Technische nota betreffende de EU-verordening over seksueel kindermisbruik (CSA- Child Sexual Abuse Act)

Deze technische nota beschrijft technologieën voor de detectie van gekende en nieuwe beelden van seksueel misbruik van minderjarigen (CSAM of Child Sexual Abuse Material) en grooming, ook in versleutelde of geëncrypteerde omgevingen. In de huidige discussies rond het voorstel tot **EU-verordening** tot het voorkomen en bestrijden van online seksueel misbruik is deze nota bedoeld om bepalingen over het toepassingsgebied te verduidelijken.

De industrie draagt de verantwoordelijkheid om bestaande technologieën in te zetten om de verspreiding van CSAM te voorkomen. Een wetgevend kader is noodzakelijk voor de omkadering van huidige technologieën en de ontwikkeling van nieuwe technologieën met respect voor de privacy van iedereen.

Scantechnologie in functie van het type beelden

Hierna onderscheiden we drie types:

1. Gekende CSAM
2. Nieuwe CSAM
3. Patronen van grooming

Op elk van deze drie types wordt een specifieke scantechnologie toegepast. Indien de beelden zich in een geëncrypteerde of versleutelde omgeving bevinden, zijn bijkomende technische tools nodig om scantechnologieën te kunnen gebruiken.

Gekende CSAM

Waarom is het belangrijk om gekende CSAM op te sporen?

De meerderheid van de beelden die gemeld worden, zijn beelden die al gekend zijn bij de gerechtelijke instanties, maar waarvan niet altijd het slachtoffer geïdentificeerd is. Van de 49.4 miljoen foto's, gemeld aan het Amerikaanse centrum NCMEC (het expertisecentrum voor de behandeling van CSAM), waren 18.8 miljoen nieuw. Voor een slachtoffer is het beeld op zich even traumatisch als het seksueel misbruik zelf, en daarom is het zaak om deze beelden zo snel mogelijk te verwijderen. Het volume is van die orde dat enkel en alleen burgermeldingen of menselijke detectie niet volstaan. Scantechnologie past een eerste filter toe, waardoor het volume behapbaar wordt voor menselijke review en de capaciteit van de gerechtelijke instanties efficiënt kan worden ingezet.

Scantechnologie voor gekende CSAM?

File hashing is een technologie om een digitale afdruk te maken van een bestand door het bestand te ontdoen van kleur en vorm. Hierdoor is de technologie in staat om alle identieke bestanden op te sporen.

Perceptual hashing gaat nog een stapje verder. Het is een technologie waarbij beelden worden omgezet in een raster, elk raster in het vierkant is een cijfermatige berekening, resulterend in verschillende digitale afdrukken. De afdrukken worden vergeleken tegenover de afdrukken van andere beelden. In tegenstelling tot de file hash, laat deze techniek toe om licht bijgesneden beelden te detecteren.

Binnen beide hashing functies zijn er verschillende types.

Hoe effectief zijn deze technologieën?
Beide technologieën zijn al 15 jaar op de markt en zijn zeer effectief. Child Focus werkt met fotoDna, een type van *Perceptual hash* binnen het project Arachnid. Project Arachnid spoort gekende CSAM op het internet aan de hand van crawlers. De Arachnid crawlers hebben al 160 miljard beelden kunnen detecteren op het net. Beelden die licht gewijzigd zijn ten opzichte van het oorspronkelijke beeld worden geverifieerd door analisten. Dit zorgt ervoor dat het aantal vals positieven beperkt wordt.

Wat zijn vals positieven?
We spreken van een vals positief wanneer scantechnologie een beeld onterecht labelt als CSAM. Vergelijk het met een verkeerscamera. Als de camera flitst bij een bestuurder die de toegelaten snelheid rijdt, dan is dat een vals positief.

Wat wordt er bedoeld met een aanvaardbaar aantal positieven?
Een betrouwbare technologie genereert in het beste geval zo weinig mogelijk vals positieven. '0' is het ultieme doel, maar onbereikbaar zelfs voor virusscanners en spamfilters. Wanneer de betrouwbaarheid op 99.9% wordt vastgelegd, zal 1 op de 1000 beelden onterecht worden weerhouden. Dit betekent wel dat de detectie van het aantal werkelijke CSAM lager zal liggen (*recall*). In het geval van 99.9 % betrouwbaarheid, is de *recall* 84%. Je kan ook de betrouwbaarheid verlagen tot 99%, waardoor je én een aanvaardbaar aantal vals positieven verkrijgt én een *recall* van 94%.

Hoe kan het platform de betrouwbaarheid verhogen?
De betrouwbaarheid van een *hashing* tool kan verhoogd worden door:

- Adequate content moderatie
- Kwaliteitsvolle hashes
- Betrouwbaarheid afstemmen op het type platform
- Regelmatig bijsturen van de hashingtool

Hoe kan het wettelijk kader bijdragen?
In het voorliggend voorstel tot een verordening tot het voorkomen en bestrijden van CSAM is het EU expertisecentrum verantwoordelijk voor de identificatie van *state-of-the-art* technologie. In deze hoedanigheid kan het zelf de benchmark van betrouwbaarheid voor scantechnologie bepalen. Ook is de oprichting van een technisch comité aangewezen, die scantechnologie test op datasets om de betrouwbaarheid en *recall* in verschillende omgevingen te bepalen.

Nieuwe CSAM

Waarom is het belangrijk om nieuwe CSAM op te sporen?
Het is belangrijk om nieuwe CSAM zo snel mogelijk op te sporen om slachtoffers, die acuut in gevaar zijn, zo snel mogelijk te identificeren en te beschermen, daders te berechten en om de exponentiële verspreiding van die beelden tegen te gaan.

Scantechnologie voor nieuwe CSAM?
Image classifiers zijn algoritmes die op basis van *machine learning* gegevens automatisch in categorieën indelen.

Hoe effectief is deze technologie?
De technologie bestaat ongeveer 5 jaar. De training van de scantechnologie gebeurt op basis van datasets van onschuldige beelden, volwassen pornografie en gekende beelden van seksueel misbruik. De *classifier* kan geen beelden analyseren of herkennen. Het werkt zoals een beslissingsboom die beelden indeelt naargelang de aanwezigheid van vastgelegde criteria.

Patronen van grooming

Waarom is het belangrijk om patronen van grooming op te sporen?
Meer en meer kinderen zijn het slachtoffer van online grooming. Grooming staat voor het proces waarbij een volwassene doelbewust minderjarigen benadert en manipuleert voor seksuele doeleinden. Bij *content driven grooming* is de groomer uit op intiem beeldmateriaal van de minderjarige. In 2022 waren er in de Europese Unie 4000 meldingen van grooming.

Scantechnologie voor grooming?
Regular expression rules is een van de meest voorkomende methoden, gebruikt door de industrie. Deze tekstregels of expressies, vergelijkbaar met trefwoorden, worden vooraf bepaald door mensen en ingegeven in een programma, opdat de computer leert deze automatisch te herkennen.

Grooming classifier is een meer geavanceerde technologie, dat gebruik maakt van taalmodellen en een tekst *classifier*. Dankzij *deep learning* technieken kan men grooming gedrag voorspellen, zonder een exacte beschrijving te hebben van het gedrag in een specifieke context.

Hoe effectief is deze technologie?
De *grooming classifier* bevindt zich momenteel in een onderzoeks- en ontwikkelingsfase, maar de eerste feedback van gebruik door ngo's is positief.

Scantechnologie in versleutelde omgevingen

Wat is encryptie?
Encryptie of versleuteling is de codificatie van een transactie. Je hebt een sleutel nodig om de code te kunnen kraken. Er bestaan vandaag heel wat toepassingen die een vorm van encryptie gebruiken: bankverrichtingen, e-mailuitwisselingen, chats,… De hoogste norm in de encryptiewereld is E2EE of end-to-end versleuteling, omdat verstuurde berichten en oproepen van zender tot ontvanger beveiligd zijn, waardoor derden geen toegang hebben tot de inhoud.

Wat is het belang van encryptie?
Door encryptie kunnen gegevens niet gestolen of bewerkt worden, met als belangrijkste voordeel de bescherming van onze privacy.

Wat betekent E2EE op vlak van de strijd tegen online seksueel misbruik?
E2EE maakt het moeilijk om te detecteren op gekende en nieuwe CSAM. In tegenstelling tot wat de privacy voorvechters beweren, zijn de reeds bestaande oplossingen geen aanfluiting van de privacy van iedereen.

Wat kunnen de digitale platformen ondernemen?
Digitale platformen kunnen bestaande oplossingen toepassen en innoveren om de spreiding van CSAM tegen te gaan én tegelijkertijd privacy garanderen.

Welke oplossingen zijn er vandaag al?
Er is niet één oplossing voor de detectie van beelden van seksueel misbruik in versleutelde omgevingen. De online platforms kunnen verschillende methoden combineren:

1. Scannen op het toestel of *client-side scanning* voor encryptie
   Hierdoor kan gekende CSAM nog voor de versleuteling worden gedetecteerd. Deze oplossing is privacy-bestendig en heeft geen invloed op de prestaties van het toestel.

2. *Secure enclave* of voor het bedrijf toegankelijke encryptie
   Scantechnologie wordt ingezet tijdens de transmissie van inhoud. Dit noodzaakt een tijdelijke decryptie van de berichten. Bij deze oplossing worden bedrijven zelf geacht hun versleutelde omgevingen te scannen op beelden van seksueel misbruik.

3. *Homomorfe versleuteling*
   Homomorfe encryptie maakt het mogelijk om complexe wiskundige bewerkingen uit te voeren op versleutelde gegevens zonder de versleuteling aan te tasten. Deze oplossing kan gebruikt worden voor gekend en nieuw CSAM. De implementatie van deze technologie op schaal moet uitgewerkt worden maar met budgetten van de big tech kan dat geen probleem zijn.

Hoe kan het wettelijk kader bijdragen tot oplossingen?
Door detectie op gekende en nieuwe CSAM te reguleren, stimuleren we verder technologisch onderzoek en vooruitgang. Een omkaderd beleid is de garantie dat deze scanmethoden de vereiste waarborgen inbouwen om de privacy van iedereen te beschermen.

Contacten en bronnen

- Dr. Hany Farid (Professor aan de Universiteit van Berkeley, Californië en ontwikkelaar van fotoDna)
  - Video by Dr Hany Farid over hoe scantechnologieën voor beelden van seksueel misbruik werken, ook in versleutelde omgevingen.
- Emily Slifer, Director of Policy at Thorn (https://www.thorn.org/)
- Arachnid (https://www.projectarachnid.ca/en/)