



ATKM conducted research into the verification of child sexual abuse material and terrorist content for the purpose of supporting databases

By Melissa Rottier

Background

The spread of Child Sexual Abuse Material (CSAM) and Terrorist Content Online (TCO) is a pressing societal problem. To quickly identify and remove such content, organizations use so-called hash databases, which are lists of digital fingerprints of previously verified material. If new content matches a hash in the database, further dissemination of the image can be stopped. However, incorrect classifications (so-called false positives or false negatives) can have unwanted consequences. To prevent this, some organizations use not one, but multiple human assessments per image. One example is triple verification, in which three reviewers each independently classify the same image. This study examines current verification practices, with a specific focus on whether and under what circumstances triple verification contributes to more accurate and reliable assessments. Legal, technical, and organizational aspects are examined.

Research Method

The research consisted of three phases. In the first phase, fourteen interviews were conducted with experts from the field, including staff from the police, government institutions, and (inter)national hotlines. A broad group of participants was intentionally selected from both the CSAM and TCO domains. These conversations provided insight into current verification practices and offered perspectives on the perceived benefits, drawbacks, and necessity of triple verification for both types of content.

Next, an experiment was explicitly conducted focused on CSAM. In this experiment, professionals assessed a large number of images and videos (2031 items), drawn from a CSAM database. Two conditions were used: in one, reviewers assessed images blindly, without seeing each other's judgments; in the other, they could see previous assessments (non-blind). This allowed the researcher to investigate whether peer influence plays a role and what the added value is of a third reviewer compared to two.

Finally, a third phase involved a group discussion with the reviewers who participated in the experiment. In this conversation, they reflected on the review process, the influence of the conditions, and the manner in which judgments were made.

Key Findings

The comparison between organizations showed that verification processes vary. Three out of five organizations use (or used) triple verification, while others deliberately opt for faster or more flexible alternatives. The way assessments are carried out also differs: in most cases, previous assessments are visible to reviewers or can be inferred from the system. This results in a non-blind or mixed condition, where reviewers may (consciously or not) be influenced by each other's judgments. Only in a few cases is fully blind reviewing applied, where reviewers classify independently without any knowledge of each other's classifications. The level of classification detail also varies: some organizations use a simple three-category system, while others use additional informational tags for features such as estimated gender, hair color, or age. These differences highlight that there is no uniform approach, and context, goals, and system design heavily influence that verification.

Notably, views on triple verification were not unanimously positive. While some viewed it as necessary, others advocated for a more flexible or lenient approach. According to these participants, triple verification is not always proportionate to the complexity of the material, particularly when it is clearly illegal or clearly not. The experiment investigated whether reviewers reached different outcomes depending on whether they could see each other's previous assessments. The difference: in the non-blind condition, where previous assessments were visible, agreement was around 89%. In the blind condition, where reviewers worked independently, the agreement rate was approximately 97%. This indicates that reviewers can support one another in resolving doubts, but that there remains room for independent judgment.

A key remark is that the effect of triple verification was only fully tested in the non-blind condition, because three assessments per item were available there. In those cases, full agreement among all three reviewers occurred in 92% of the items. However, in the blind condition, only two assessments per item were available. Due to technical constraints in the system, it was not possible to test what a third independent vote would have added under those conditions. As a result, the effects of triple verification could not be compared equally across both conditions. This limits the extent to which firm conclusions can be drawn about the added value of a third reviewer in all assessment situations.

The focus group provided more profound insights into how reviewers made their choices. One important theme was the role of image series. If an image was part of a series that included previously flagged illegal material, then even a seemingly neutral image was more likely to be assessed as illegal. This aligns with the legal interpretation set out by the Dutch Supreme Court, which states that an entire series may be classified as CSAM if it contains internal coherence and includes multiple images with an unmistakably sexual nature. In such cases, the more neutral image is judged in the context of the overall series.

Interestingly, pose and framing, such as the posture of the child or the camera angle, were often decisive in the assessment. It was not so much the amount of nudity, but the way the image was presented that determined whether it was perceived as sexual or not. Because these visual elements are more open to subjective interpretation than, for instance, nudity itself, discussions regularly arose among reviewers about how to interpret an image.

Finally, age estimation remained a difficult issue. Reviewers relied on visual features such as body posture and skin texture, but indicated that this remained uncertain, especially in cases with lower image quality or ethnic diversity.

Recommendations:

Tailor the verification process to the organization's context. There is no one-size-fits-all solution. Factors such as content volume, complexity, and the purpose of the hash database play a role.

Match the number of assessments to the complexity of the material. Two assessments may be sufficient for clearly illegal or clearly legal material; ambiguous or complex cases require more reviewers.

Make conscious choices about blind versus non-blind reviewing. Non-blind reviewing promotes convergence, while blind reviewing safeguards independence.

Continue investing in reflection. Organize regular meetings or feedback sessions to discuss areas of uncertainty and adjust practices as needed.

Conclusion

Triple verification proves valuable, especially in situations where carefulness and legal accountability are crucial. At the same time, it is not a one-size-fits-all solution: flexibility is key. The decision to deploy three reviewers depends strongly on the organization's context, including the type of content, scale of operations, and the use of hash databases.

This study forms an exploratory first step in understanding these processes. Further research is needed, especially across national, sectoral, and organizational boundaries. Only through collaboration between governments, companies, NGOs, and other stakeholders can we work toward an internet that is truly clean and safe for everyone. You can read the full thesis here:

<https://repository.tudelft.nl/record/uuid:32f36d04-1f38-4cf4-9296-3eace291f7d1>