



# ATKM deed onderzoek naar de verificatie van kinderpornografisch en terroristisch materiaal ten behoeve van databases

Door Melissa Rottier

## Aanleiding

De verspreiding van kinderpornografisch materiaal (KP) en terroristische online inhoud (TOI) op het internet is een urgent, groot en maatschappelijk probleem.

Om deze beelden snel te kunnen herkennen en verwijderen, maken organisaties gebruik van zogenaamde hash-databases: lijsten met digitale vingerafdrukken van reeds geverifieerd materiaal. Als materiaal overeenkomt met een hash in de database, kan verdere verspreiding van het beeld gestopt worden. Echter kunnen foutieve classificaties (zogenaamde vals positieven of negatieven) ongewenste gevolgen hebben. Om dit te voorkomen hanteren sommige organisaties niet één, maar meerdere menselijke beoordelingen per beeld. Een voorbeeld hiervan is triple verificatie, waarbij drie beoordelaars elk beeld classificeren. Dit onderzoek verkent de huidige werkwijzen rond verificatie van materiaal. Het richt zich specifiek op de vraag of en wanneer triple verificatie bijdraagt aan een zorgvuldigere en betrouwbaardere beoordeling. Hierbij is gekeken naar zowel juridische, technische, als organisatorische aspecten.

# Onderzoeksmethode

Het onderzoek bestaat uit drie fases. In de eerste fase zijn veertien interviews gehouden met experts uit het werkveld, waaronder medewerkers van politie, overheidsinstanties en (inter)nationale meldpunten. Hierbij is bewust gekozen voor een brede groep gesprekspartners die werkzaam zijn op het gebied van zowel KP als TOI. Deze gesprekken boden inzicht in de huidige werkwijzen rond verificatie, en gaven zicht op de ervaren voordelen, nadelen en de noodzaak van triple verificatie voor beide typen content.

Vervolgens is een experiment opgezet dat zich specifiek richt op KP. Hierin beoordeelden professionals een groot aantal afbeeldingen en video's (2031 items), afkomstig uit een database voor KP-detectie. Er is gewerkt met twee condities: in de ene conditie beoordeelden deelnemers de beelden blind, zonder kennis van elkaars oordeel; in de andere conditie zagen zij juist wel de eerdere beoordelingen (niet-blind). Op deze manier is onderzocht of onderlinge beïnvloeding plaatsvindt, en wat de toegevoegde waarde is van een derde beoordelaar ten opzichte van twee.

Tot slot vond een derde fase plaats in de vorm van een groepsgesprek met de beoordelaars die deelnamen aan het experiment. In dit gesprek werd gereflecteerd op het beoordelingsproces, de invloed van de beoordelingscondities, en de manier waarop tot een oordeel werd gekomen.

# Belangrijke bevindingen

Uit de vergelijking tussen organisaties bleek dat verificatieprocessen uiteenlopen. Drie van de vijf organisaties gebruiken (of gebruikten) triple verificatie, terwijl anderen bewust kiezen voor snellere of flexibelere alternatieven. Ook de manier waarop de beoordeling zelf plaatsvindt verschilt: in de meeste gevallen zijn eerdere oordelen van collega's zichtbaar voor beoordelaars, of kunnen deze worden afgeleid via het systeem. Hierdoor is er sprake van een niet-blinde of gemengde stemconditie, waarin beoordelaars (al dan niet bewust) door elkaars oordeel beïnvloed kunnen worden. Slechts in enkele gevallen wordt gewerkt met volledig blinde beoordelingen, waarbij beoordelaars volledig onafhankelijk van elkaar classificeren.

Ten slotte varieert ook de mate van detail in classificatie: sommige organisaties hanteren een eenvoudige verdeling in drie categorieën, terwijl andere organisaties daarnaast gebruikmaken van aanvullende informatietags over kenmerken zoals geschat geslacht, haarkleur, of leeftijd. Deze verschillen onderstrepen dat er geen uniforme aanpak is, en dat verificatie sterk wordt beïnvloed door context, doelen en systeemontwerp.

Opvallend was dat men niet unaniem positief was over de inzet van drie beoordelaars. Waar sommigen het zien als noodzakelijk, pleitte een ander deel juist voor een flexibelere of zelfs lichtere aanpak. Volgens deze participanten staat triple verificatie niet altijd in verhouding tot de complexiteit van het material proportioneel, zeker niet bij beelden die overduidelijk niet strafbaar of juist evident strafbaar zijn. In het experiment werd getest of beoordelaars tot andere uitkomsten komen wanneer ze elkaars eerdere beoordelingen wel of niet kunnen zien. Het verschil: bij niet-blinde beoordeling, waarin de beoordelaars de eerdere oordelen konden inzien, kwamen zij in ongeveer 89% tot overeenstemming. Bij blinde beoordeling, waarin zij onafhankelijk werkten, lag dat percentage op ongeveer 97%. Dit laat zien dat beoordelaars elkaar kunnen helpen om twijfels weg te nemen, maar ook dat er nog steeds ruimte blijft voor zelfstandig oordeel.

Een belangrijke kanttekening is dat het effect van triple verificatie alleen volledig is getest in de niet-blinde conditie, omdat daar drie beoordelingen per beeld beschikbaar waren. In die gevallen was er in 92% van de gevallen sprake van overeenstemming tussen alle drie de beoordelaars. In de blinde conditie waren echter slechts twee beoordelingen per beeld beschikbaar. Vanwege de technische beperkingen in het systeem was het niet mogelijk om in deze conditie te onderzoeken wat een derde, onafhankelijke stem zou hebben toegevoegd. Hierdoor konden de effecten van triple verificatie niet op een gelijke manier tussen de twee condities worden vergeleken. Dit beperkt de mate waarin algemene conclusies getrokken kunnen worden over de toegevoegde waarde van een derde beoordeling in alle beoordelingssituaties.

De focusgroep bood verdiepende inzichten in hoe beoordelaars tot hun keuzes kwamen. Een belangrijk thema was de rol van beeldreeksen: als een beeld onderdeel was van een reeks waarin eerder strafbaar materiaal zat, dan werd ook een op zichzelf neutraal beeld eerder als strafbaar beoordeeld. Dit hing samen met de juridische interpretatie zoals vastgesteld door de Hoge Raad, waarin is bepaald dat een hele reeks als strafbaar kan worden aangemerkt als er binnen die reeks sprake is van onderlinge samenhang en meerdere beelden met een onmiskenbaar seksueel karakter. In dat geval wordt het neutralere beeld gezien als onderdeel van een context die als geheel strafbaar is.

Opmerkelijk was ook dat pose en framing, bijvoorbeeld de houding van het kind of de camerahoek, vaak doorslaggevend waren in de beoordeling. Niet zozeer de hoeveelheid naaktheid, maar de manier waarop het beeld gepresenteerd werd, bepaalde of iets als seksueel of juist niet seksueel werd gezien. Omdat deze visuele elementen vatbaarder zijn voor subjectieve interpretatie dan bijvoorbeeld naaktheid op zich, leidde dit regelmatig tot discussies onder beoordelaars over hoe het beeld geïnterpreteerd moest worden.

Tot slot bleek leeftijdsinschatting een lastig punt. Beoordelaars gebruikten visuele kenmerken zoals lichaamshouding en huidstructuur, maar gaven aan dat dit onzeker bleef, vooral bij lagere beeldkwaliteit of bij etnische diversiteit.

### Aanbevelingen:

Pas verificatie proces aan op de context van de organisatie. Er is niet één juiste manier. De inzet ervan hangt af van factoren zoals het volume van content, het type content (complexiteit), en waarvoor de hash database wordt gebruikt.

Stem het aantal beoordelingen af op de complexiteit. Bij evident strafbaar materiaal kan twee beoordelingen volstaan; ambigue of complexe gevallen vragen om meer ogen.

Maak bewuste keuzes over blind versus niet-blind beoordelen. Niet blind beoordelen bevordert convergentie, maar blinde beoordeling waarborgt onafhankelijkheid.

Blijf investeren in reflectie. Organiseer regelmatig overleg of feedbacksessies om grijze gebieden en afwegingen bespreekbaar te maken en bij te sturen.

## Conclusie

Triple verificatie blijkt waardevol, vooral in situaties waarin zorgvuldigheid en juridische verantwoording cruciaal zijn. Tegelijkertijd is het geen one-size-fits-all oplossing: flexibiliteit staat voorop. De keuze om drie beoordelaars in te zetten hangt sterk af van de context van een organisatie, denk aan het type content, de schaal van werkzaamheden en het gebruik van hash-databases.

Dit onderzoek vormt een eerste verkennende stap in het begrijpen van deze processen. Om tot duurzame en passende werkwijzen te komen, is vervolgonderzoek nodig, met name over de grenzen van landen, sectoren en organisatievormen heen. Alleen door samenwerking tussen overheden, bedrijven, NGO's en anderen betrokkenen kunnen we bouwen aan een internet dat écht schoon en veilig is voor iedereen. Lees hier de gehele scriptie: <https://repository.tudelft.nl/record/uuid:32f36d04-1f38-4cf4-9296-3eace291f7d1>