or informed environment in the second condition may have helped raters converge.

Voting order also played a role. A Chi-square test revealed a statistically significant association between a rater’s voting position and the likelihood of agreement $\chi^2(1, N = 1373) = 4.151, p = 0.042$. Specifically, when Rater A voted third, agreement with Rater B was slightly higher (98.0%) than when Rater A voted second (96.2%), suggesting that seeing one prior label may have had a subtle priming or consensus effect. Interestingly, the opposite pattern was observed for Rater B: when B voted second, agreement was 98.0%, but when voting third, it dropped slightly to 96.2%. This asymmetry indicates that the effect of voting order on agreement was not uniform across raters, suggesting that individual differences may interact with positional effects.

Finally, in the non-blind condition, it was also possible to assess three-rater agreement. This analysis showed that 91.6% of all files received full agreement from all three raters. In 5.5% of cases, Raters A and B agreed, but the third rater disagreed, and 2.9% of cases remained fully unresolved. These figures highlight the strength of agreement under the non-blind setting, while also highlighting the persistence of ambiguity in a small but relevant portion of the dataset.

Key Takeaways from the Quantitative Analysis

- **Non-blind conditions yield significantly higher agreement.** Cohen's Kappa rose from 0.670 (blind) to 0.893 (non-blind), suggesting contextual cues enhance consistency.
- **Video content is classified more reliably than images.** Agreement was consistently higher for videos, indicating clearer classification cues.
- **Disagreements persist primarily between CSAM and Other.** These categories caused most confusion, though mismatches dropped from 117 (blind) to 37 (non-blind).
- **Voting order subtly affects agreement.** Raters voting third had slightly higher agreement rates, suggesting a mild positional influence ($p = 0.042$).

4.4. Results Phase 2b: Focus Group

In the final phase of the analysis, a follow-up discussion was conducted with the raters to further investigate the disagreements that emerged from the data. The results from phase 2a are summarized in Table 4.10. It shows a total of 121 mismatches in the blind condition and 40 in the non-blind condition. Most differences were observed in image files, accounting for 103 (blind) and 38 (non-blind) cases. The most prevalent disagreement concerned files labeled as *CSAM* by one rater and *Other* by the other.

Table 4.10: Overview of mismatched file type and category combinations per condition

Category Combination	Blind Condition	Non-Blind Condition
Video files		
CSAM vs *Other*	17	2
CSAM vs *Animal Porn.*	1	0
Subtotal video mismatches	**18**	**2**
Image files		
CSAM vs *Animal Porn.*	2	0
Animal Porn. vs *Other*	1	1
CSAM vs *Other*	100	37
Subtotal image mismatches	**103**	**38**
Total mismatches	**121**	**40**

A smaller number of disagreements involved the *Animal Pornography* category. In the blind condition, four such cases were recorded: two where one rater labeled a file as *CSAM* and the other as *Animal Pornography* (both in images), one *CSAM vs Animal Pornography* in a video file, and one *Animal Pornography vs Other* in an image file. In the non-blind condition, only a single disagreement involving *Animal Pornography* was observed (*Animal Pornography vs Other* in an image file.

The review of 49 items revealed key themes (See Table 4.11). The table presents the themes, along with the number of times each theme was referenced by the reviewers when discussing the items. These frequencies offer insight into which aspects were most mentioned in the reviewers' reasoning. Series Recognition and Age Estimation were the most frequently mentioned themes. In contrast, themes such as Blind vs. Non-Blind and Volume Pressure were not directly mentioned in the annotation text, but were discussed afterwards.

It should be noted that these themes are not mutually exclusive per item. In several instances, different reviewers identified distinct interpretive challenges within the same item. For example, one reviewer might have perceived an image as part of a larger illegal series (i.e., access to other frames in a series), while another judged it as a standalone and non-criminal photograph.

Table 4.11: Overview of interpretive themes in content classification

Theme	**Key Insight**	**Times Mentioned**
Series Recognition	Contextual cues often override content. Series were not always visible due to technical limitations.	22
Age Estimation	Raters struggled with low-resolution images and ethnic variation. Proxy indicators (e.g., teeth, skin) were sometimes used.	20
Pose, Framing, and Intent	Sexualized framing influenced judgments. Disagreements centered on intent vs. context.	17
Animal Pornography	Visuals involving animals created legal uncertainty. Classification varied due to framing, realism, and interpretation of sexual contact.	6
Textual Cues	Embedded text altered item interpretation. Legal relevance of captions remains uncertain.	7
Blind vs. Non-Blind	Peer visibility aided some, but autonomy and legal accountability remained central. Trust varied by familiarity.	-
Volume Pressure	High throughput demands limit depth. Speed may lead to misclassification.	-

4.4.1. Series Recognition

Within the Dutch National Police, when assessing the seized material, the case law of the Supreme Court has been taken into account (Supreme Court, 2010). Herein, the Supreme Court ruled that, if there is a series of visual representations, within which there is a coherence in terms of substantive characteristics and/or the manner of creation and within which a number of visual representations have an unmistakably sexual purport, the entire series may be deemed CSAM because of this mutual connection. Therefore, if such a connection was established in this investigation within a series of visual representations, the entire series was classified as CSAM.

Across the annotation sessions, both raters repeatedly emphasized that single images, when assessed independently, often did not meet the legal threshold for CSAM. However, their classification shifted when the image was known to be part of a broader, identifiable series. However, this reliance on series information introduced inconsistencies due to technological and procedural limitations. Raters pointed out that they did not always have access to other images in a series or could not confirm their inclusion due to system constraints.

In multiple cases, the visible content of the photo was ambiguous or non-sexual in isolation, but knowledge that it belonged to a recognized CSAM series led to its classification as illegal. For example, one annotator explained, *'On its own this image is not punishable, but I know it's part of a series. That makes it illegal.'* Another added, *'This is part of a known child pornographic series. We recognize the setup and the logo.'*

This reasoning highlights a critical mechanism in classification: legal status can be constructed retrospectively through series-based inference. In some instances, raters disagreed specifically because only one had seen or recognized the series context, leading to a mismatch in classification. One stated, *'With just this image, I don't find it punishable. But if another zoomed-in photo exists in the same batch,* *then this one becomes problematic too.'*

This reliance on series-based context, while central to the classification decisions observed in this study, appears to be highly specific to the operational realities of law enforcement workflows. Unlike hotlines or platform-based moderators who typically assess individual items in isolation, police investigators often work with seized devices containing millions of files, including large volumes of screenshots and images that form identifiable series. In such settings, the legal classification of an image may shift retrospectively based on its association with other files.

Notably, this dynamic has not been widely addressed in existing academic literature or surfaced in prior findings of this thesis, suggesting that it remains an underexplored but significant feature of forensic moderation practice. Its emergence during the focus group, rather than earlier phases of the study, highlights how context-sensitive these insights are.

4.4.2. Age Estimation

Another dominant theme was the difficulty of age estimation, especially in images where the subject's development, attire, or ethnicity complicated visual assessment. Annotators expressed repeated uncertainty in estimating whether individuals were minors, particularly when the image quality was poor or when the subject displayed ambiguous physical maturity.

A critical factor mentioned was the challenge of assessing individuals of Asian descent, as both annotators noted that smoother skin, smaller body frames, and delayed development of secondary sex characteristics made age harder to identify. One rater commented, *'I can't estimate the age properly. It's also someone of Asian appearance, which makes it more difficult.'*

Some annotators used skin clarity and body proportions as indicators for age. One noted, *'You can see it's a minor from the even, clean skin. No blemishes. That's how I recognize youth.'* Others relied on indirect cues: *'Based on the size of her face and the adult hand next to it, I estimate she's underage.'* This reliance on heuristics reflects both the visual limitations of image content and the absence of objective age indicators.

Physical development cues like breast formation or the presence of pubic hair were considered, but were often obscured or not visible due to bad image or video quality. As another rater explained, *'I can't even tell if it's a boy or a girl. The resolution is too low, and the pose doesn't help.'*

In some cases, subtle indicators such as dental development or skin texture were cited as indicators for age. One rater shared: *'For me, it was the teeth—clearly a minor. The development stage of the teeth gave it away.'*

The implications of such uncertainty are profound: in the absence of definitive cues, annotators defaulted to the category *Other.* As one noted, *'Even if I suspect she's underage, I can't say it with certainty. So I won't classify it as CSAM.'*

4.4.3. Pose, Framing, and Intent

Beyond visible nudity or sexual acts, a number of discussions centered on inferred intent, particularly in cases where subjects were clothed or partially clothed. Raters evaluated body position, gaze direction, and camera angle to determine whether the image was constructed with sexual intent.

In multiple cases, images of minors seated with legs spread or partially lifting clothing prompted concern due to the framing, even in the absence of nudity. Annotators noted that the focal point of the image (what the camera emphasized) played a central role. One rater observed, *'She is sitting with her legs apart. The way it's photographed really draws attention to her underwear.'*

Yet, interpretation varied. Some argued that such poses could occur naturally in non-sexual contexts. *'It could just be a girl standing at a nudist campsite,'* said one. 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 interpretive differences highlight the grey area in CSAM classification: sexualization is often inferred, not observed. Without accompanying metadata or textual cues, raters must rely on implicit visual signals, which can be influenced by personal and cultural norms.

4.4.4. Ambiguity in Animal Pornography Cases

Images involving animals introduced another layer of complexity. While direct sexual contact between children and animals was consistently classified as *CSAM*, several cases presented ambiguous situations. These included unclear spatial relationships, suggestive positioning, or content that lacked visible interaction due to framing or resolution.

One case involved an image showing a sexual act between a human and an animal, which one rater labeled as *Animal Pornography* and another as *Other*. The disagreement appeared to stem from the image's virtual or drawn nature, which complicated its legal interpretation. Without photographic realism, the file was not clearly punishable, even though the depicted act was sexual in nature.

In another case, a rater suggested the image was likely a still from a video and explained, *'This picture alone isn't illegal, but I suspect it's part of a series that is'*. Another countered, *'If there's no visible act, it might be inappropriate but not punishable.*

Another example concerned a video in which a dog licked a seemingly underage girl. One rater classified it as *CSAM*, citing visible breast development below the age of 18. Another rater hesitated, pointing to visible pubic hair and uncertainty about age due to poor resolution.

Two final cases initially led to classification as *Animal Pornography* but were later reinterpreted as *CSAM* after closer inspection. Upon zooming in, the rater noted, *"I now see that it's actually CSAM,"* emphasizing the small, fragile appearance of the person involved. Another described, *"You clearly see a dog licking a woman, very small and delicate person"*. The image also consisted of many smaller thumbnail photos, which further influenced the judgment.

4.4.5. Textual Cues

Another emergent theme was the impact of text embedded in or surrounding images. Some raters noted that while the visual content might be ambiguous, accompanying text could create a suggestive or sexual context.

For instance, one case involved a minor holding a banana, accompanied by a caption: *'She's learning to take it deep in her mouth.'* A rater said, *'Without the text, I'd classify it as neutral. But the caption makes it sexual.'*

However, there was disagreement about whether text alone should elevate a photo to illegal status. *'I know there's jurisprudence on this, but I don't remember the details,'* one said. *'If someone types that in a chat, it makes the image CSAM. But if it's embedded text, I'm not sure.'*

This points to a legal and ethical grey area: textual framing can shift interpretation, but at this point the raters were not certain whether such cues satisfy legal definitions.

4.4.6. Volume Pressure and Rating Conditions

After discussing how they approached the classification of the content, the raters reflected on the conditions under which they performed the experiment: blind and non-blind. During this discussion, participants also spontaneously raised concerns about volume pressure and time constraints, highlighting how these factors shaped their ability to make careful decisions.

Annotators reflected on how these setups influenced their decisions. *'In some doubt cases, I looked at what the others answered,'* one explained. However, they emphasized that this did not lead to automatic conformity: *'Even if others say CSAM, if I can't defend that in court, I won't follow.'*. Importantly, participants stressed that professional responsibility overrides consensus. *'I'm the one who has to raise two fingers in court. So I need to stand by my decision.'*

Trust in peer judgment varied. Familiar colleagues were seen as more credible. *'If it's someone I know and trust, I'm more likely to follow. But if it's a new reviewer, I'm more critical.'*. Another described a selective trust in certain colleagues: *'It depends on who gave the vote. If it's someone I know and trust, I'm more likely to align.'*

The data suggests that while the non-blind condition may support consistency and reduce oversight in ambiguous cases, it also risks introducing conformity bias. Still, raters saw value in the method when applied judiciously.

The raters also commented on the high volume of content and its impact on judgment. One noted, *'I reviewed 120,000 images yesterday. I probably missed something, but it's unavoidable'*. Speed affected the depth of evaluation, particularly for borderline images or videos requiring multiple frames to assess. *'When it's a million pictures, you don't have time to zoom in on everything,'* a participant said. Another added, *'Sometimes you just don't have the time to see if someone's wearing underwear.'*

Key Takeaways from the Focus Group Analysis

- **Series are defined by the presence of punishable content.** Raters could only classify an on its own neutral image as CSAM when they had encountered at least one punishable item from the same series.
- **Age estimation is highly uncertain, especially across ethnicities.** Raters relied on proxies like skin texture, dental development, and body proportions, but image quality and ethnicity complicated judgments.
- **Sexualized framing, not nudity, often drove CSAM classification.** Pose, angle, and visual focus influenced perceptions of intent. Disagreements arose where interpretations of natural vs. sexualized imagery diverged.
- **Non-blind conditions provided support, but not at the cost of autonomy.** Raters used peer input to resolve doubts, but retained final responsibility for decisions. Familiarity with the peer increased perceived credibility.

5 Discussion

This chapter discusses the findings and examines their theoretical and practical implications in depth. The results contribute to the existing body of knowledge and hold value for real-world applications. Additionally, the limitations encountered during the research process are addressed, acknowledging how these factors may influence the interpretation of the results. This reflection also paves the way for suggestions for future research.

5.1. Discussion of Findings

This study aimed to explore the different characteristics of verification processes of CSAM and TCO hash databases. Through a multiphase mixed-methods approach combining qualitative interviews, a controlled experiment, and a follow-up conversation, the findings offer a comprehensive view of both perspectives and actual verification behaviour.

The interviews revealed that verification processes vary across organizations, with differences in terminology, workflows, and voting thresholds. Participants consistently emphasized the need for accuracy and accountability but expressed concern over practical limitations such as emotional burden, classification ambiguity, and scalability. These challenges framed the perceived need for triple verification as both a safeguard against misclassification and a source of operational friction. While many participants valued the principle of multiple human checks, opinions diverged regarding its feasibility and necessity in every case.

These concerns were echoed in the quantitative results. The blind condition, which reflects a typical independent classification process, showed substantial agreement but also highlighted recurring points of confusion. Most disagreements occurred between the *CSAM* and *Other* categories, pointing to an interpretive grey zone that raters struggled to consistently navigate. In contrast, the non-blind condition, in which information about previous votes was available, led to significantly higher agreement. The increase in Cohen's Kappa from 0.670 to 0.893, along with a reduction in unresolved or mismatched cases, suggests that contextual input can support convergence and reduce the cognitive burden associated with independent decision-making.

Interestingly, agreement rates were not only affected by visibility but also by vote order. When Rater A voted third rather than second, agreement was statistically higher, suggesting that the order of exposure to prior labels may have an anchoring or alignment effect. This confirms what several interviewees described: even subtle forms of contextual awareness, such as non-blind voting or group-based review, can influence decision outcomes, for better or worse.

The analysis of three-rater outcomes further emphasized this point. While most cases reached full agreement, a meaningful portion disagreed or remained unresolved. These cases demonstrate the limits of consensus even in collaborative settings and reinforce the qualitative findings around the necessity of balancing consistency with efficiency.

The follow-up discussion with raters added interpretive depth to these statistical patterns. Raters confirmed that seeing others' votes helped resolve uncertainties, but were adamant that final responsibility remained individual: "I'm the one who has to raise two fingers in court." This dual commitment was echoed in statements about trust in specific colleagues, the weight of one's own experience, and selective attention to prior votes. While non-blind workflows may support convergence, they do not guarantee agreement, nor do they fully erase subjective interpretation.

The focus group also offered detailed insight into how raters negotiate ambiguity in real-world classification. Participants described concrete challenges such as inconsistent series visibility, difficulty estimating age, and uncertainty about whether suggestive text should influence classification. In many cases, disagreement was not due to a lack of expertise, but to differences in interpretation when visual cues were subtle or absent. Some raters used physical features like teeth or skin tone to estimate age, while others relied on photographic context or prior investigative knowledge.

Intent was similarly interpreted in diverse ways, with pose, gaze, or framing seen as either neutral or sexualized depending on individual thresholds. These discussions confirmed that ambiguity is an unavoidable part of content moderation, and that disagreement often reflects reasonable differences in professional judgment rather than error.

5.2. Implications

The findings of this study offer two types of implications: theoretical and practical implications.

Theoretical Implications

Based on the identified knowledge gap in Chapter 2, three gaps were identified: (1) operationalization of verification processes, (2) perspectives on the triple verification process, and (3) quantitative evidence on the operationalized processes.

1. Operationalization of Verification Processes

This research contributes to the theoretical understanding of CSAM and TCO hash database verification processes. It provides an overview of the varied operational processes employed by different organizations. It highlights the variability in these processes, including the number of analysts involved, the classification criteria for content, and the types of voting mechanisms used. Thus, it fills a gap in existing literature on the operationalization of verification processes.

2. Perspectives on the Triple Verification

This research contributes to the theoretical understanding of the perspectives on triple verification processes. The findings illustrate how organizations perceive the triple verification processes in combating online harmful content, shedding light on the perceived benefits of enhanced accuracy and accountability, the inherent challenges faced, and the ethical considerations that emphasize the necessity of such processes. Additionally, the examination of future opportunities related to technological advancements highlights the evolving nature of verification methodologies. Lastly, the comparative analysis with terrorist content highlights differences in verification practices and the varying levels of understanding among participants, thereby contributing valuable insights to the field of content moderation and harm reduction online.

3. Empirical Evaluation

This study contributes empirical evidence on the effects of blind and non-blind verification workflows in the context of CSAM moderation. Systematically comparing inter-rater agreement metrics under controlled conditions provides quantitative support for the theoretical claim that verification structures shape the consistency of classification outcomes. Unlike much of the existing literature, which is qualitative or conceptual, this study provides a measurable analysis of accuracy and agreement across varying configurations of verification.

Practical Implications

The findings of this study offer several practical implications for organizations involved in detecting and moderating harmful content, particularly those that rely on hash databases to classify CSAM or TCO material.

1. Differentiating Verification Depth to Use, Volume, and Content Type

A central implication of this study is that verification protocols for classifying potentially illegal content must be calibrated in accordance with three parameters: the intended use of the hash, the volume and structure of the material processed, and the legal classification of the content.

Findings across all phases of the research indicate that applying a uniform verification standard across all cases fails to accommodate the legal diversity, operational demands, and interpretive challenges of classifying potentially illegal content.

The first parameter is the intended use of the hash. Hashes do not serve a singular function. In practice, they are deployed across a range of legal and operational settings, each of which carries different implications for the necessary verification depth. Certain hashes are directly linked to legal processes. These include cases where the classification serves as input for criminal investigations, evidentiary files in prosecution, or mandatory registration in national law enforcement databases. In such situations, the consequences of a wrong classification can be severe. Mislabeling content may lead to procedural violations, wrongful accusations, or the undermining of judicial processes. As a result, these applications require rigorous verification protocols, such as blind triple review conducted by independent expert raters.

By contrast, many hashes are used in the context of preventive or large-scale moderation, such as automated takedown systems or proactive detection frameworks implemented by platforms. Although important, these applications typically involve lower legal thresholds. Here, the classification is not used as legal evidence but rather to suppress or prevent the distribution of harmful content. In such cases, lighter verification protocols may be appropriate, such as double verification.

The second consideration is the volume and structure of the material that must be reviewed. While both law enforcement agencies and content platforms deal with high volumes of potentially illegal content, the characteristics of this volume differ significantly. Law enforcement often processes large quantities of material on a per-case basis, particularly when investigating seized digital devices. A single case may involve the classification of hundreds of thousands of files, many of which must be processed under strict legal constraints. This results in a form of high-volume, bounded-case analysis, where verification must be thorough.

Platforms, hotlines, and NGOs also handle large-scale intake. However, their volume is frequently continuous and externally driven, consisting of user reports, automated detections, or referrals from third parties. Although these contexts also involve scale, the operational structure and review urgency differ. Verification strategies must therefore be responsive not only to the total amount of material processed, but also to the mode of intake, the time sensitivity of decisions, and the available resources within the organization.

The third parameter is the type of content, particularly in terms of its legal classification. This study draws a distinction between material that falls under the INTERPOL-defined baseline and content that is governed by national law. Baseline CSAM refers to material involving individuals under the age of thirteen depicted in sexual acts or explicit poses. This category is globally recognized as illegal and is generally more straightforward to classify. Experimental findings and interview data confirm that such material tends to generate higher agreement among expert reviewers.

In contrast, content involving individuals between the ages of thirteen and seventeen does not fall under the INTERPOL baseline. Its legality is governed by national legislation, which varies significantly across jurisdictions. In the Netherlands, for example, such material may still qualify as illegal under Article 252 of the Dutch Criminal Code, even if it does not involve nudity or overt sexual activity.

These cases are far more likely to involve interpretive complexity and disagreement among raters, particularly in the absence of contextual information. As indicated by the focus group, this age range yielded the highest rates of disagreement, highlighting the importance of reviewer independence and structured deliberation.

Taken together, these three parameters must be considered jointly when designing verification workflows. The findings of this study suggest that verification depth should increase as the legal consequences of the hash increase, as the content becomes more interpretively ambiguous, and as the risk associated with misclassification intensifies.

2. Managing Ambiguity with Structured Protocols and Institutional Awareness

A second implication of this study is the importance of addressing ambiguity more deliberately in the classification process. Ambiguity is not a rare or accidental feature of this work. It is a recurring reality, particularly in cases where content is ambiguous or where legal definitions are unclear. Rather than expecting reviewers to resolve these cases on their own, organizations should develop structured methods to handle ambiguity and incorporate them into the verification process.

This study revealed that ambiguity frequently arises in three areas: estimating the age of the person depicted, determining whether a pose is sexualized, and interpreting the overall intent or message of the image. These factors are especially difficult to judge in cases involving teenagers, where the person may be between thirteen and seventeen years old, and where there is no clear nudity or explicit sexual act.

In these situations, reviewers described using personal cues or heuristics, such as skin appearance, clothing, facial development, or posture. Some reviewers also relied on their background knowledge of similar cases or patterns in series. However, these methods varied from person to person and were not always consistent, which explains the higher rate of disagreement in such cases.

Instead of trying to eliminate this type of variation, organizations should recognize that it is part of the work and respond to it with clear structures. Verification systems should include options for reviewers to flag cases they find difficult, uncertain, or open to interpretation. This can be done through built-in features such as uncertainty markers, disagreement tracking, or predefined categories for ambiguous content. These cases can then be routed to a different review process, such as involving an additional reviewer, a group discussion, or a more thorough legal check.

Ambiguous cases should also be used in training. Institutions should organize regular calibration sessions where reviewers review difficult or previously disputed cases together. These sessions should focus on explaining how decisions are made, comparing different interpretations, and discussing how legal definitions apply. The goal is not to reach full agreement in every case, but to help reviewers understand where and why they differ, and how to make their reasoning more consistent over time.

3. Leveraging Non-Blind Verification as a Reflective Tool, Not a Rule

A third implication of this study relates to how reviewers interact with each other during the classification process. In particular, it concerns the practice of non-blind verification, where reviewers can see how their colleagues have voted before making their own decision. This method is already used in some operational settings, often to help reviewers align more quickly or resolve disagreements. However, it also introduces risks that organizations need to manage carefully.

The study showed that non-blind verification is not unusual in practice. Some systems are designed to show previous votes automatically, especially in workflows where speed is important or where reviewers are expected to check each other's work.

In the experiment, reviewers had access to each other's votes in the non-blind condition, and they often mentioned using this information as a way to reflect on their own decision. They did not blindly copy the earlier vote, but seeing how a colleague had classified a file sometimes made them reconsider or look again. This suggests that non-blind setups can support learning and reflection, especially when reviewers trust the expertise of the person who voted before them.

At the same time, non-blind verification can have unwanted effects. Research in other domains shows that people are often influenced by the decisions of others, especially when they feel unsure or tired. This is known as conformity bias. In the context of CSAM classification, such bias could mean that reviewers agree with a vote they do not fully support, just to avoid conflict or speed up the process. This is especially risky in ambiguous cases, where disagreement is common and independent judgment is essential. In such situations, the visibility of previous votes may reduce the quality of decisions by discouraging critical thinking.

Because of these mixed effects, non-blind verification should not be used as the default method for all cases. It can be helpful in some situations, but it must be handled with care. For instance, within law enforcement agencies, employees are generally highly aware of the weight of their votes due to the sensitive nature of their work. As indicated during the focus group, this awareness may reduce vulnerability to conformity effects, making non-blind verification less prone to convergence in such environments.

Similarly, in training contexts, exposing the outcome of the vote without disclosing individual voters could help reveal patterns of disagreement. This can stimulate discussion, promote consistency in judgment, and improve decision quality over time without enforcing convergence. These examples suggest that while non-blind verification has situational value, it must be applied deliberately, with attention to institutional culture, individual expertise, and the specific goals of the verification task.

If non-blind voting is used, the system should include safeguards. For example, it can include reminders that reviewers should think independently, even if they see another vote. It can also allow reviewers to add a short comment explaining why they agreed or disagreed with a previous decision. This promotes accountability and helps identify cases where agreement may not reflect true consensus. Additionally, systems should track how often reviewers follow or diverge from earlier votes, so that patterns of influence can be studied over time.

5.3. Limitations and Future Research

This section outlines the limitations of the research and provides suggestions for future research.

5.3.1. Research Limitations

As this study combines qualitative and experimental approaches, it is important to reflect on the methodological limitations of each phase in detail. Acknowledging these constraints provides clarity about the scope and reliability of the findings. In all phases, design considerations and decisions were made. The following sections will discuss those and reflect on them.

Phase 1: Interviews

While the interview phase was designed to capture expert perspectives from across the field, it was inevitably shaped by several methodological and contextual constraints.

First, the selection of participants was based on purposive sampling. While this ensured that all participants had relevant expertise, it inherently limited the diversity of perspectives. Additionally, several internationally relevant organizations, including NCMEC and Tech Against Terrorism, were not included due to non-response. As a result, while the sample provides operational depth, it does not fully represent the range of perspectives.

Additionally, the structure and number of interviews were limited by availability, scheduling, and the study's short timeframe. Although a total of 14 interviews were conducted, this sample size remains modest given the complexity of the subject matter and the global scale of the problem. One interview took place in a group setting with four individuals, which may have influenced the openness and individual expression of participants due to social dynamics or institutional context.

Another limitation lies in the semi-structured format of the interviews. While this format was well-suited to exploring complex processes and adapting to different organizational contexts, it introduced variability in how themes were probed and discussed. Not all topics were explored equally across participants, depending on their specific expertise, comfort level, and the direction of the conversation.

Furthermore, the interviews were conducted in both Dutch and English, depending on participant preference. All quotes were translated into English for reporting, which introduces the risk of minor semantic shifts. Although the transcriptions were verified against the audio, some nuance may have been lost in translation.

Lastly, the thematic analysis was carried out through inductive coding by a single researcher. While this ensured consistency and familiarity with the material, it also introduces interpretive subjectivity. No second coder was involved. As a result, the codes and themes reflect a single interpretive lens, shaped by the researcher's understanding of the domain and theoretical sensitivity.

Phase 2a: Experiment

The experimental design was also shaped by constraints beyond the researcher's control. These arose from the sensitive nature of the material, the institutional environment of law enforcement, and the technical properties of the systems in which the study was embedded.

First, the classification system used was not developed by the researcher but is legally and operationally defined by Dutch law and police standards. This ensured validity but limited conceptual flexibility. Similarly, the dataset itself was not randomly constructed. Instead, it was composed of real-world materials selected by the Dutch National Police. While the selection aimed to ensure diversity and avoid overrepresentation of series content, full control over the distribution, ordering, or thematic spread of items was not possible.

The study relied on existing content that had already received a first classification. This decision was practical and enabled a realistic simulation of double and triple verification with only two active raters. However, it introduces the potential for cognitive bias if the nature of the initial classification implicitly influenced raters' expectations, even though the prior vote was not shown in the blind condition.

Random assignment of voting roles was not possible due to technical restrictions in the police system. Instead, a counterbalanced sequence was constructed, where both raters alternated between acting as the second and third voter across the dataset. This approach introduced structure and fairness, yet it does not replicate true randomization, leaving open the possibility of order effects or fatigue influencing outcomes in undetected ways.

Finally, another limitation lies in the operationalization of ground truth as full agreement among three raters. This reflects standard practice in the Dutch National Police but should not be mistaken for an objective or uncontested truth. Full agreement may reflect alignment in interpretation rather than certainty.

Phase 2b: Focus Group

The focus group served as a supplementary component, designed to contextualize the findings from the experimental phase. While this added qualitative depth to the interpretation of disagreements, its methodological limitations should also be acknowledged. The session involved only the two raters who participated in the experiment, and the discussion covered a limited subset of 49 items. Due to logistical constraints, such as the hybrid setup and time spent retrieving files, not all disagreements could be reviewed.

Additionally, the session was informal and exploratory in nature, with no formal coding or validation procedure applied to the insights gathered. As such, the findings from the focus group should be interpreted as illustrative and indicative rather than comprehensive or generalizable. Nonetheless, the session offered valuable insight into how professional reasoning unfolds and complements the experimental data.

5.3.2. Future Research

Building on the findings and limitations of this study, several directions for future research are proposed.

A first set of recommendations relates to the diversity of participants and settings involved in future studies. This research primarily focused on operational experts recruited through internal networks, which offered depth but limited diversity. Future studies should adopt more inclusive sampling strategies to reach underrepresented groups, such as international policy actors or victim support organizations.

Second, the organizational context influences content moderation practices. Earlier research focused on the UK's A/B/C system, whereas this study highlights the approach of the Dutch National Police. Given the observed variation in decision-making workflows, additional cross-organizational studies are needed. Research that investigates the inter-annotator agreements across organizations with different categorical systems can provide valuable insights into how institutional design affects moderation quality and consistency.

Third, while this study examined triple verification, the blind condition has not yet been investigated. Future research could investigate how blind triple verification impacts the agreement between reviewers. It would also be helpful to include a focus group afterwards to understand how reviewers experienced the process and what influenced their decisions. This would be a useful addition to the current study, as this missing piece makes it difficult to draw strong conclusions about how well triple verification works.

Fourth, the findings of this study suggest that professional background, seniority, and specific training may influence decision-making in moderation processes. Future research should investigate whether factors such as age, expertise, institutional hierarchy, or specialized training influence the classifications. This could include experimental setups that simulate mixed-experience groups or surveys assessing perceived confidence and actual accuracy.

Fifth, several interviewees highlighted both the necessity and the challenges of combining algorithmic support with human decision-making. While automation can assist with filtering, categorization, or highlighting ambiguous cases, participants expressed concerns about overreliance on algorithms, lack of contextual understanding, and the risk of false confidence in AI-generated suggestions. These insights highlight the importance of future research in designing collaborative systems that effectively integrate human judgment with automated/AI/ML support.

Lastly, a recommendation is to evaluate how national legal frameworks influence content classification decisions. This study found that disagreement did not stem from the baseline criteria used to assess content but rather from how legal thresholds are applied in practice. Although age boundaries are clearly defined in national laws, moderators often struggle to determine whether content meets these thresholds, particularly when evaluating whether a pose constitutes sexual content under the law. Future research should therefore examine whether current legal definitions provide sufficient clarity and consistency for practical application.

5.4. Recommendations ATKM

This research was on behalf of ATKM. While the primary aim was to examine verification practices within the broader context of content moderation and the detection of illegal material, the findings also carry relevance for the ATKM itself. As such, this section concludes with recommendations for the ATKM.

Given that this study revealed the ambiguity in classifying specific types of CSAM, there is a clear need to collect and learn from such material systematically. It is recommended that the ATKM create a central repository of ambiguous or disputed cases (e.g., flagged content with internal disagreement or divergent classifications across organizations). This repository would serve both as an internal knowledge base and as a training resource for future employees.

A recurring insight from this research is that ambiguity in classification is not a failure of process but a characteristic of the material itself. To navigate this interpretive complexity, ATKM should actively facilitate peer learning with partner institutions, such as the Dutch National Police and/or Offlimits, which often deal with the same material. Beyond knowledge transfer, it would help align interpretive logics between the three key actors responsible for addressing this type of content in the Netherlands.

Ultimately, this research has emphasized the variability in how individual raters interpret the same material, even when protocols are clear and legal definitions are readily available. To support consistency, it is recommended that ATKM implement regular calibration sessions. These sessions should involve classifying a small set of complex cases, followed by reflective comparison of judgments and rationales. The focus should not only be on reaching consensus, but on understanding how disagreement arises, what constitutes acceptable variance, and how to get consistent classification.

6

Conclusion

This thesis explored how verification processes for CSAM and TCO hash databases are designed and applied. Using a mixed-methods approach with expert interviews, a large-scale annotation experiment (N=2,031), and a focus group, the study examined both organizational workflows and individual classification practices. It looked at how blind and non-blind voting affect agreement, and how annotators deal with ambiguous content.

The results show that verification procedures differ greatly across organizations in both terminology and setup. Triple verification is often seen as a safeguard, but its use depends on time, volume, and case complexity. The experiment showed that visibility of prior votes increased agreement, while the qualitative phases revealed that classification often involves subjective judgment. The focus group stressed the role of context, including image series, age estimation, and framing. Series recognition, in particular, seemed specific to law enforcement and only came up when raters reflected on their workflow together.

The discussion highlights how disagreement often stems from ambiguity rather than error. Age, intent, and context are not always clear, and decisions depend on the setup of the review process and the environment in which it happens. While triple verification could be a valuable mechanism, however, not all verification tasks carry the same operational or legal weight. The study points to the need for flexible systems, especially when dealing with large volumes and complex cases.

Declaration

I affirm that this thesis is my original work, and all contributions are my own unless otherwise noted. During the preparation of this document, I used Grammarly to maintain consistency in language and to improve its flow and readability. Furthermore, I utilized ChatGPT 4.0 to help identify and correct any LaTeX errors, create the tables and creative inspiration. Lastly, the Canva AI Image Creator was used to create the front page. The entire document has been carefully reviewed, edited, and revised, and I take complete responsibility for its content.

Bibliography

Adams, W. C. (2015). Conducting semi-structured interviews. *Handbook of practical program evaluation*, 492–505.

Aiken, M., Mc Mahon, C., Haughton, C., O'Neill, L., & O'Carroll, E. (2019). A consideration of the social impact of cybercrime: Examples from hacking, piracy, and child abuse material online. In *Crime and society* (pp. 91–109). Routledge.

Al-Nabki, M. W., Fidalgo, E., Alegre, E., & Alaiz-Rodriguez, R. (2023). Short text classification approach to identify child sexual exploitation material. *Scientific Reports*, *13*(1), 16108.

Anda, F., Le-Khac, N.-A., & Scanlon, M. (2020). Deepuage: Improving underage age estimation accuracy to aid csem investigation. *Forensic Science International: Digital Investigation*, *32*, 300921.

Appelman, N., & Leerssen, P. (2022). On" trusted" flaggers. *Yale JL & Tech.*, *24*, 452.

Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., et al. (2023). Detecting harmful content on online platforms: What platforms need vs. where research efforts go. *ACM Computing Surveys*, *56*(3), 1–17.

Baines, V. (2019). Online child sexual exploitation: Towards an optimal international response. *Journal of Cyber Policy*, *4*(2), 197–215.

Banko, M., MacKeen, B., & Ray, L. (2020). A Unified Taxonomy of Harmful Content. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 125–137. https://doi.org/10.18653/v1/2020.alw-1.16

Barrett, P. M. (2020). Who moderates the social media giants. *Center for Business*, *102*.

Bellanova, R., & De Goede, M. (2022). Co-producing security: Platform content moderation and european security integration. *JCMS: journal of common market studies*, *60*(5), 1316–1334.

Bleakley, P., Martellozzo, E., Spence, R., & DeMarco, J. (2024). Moderating online child sexual abuse material (csam): Does self-regulation work, or is greater state regulation needed? *European Journal of Criminology*, *21*(2), 231–250.

Bonagiri, A., Li, L., Oak, R., Babar, Z., Wojcieszak, M., & Chhabra, A. (2025). Towards safer social media platforms: Scalable and performant few-shot harmful content moderation using large language models. *arXiv preprint arXiv:2501.13976*.

Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., & Walker, K. (2020). Purposive sampling: Complex or simple? research case examples. *Journal of research in Nursing*, *25*(8), 652–661.

Canadian Centre for Child Protection. (2023). Home. https://www.projectarachnid.ca/en/#top

Chametzky, B., et al. (2016). Coding in classic grounded theory: I've done an interview; now what? *Sociology Mind*, *6*(04), 163.

Christensen, L. S., Rayment-McHugh, S., Prenzler, T., Chiu, Y.-N., & Webster, J. (2021). The theory and evidence behind law enforcement strategies that combat child sexual abuse material. *In-*

ternational Journal of Police Science Management, *23*(4), 392–405. https://doi.org/10.1177/14613557211026935

Church, P., & Pehlivan, C. N. (2023). The digital services act (dsa): A new era for online harms and intermediary liability. *Global Privacy Law Review*, *4*(1).

Claussen, V. (2018). Fighting hate speech and fake news. the network enforcement act (netzdg) in germany in the context of european legislation. *Media Laws*, *3*(3), 110–136.

Clune, C., & McDaid, E. (2024). Content moderation on social media: Constructing accountability in the digital space. *Accounting, Auditing & Accountability Journal*, *37*(1), 257–279.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., Smith, K. C., et al. (2011). Best practices for mixed methods research in the health sciences. *Bethesda (Maryland): National Institutes of Health*, *2013*, 541–545.

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, *10*, 92–110.

de Macedo Neto, J. J., et al. (2019). Csam detection based on age estimation from faces.

De Streel, A., Defreyne, E., Jacquemin, H., Ledger, M., Michel, A., Innesti, A., Goubet, M., & Ustowski, D. (2020). *Online platforms' moderation of illegal content online*. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf

Dos Santos, H., Martins, T. S., Barreto, J. A., Nakamura, L. H., Ranieri, C. M., Robson, E., Geraldo Filho, P., & Meneguette, R. I. (2024). Chasam: An architecture based on perceptual hashing for image detection in computer forensics. *IEEE Access*.

Draper, L. (2022). Protecting children in the age of end-to-end encryption.

Du, K., Shah, S., Bollepalli, S. C., Ibrahim, M. N., Gadari, A., Sutharahan, S., Sahel, J.-A., Chhablani, J., & Vupparaboina, K. K. (2024). Inter-rater reliability in labeling quality and pathological features of retinal oct scans: A customized annotation software approach. *PloS one*, *19*(12), e0314707.

European Commission. (2024, November). Trusted flaggers under the Digital Services Act (DSA). https://digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa

European Commission. (2025a, January). Het wetgevingspakket inzake digitale diensten. https://digital-strategy.ec.europa.eu/nl/policies/digital-services-act-package

European Commission. (2025b, January). List of national competent authority (authorities) and contact points. https://home-affairs.ec.europa.eu/policies/internal-security/counter-terrorism-and-radicalisation/prevention-radicalisation/terrorist-content-online/list-national-competent-authority-authorities-and-contact-points_en#netherlands

European Union. (2024, February). Directive - 2000/31 - EN - e-commerce directive - EUR-Lex. https://eur-lex.europa.eu/eli/dir/2000/31/oj/eng

Farid, H. (2021). An overview of perceptual hashing. *Journal of Online Trust and Safety*, *1*(1).

Fasel, M., & Weerts, S. (2024). Can facebook's community standards keep up with legal certainty? content moderation governance under the pressure of the digital services act. *Policy & Internet*.

Galli, F., Loreggia, A., & Sartor, G. (2022). The regulation of content moderation. *International conference on the legal challenges of the fourth industrial revolution*, 63–87.

Gewirtz-Meydan, A., Mitchell, K. J., & O'Brien, J. E. (2024). Trauma behind the keyboard: Exploring disparities in child sexual abuse materials exposure and mental health factors among investigators and forensic examiners–a network analysis. *Child Abuse & Neglect*, *152*, 106757.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data &Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Global Internet Forum to Counter Terrorism. (2023, April). GIFCT's Hash-Sharing Database | GIFCT. https://gifct.org/hsdb/

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

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data &Society*, *7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Grippo, V. (2024, December). *Regulating content moderation on social media to safeguard freedom of expression* (tech. rep.). Council of Europe.

Guerra, E., & Westlake, B. G. (2021). Detecting child sexual abuse images: Traits of child sexual exploitation hosting and displaying websites. *Child Abuse &Neglect*, *122*, 105336. https://doi.org/10.1016/j.chiabu.2021.105336

Gutfeter, W., Gajewska, J., & Pacut, A. (2023). Detecting sexually explicit content in the context of the child sexual abuse materials (csam): End-to-end classifiers and region-based networks. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 154–168.

He, Q., Hong, Y., & Raghu, T. (2024). Platform governance with algorithm-based content moderation: An empirical study on reddit. *Information Systems Research*.

INHOPE. (2025). What is Baseline? https://inhope.org/EN/articles/what-is-baseline

International Association of Internet Hotline Providers. (2025). INHOPE - Association of Internet Hotline Providers | The Facts. https://www.inhope.org/EN/the-facts

Internet Watch Foundation. (2025). Image Hash list. https://www.iwf.org.uk/our-technology/our-services/image-hash-list/

Interpol. (2025). International Child Sexual Exploitation database. https://www.interpol.int/Crimes/Crimes-against-children/International-Child-Sexual-Exploitation-database

Jain, P., & Sengar, S. (2024). Unraveling the role of ibm spss: A comprehensive examination of usage patterns, perceived benefits, and challenges in research practice. *Educational Administration: Theory and Practice*, *30*(5), 9523–9530.

Jansen, R. J., van der Kint, S. T., & Hermens, F. (2021). Does agreement mean accuracy? evaluating glance annotation in naturalistic driving data. *Behavior research methods*, *53*, 430–446.

Joffe, H. (2011). Thematic analysis. *Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners*, 209–223.

Jones, S. (2022). Interpreting themes from qualitative data: Thematic analysis. *Eval Academy*.

Kloess, J. A., Woodhams, J., & Hamilton-Giachritsis, C. E. (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*, 105166.

Kloess, J. A., Woodhams, J., Whittle, H., Grant, T., & Hamilton-Giachritsis, C. E. (2019). The challenges of identifying and classifying child sexual abuse material. *Sexual Abuse*, *31*(2), 173–196.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Langvardt, K. (2017). Regulating online content moderation. *Geo. LJ*, *106*, 1353.

Lanza, E., & Jackson, M. (2021). Content moderation and self-regulation mechanisms. *The Facebook Oversight Board and its Implications for Latin America. Inter-American Dialogue, Washington, DC.*

Laranjeira da Silva, C., Macedo, J., Avila, S., & dos Santos, J. (2022). Seeing without looking: Analysis pipeline for child sexual abuse datasets. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2189–2205.

Larroza, A., Pérez-Benito, F. J., Tendero, R., Perez-Cortes, J. C., Román, M., & Llobet, R. (2025). Three-blind validation strategy of deep learning models for image segmentation. *Journal of Imaging*, *11*(5), 170.

Lee, H.-E., Ermakova, T., Ververis, V., & Fabian, B. (2020). Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation*, *34*, 301022.

Link, D., Hellingrath, B., & Ling, J. (2016). A human-is-the-loop approach for semi-automated content moderation. *ISCRAM*.

Macdonald, S., Mattheis, A., & Wells, D. (2024). Using artificial intelligence and machine learning to identify terrorist content online.

Malina, M. A., Nørreklit, H. S., & Selto, F. H. (2011). Lessons learned: Advantages and disadvantages of mixed method research. *Qualitative Research in Accounting & Management*, *8*(1), 59–71.

Marchal, M., Scholman, M., Yung, F., & Demberg, V. (2022). Establishing annotation quality in multi-label annotations. *Proceedings of the 29th international conference on computational linguistics*, 3659–3668.

Marsoof, A., Luco, A., Tan, H., & Joty, S. (2022). Content-filtering AI systems–limitations, challenges and regulatory approaches. *Information & Communications Technology Law*, *32*(1), 64–101. https://doi.org/10.1080/13600834.2022.2078395

McGarvie, J. (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), 4.

Meggyesfalvi, B. (2024). Challenges in investigating self-generated online child sexual abuse material. *BELÜGYI SZEMLE/ACADEMIC JOURNAL OF INTERNAL AFFAIRS: A BELÜGYMINISZTÉRIUM SZAKMAI TUDOMÁNYOS FOLYÓIRATA (2010-)*, *72*(2), 329–339.

Mei, J., & Frank, R. (2015). Sentiment crawling: Extremist content collection through a sentiment analysis guided web-crawler. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 1024–1027.

Morais Carvalho, J., Arga e Lima, F., & Farinha, M. (2021). Introduction to the digital services act, content moderation and consumer protection. *Revista de Direito e Tecnologia*, *3*(1), 71–104.

Mulligan, D. K., & Bamberger, K. A. (2021). Allocating responsibility in content moderation: A functional framework. *Berkeley Tech. LJ*, *36*, 1091.

National Center for Missing & Exploited Children. (2022). *2021 CyberTipline reports by electronic service providers* (tech. rep.). https://www.missingkids.org/content/dam/missingkids/pdfs/2021-reports-by-esp.pdf

National Center for Missing & Exploited Children. (2023). CyberTipline Data. https://www.missingkids.org/cybertiplinedata

National Center for Missing & Exploited Children. (2025). About us. https://www.missingkids.org/footer/about

Ngo, V. M., Gajula, R., Thorpe, C., & Mckeever, S. (2024). Discovering child sexual abuse material creators' behaviors and preferences on the dark web. *Child Abuse & Neglect*, *147*, 106558.

Oronowicz-Jaśkowiak, W., Kozłowski, T., Polańska, M., Wojciechowski, J., Wasilewski, P., Ślęzak, D., & Kowaluk, M. (2024). Using expert-reviewed csam to train cnns and its anthropological analysis. *Journal of Forensic and Legal Medicine*, *101*, 102619.

Ozanne, M., Bhandari, A., Bazarova, N. N., & DiFranzo, D. (2022). Shall ai moderators be made visible? perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, *9*(2), 20539517221115666.

Parker, O. (2024). Navigating the privacy-freedom dilemma: The impact of ai on content moderation and free speech.

Parmar, M., Mishra, S., Geva, M., & Baral, C. (2022). Don't blame the annotator: Bias already starts in the annotation instructions. *arXiv preprint arXiv:2205.00415*.

Pereira, M., Dodhia, R., Anderson, H., & Brown, R. (2023). Metadata-based detection of child sexual abuse material. *IEEE Transactions on Dependable and Secure Computing*.

Popović, M. (2021). Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 234–243.

Ramesh Babu, B., Usha Rani, T., & Naga Kumari, Y. (2024). Ai's watchful eye: Protecting children from sexual abuse with artificial intelligence. In *Child sexual abuse: A public health problem in india* (pp. 441–455). Springer.

Rastogi, C., Song, X., Jin, Z., Stelmakh, I., Daumé III, H., Zhang, K., & Shah, N. B. (2024). A randomized controlled trial on anonymizing reviewers to each other in peer review discussions. *PloS one*, *19*(12), e0315674.

Reid, E., Qin, J., Zhou, Y., Lai, G., Sageman, M., Weimann, G., & Chen, H. (2005). Collecting and analyzing the presence of terrorists on the web: A case study of jihad websites. *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2005, Atlanta, GA, USA, May 19-20, 2005. Proceedings 3*, 402–411.

Rimez, D., Legay, A., & Macq, B. (2024). Ensuring data security and annotators anonymity through a secure and anonymous multiparty annotation system. *Novel & Intelligent Digital Systems Conferences*, 620–631.

Roberts, S. T. (2017, January). *Content moderation*. Springer eBooks. https://doi.org/10.1007/978-3-319-32001-4\{_}44-1

Rondeau, J. (2019). *Deep learning of human apparent age for the detection of sexually exploitative imagery of children*. University of Rhode Island.

Saleem, R., & Kamande, J. (2024). Redefining free speech: The impact of ai-driven content moderation on privacy and expression.

Salter, M., & Richardson, L. (2021). The Trichan takedown: Lessons in the governance and regulation of child sexual abuse material. *Policy & Internet*, *13*(3), 385–399. https://doi.org/10.1002/poi3.256

Sanchez, L., Grajeda, C., Baggili, I., & Hall, C. (2019). A practitioner survey exploring the value of forensic tools, ai, filtering, & safer presentation for investigating child sexual abuse material (csam). *Digital Investigation*, *29*, S124–S142.

Sang, Y., & Stanton, J. (2022). The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. *International Conference on Information*, 425–444.

Schmidt, T., Winterl, B., Maul, M., Schark, A., Vlad, A., & Wolff, C. (2019). Inter-rater agreement and usability: A comparative evaluation of annotation tools for sentiment annotation.

Schneider, P. J., & Rizoiu, M.-A. (2023). The effectiveness of moderating harmful online content. *Proceedings of the National Academy of Sciences*, *120*(34), e2307360120.

Scrivens, R., Gaudette, T., Davies, G., & Frank, R. (2019). Searching for extremist content online using the dark crawler and sentiment analysis. In *Methods of criminology and criminal justice research* (pp. 179–194, Vol. 24). Emerald Publishing Limited.

Seigfried-Spellar, K. C., Rogers, M. K., Matulis, N. L., & Heasley, J. S. (2024). Testing a hybrid risk assessment model: Predicting csam offender risk from digital forensic artifacts. *Child Abuse & Neglect*, *154*, 106908.

Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2023). SOK: content moderation in social media, from guidelines to enforcement, and research to practice. *2023 IEEE 8th European Symposium on Security and Privacy*, 868–895. https://doi.org/10.1109/eurosp57164.2023.00056

Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., & DeMarco, J. (2023). The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology Journal of Psychosocial Research on Cyberspace*, *17*(4). https://doi.org/10.5817/cp2023-4-8

Spiller, T. R., Rabe, F., Ben-Zion, Z., Korem, N., Burrer, A., Homan, P., Harpaz-Rotem, I., & Duek, O. (2023). Efficient and accurate transcription in mental health research-a tutorial on using whisper ai for audio file transcription. *OSF Preprint. https://doi. org/10.31219/osf. io/9fue8.*

Supreme Court. (2010, December). The Case Law. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:HR:2010:BO6446

Thakor, M., Sabnam, S., Ueno, R., & Zaslow, E. (2023). To search and protect? Content moderation and platform governance of explicit image material. *IT Case Studies in Social and Ethical Responsibilities of Computing*, (Summer 2023). https://doi.org/10.21428/2c646de5.cdecbadf

Trivison, A. (2024). Understanding the line between art and abuse: How generative ai changes the landscape of child sexual abuse materials. *Catholic University Journal of Law and Technology*, *33*(1), 87–114.

Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. *Big Data &Society*, *10*(1). https://doi.org/10.1177/20539517231172424

Van De Kerkhof, J., van Es, K., Helmond, A., & van der Vlist, F. (2024). Constitutional Aspects of Trusted Flaggers in The Netherlands. *Governing the Digital Society*. https://doi.org/10.2139/ssrn.4943851

Wagner, B., Rozgonyi, K., Sekwenz, M.-T., Cobbe, J., & Singh, J. (2020). Regulating transparency? facebook, twitter and the german network enforcement act. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 261–271.

Westlake, B., Bouchard, M., & Frank, R. (2012). Comparing methods for detecting child exploitation content online. *2012 European intelligence and security informatics conference*, 156–163.

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

Westlake, B., Brewer, R., Swearingen, T., Ross, A., Patterson, S., Michalski, D., Hole, M., Logos, K., Frank, R., Bright, D., et al. (2022). Developing automated methods to detect and match face and voice biometrics in child sexual abuse videos. *Trends and issues in crime and criminal justice*, (648), 1–15.

Wolbers, H., Cubitt, T., & Cahill, M. J. (2025). Artificial intelligence and child sexual abuse: A rapid evidence assessment. *Trends and Issues in Crime and Criminal Justice*, (711), 1–18.

Woodie, A. (2016). Yahoo shares algorithm for identifying "nsfw" images. *Datanami*.

Yan, Y., Rosales, R., Fung, G., Subramanian, R., & Dy, J. (2014). Learning from multiple annotators with varying expertise. *Machine learning*, *95*, 291–327.

A

List of National Competent Authorities

Table A.1: List of National Competent Authorities for Content Moderation in the EU

Country	**Competent Authority**
Austria	Kommunikationsbehörde Austria (KommAustria)
Belgium	Federal Prosecution Service; I2-IRU Section DJSOC; BIPT
Bulgaria	Ministry of Interior - GD Combating Organised Crime
Croatia	Ministarstvo unutarnjih poslova; HAKOM; Općinski prekršajni sud u Zagrebu
Cyprus	Cyprus Police; Ministry of Energy, Commerce and Industry
Czech Republic	National Counterterrorism, Extremism and Cybercrime Agency; Ministry of Interior; Czech Telecommunication Office
Denmark	Danish National Police; Prosecution Service; Courts of Denmark
Estonia	Estonian Internal Security Service; Estonian Technical Surveillance Authority
Finland	National Bureau of Investigation; Finnish Transport and Communications Agency Traficom; Sanctions Board at the National Police Board
France	OCLCTIC; ARCOM; Le juge judiciaire
Germany	Bundeskriminalamt; Bundesnetzagentur
Greece	Prosecutor of the Anti-terrorist Unit; National Intelligence Service; Telecommunications and Postal Committee
Hungary	Nemzeti Média- és Hírközlési Hatóság
Ireland	An Garda Síochána; Coimisiún na Meán
Italy	Uffici del Pubblico Ministero; Ministero dell'Interno; Ministero delle imprese e del Made in Italy
Latvia	State Security Service
Lithuania	Lithuanian Police; Communications Regulatory Authority

Continued on next page

Table A.1 – *Continued from previous page*

Country	**Competent Authority**
Luxembourg	Police Grand-Ducale; Haut-Commissariat à la Protection nationale
Malta	The Court of Justice; The Police; Critical Information Infrastructure Protection Unit
Netherlands	Autoriteit online Terroristisch en Kinderpornografisch Materiaal (ATKM)
Poland	Head of the Internal Security Agency
Portugal	(Specific authority not listed)
Romania	ANCOM; Serviciul Român de Informații; Inspectoratul General al Poliției Române
Slovakia	Police Force of the Slovak Republic; Council for Media Services
Slovenia	District Court Nova Gorica; Agency for Communication Networks and Services
Spain	The Centre for Intelligence against Terrorism and Organised Crime; Secretary of State for Security
Sweden	Polismyndigheten

B

Verbal Consent

You are being invited to participate in a research study titled 'Researching the effectiveness and reliability of triple verification in CSAM/TCO detection'. This study is conducted by myself, Melissa Rottier, from TU Delft in collaboration with the Autoriteit online Terroristisch en Kinderpornografisch Materiaal. The aim of my research is to explore the necessity and efficiency of triple verification in moderating Child Sexual Abuse Material and Terrorist Content Online.

This interview will take approximately 45-60 minutes. I will be asking you to share your insights and perspectives related to the verification processes of online harmful content databases. The data will be used for Melissa Rottier's master's thesis.

Your interview will be recorded in audio format solely for research purposes. Access to the recordings will be limited to the research team, and your identity will remain confidential in all reports and publications. The audio recording will be transcribed initially. Subsequently, the transcriptions will undergo anonymisation and will be securely stored. Data will be held for a period of two months following the conclusion of the study, after which it will be securely destroyed.

Participating in this study involves no anticipated risks beyond those encountered in everyday life. You have the right to choose not to participate or to withdraw from the study at any point without any penalty. You are also free to skip any questions during the interview that you do not wish to answer. Do you have any questions about the study or the procedures that we will be using?

[Wait for any questions and provide answers]

If everything is clear and you feel comfortable, can you please confirm that you consent to participate in this study, understanding that you can withdraw your consent at any time without any consequences to you?

[Start interview after consent confirmation]

C

Interview Protocols

C.1. General Interview Protocol

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?)

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?

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.2. 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 TC?

C.3. Other 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 TC?

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?

D

Qualitative Coding

Table D.1: Summary of qualitative coding scheme for 'Benefits'

Sub-Theme	**Code examples**	**Respondents (n=14)**
Legal Considerations	Judgement according to the law and jurisdiction, Law can sometimes be indistinct	14 (100%)
Content Assessment	Estimation of sexual act is difficult, Age estimation in CSAM is difficult	10 (71%)
Human Factors	Bias, Consistency	9 (64%)
Quality	Accuracy of hash database, Quality of hash database	8 (57%)
Operations	Tribal knowledge, Triple verification as evidence in criminal law	8 (57%)

Table D.2: Summary of qualitative coding scheme for 'Challenges'

Sub-Theme	**Code examples**	**Respondents (n=14)**
Human Impact	Impact assessing CSAM content on moderator, Continuous work of assessing illegal content online	14 (100%)
Content Volume	The amount of content is growing, Lot of duplicates	13 (93%)
Resource and Costs	Expensive process, Waste of analysts	10 (71%)
Technological Limitations	AI-generated content is becoming unmanageable, Different hash value if picture changes	6 (43%)
Accuracy	After match database some content needs rechecking, there is no 100% certainty in triple verification	5 (36%)

Table D.3: Summary of qualitative coding scheme for ‘Necessity‘

Sub-Theme	**Code examples**	**Respondents (n=14)**
Process	No understanding for the triple verification process, Understanding for the triple verification process	14 (100%)
Utility and Impact	Hash database is used by other parties and organizations, Hash databases can have major consequences	8 (57%)
Human Factors	Assessing content remains subjective, Experience of analyst can influence assessment	7 (50%)
Ethical Considerations	Baseline is easier, Classifying content can be obvious sometimes	7 (50%)

Table D.4: Summary of qualitative coding scheme for ‘Opportunities and Future Perspectives‘

Sub-Theme	**Code examples**	**Respondents (n=14)**
AI and ML	AI or ML models trained well enough, Reliability of AI remains a question	12 (86%)
Oversight	There should always be human oversight, What error rate is acceptable	10 (71%)
Automation	Automation is going to happen, Different type of hashing	9 (64%)
Technological Diversification	Perceptual hashing, Trade-off	5 (36%)

Table D.5: Summary of qualitative coding scheme for ‘Comparison to TCO‘

Sub-Theme	**Code examples**	**Respondents (n=14)**
Classification Criteria	Classification criteria TCO, Different requirements	7 (50%)
Verification	Verification process of TCO hash database, TCO is multi-interpretable	7 (50%)
Knowledge Gap	No understanding of verification process of TCO hash database, Uncertain how TCO content works	7 (50%)

Table D.6: Thematic Analysis of Codes and Quotes

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” (P7), ”For accuracy for their database as a standing argument, I get it” (P1)

Continued on next page

Table D.6 Continued from previous page

Main Theme	Sub-Theme	Code	Example Quotes
	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” (P5), ”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?” (p6)
	Inconsistent assessment	Humans make mistakes, Context is important and hard to classify	”Because yes, people make mistakes too” (P2), ”People just make mistakes, just stupid mistakes” (P5)
	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” (p5)
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” (P7)
	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” (P5)
	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” (P1)

Continued on next page

Table D.6 Continued from previous page

Main Theme	**Sub-Theme**	**Code**	**Example Quotes**
	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” (P2)
	Accuracy	There is no 100% certainty in triple verification, After match 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” (P6)
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” (P4), ”I don’t want people to suffer from this” (P5)
	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)
	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” (P4)
	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?” (P6)

Continued on next page

Table D.6 Continued from previous page

Main Theme	Sub-Theme	Code	Example Quotes
	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” (P8)
	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” (P2)
	Technological Diversification	Different type 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” (P7)
Comparison to TCO	Classification criteria	Classification criteria TCO (evident, grey, or not TCO), Example of difficult classifying content in TCO, TCO is multi-interpretable	”Because there is not one taxonomy or one certain criteria list, such as CSAM baseline, which makes it harder” (P6)
	Verification	Lots of variables in TCO that decide whether it is illegal, TCO hashes are added directly to database	”There are a lot of variables in classifying terrorist content that makes it very difficult” (P2)
	Knowledge Gap	No understanding of verification process of TCO hash database, No knowledge of this type of content, Uncertain how TCO content works	”I’ve never seen such content in my life so I’m not aware of how these processes work” (P10)