Het welzijn van de kinderen, zienswijze Bert Hubert ten behoeve van rondetafelgesprek Client Side Scanning

Als achtergrond, ik heb de afgelopen 20 jaar veelvuldig bijgedragen aan overleggen tussen overheid en de internetwereld over hoe kinderporno bestreden kan worden. Ook leverde mijn bedrijf software aan opsporingsdiensten voor dit doeleinde. Ik ben tevens van mening dat de Nederlandse hostingindustrie nu veel te gemakkelijk wegkomt met het huisvesten van zeer ongewenste content.

Andere position papers voor dit gesprek hebben al hun afgrijzen uitgesproken over het Client Side Scanning voorstel en ik sluit me daar volledig bij aan. Het idee dat we alle communicatie van *alle* Europeanen automatisch gaan scannen en de bevindingen live doorsturen naar de EU is verwerpelijk.

Met deze voorgenomen wetgeving heeft u (Kamerleden, kabinet) de kans om deel te nemen aan "het moment dat het allemaal mis ging" met onze rechtsstaat, of u juist in te zetten voor wijzere regelgeving.

Ik ben gevraagd in te gaan op de cybersecurity aspecten van de wetgeving, waaronder ik ook vat hoe het dan precies werkt. Mijn voornaamste focus is hierbij op het gestelde doel: het welzijn van de kinderen, wordt dit netto beter of niet.

Achtergrond wetsvoorstel

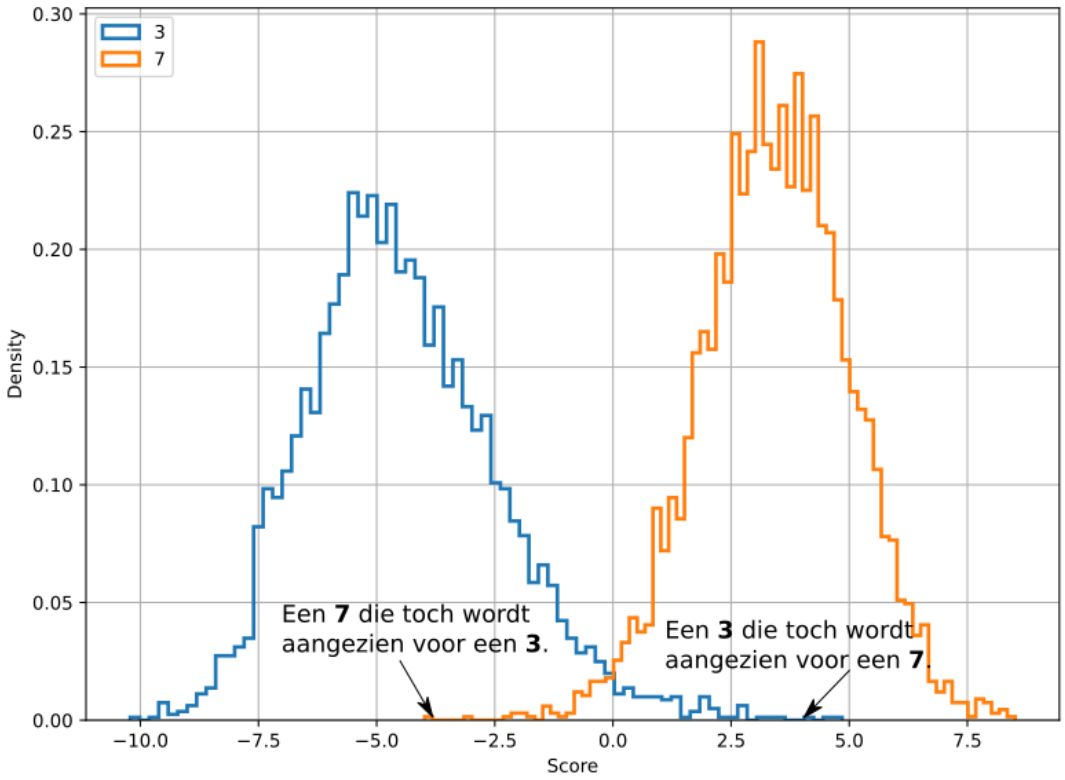
Het voorstel omvat veel verschillende dingen (sommige goed), maar deze rondetafelbijeenkomst beperkt zich tot het *opsporingsbevel*, dat uitgevaardigd kan worden tegen aanbieders van "interpersoonlijke communicatiediensten".

Deze "detection order" houdt in dat aanbieders van chatdiensten (Signal, WhatsApp, Facebook, iMessage/Apple, Snapchat, Telegram etc) verplicht worden om software te installeren die:

"doeltreffend [is] voor het opsporen van de verspreiding van bekend of nieuw materiaal van seksueel misbruik van kinderen dan wel het benaderen van kinderen"

Dit gaat dus om drie dingen – reeds bekend pornografisch materiaal, nog onbekend materiaal en het "groomen" van kinderen. Bij dit laatste gaat het om gesprekken, zowel als tekst of als spraak.

Voor dit soort detectie gebruikt de computerwereld zogeheten 'classifiers', mogelijk het best bekend van de spamfilter. Een stuk informatie wordt bekeken, en een algoritme kent een score toe. Als de score boven een bepaalde waarde is zeggen we dat iets bijvoorbeeld spam is.



Een tweezijdige classifier van handgeschreven cijfers 3 en 7. Een score boven de 2,5 is bijna zeker een 7. Een score onder -2,5 is bijna zeker een 3. Daartussen is twijfel. Noteer bij de pijltjes cijfers die verkeerd geklassificeerd zijn!

Er is hierbij altijd een keuze - als de grenswaarde voor de score te hoog is glipt er te veel spam door de filter. Maar, bij een te lage grens komt er te veel belangrijke mail in de spamfolder.

Iedere mailprovider maakt hier een afweging, en we hebben allemaal weleens meegemaakt dat die afweging niet goed was: echte mail in de spamfolder, en toch spam in de inbox. Fundamenteel is daar ook heel weinig aan te doen, 99% correct is al heel knap, 99,9% is vreselijk veel moeilijker.

Details

Aangenomen moet worden dat alle grote aanbieders van chat en bel-diensten een opsporings-bevel zullen krijgen. Dit houdt in dat classifiers op onze telefoons en computers geïnstalleerd zullen worden door Facebook/Meta/WhatsApp, Snapchat etc.

Als de classifier een te hoge ‘Kind exploitatie-score’ berekent van een bericht, foto of video gaat er automatisch een kopie naar het zogeheten EU-centrum, een club te vestigen hier in Den Haag in of naast het gebouw van Europol.

Bij het EU-centrum bekijkt een EU ambtenaar de foto/video of het bericht vervolgens. Mocht de melding “evident onjuist zijn” dan stopt het proces daar. Maar als er ook maar enige twijfel

is stuurt de ambtenaar de melding door naar Europol EN een lokaal politiebureau nabij de persoon van wie de foto of bericht kwam. De locatiegegevens van de gebruiker werden eerder meegestuurd om dit makkelijk te maken.

Bij het lokale politiebureau zal een agent de foto, video of het bericht evalueren, en besluiten of er onderzoek ingesteld moet worden.

Overigens kan dit politiebureau dus ook heel goed de politie van het dorp van je Griekse vakantiebestemming zijn.

Het EU-centrum legt ondertussen een database aan van al deze bestanden, voor onderzoek naar trends en technologieën. Deze database omvat ook alles wat uiteindelijk *niet* kind seksueel-gerelateerd bleek. Deze data wordt ook beschikbaar gesteld aan Europol voor verder onderzoek.

“False positives”

Spamfilters hebben het al zwaar, maar zelfs mensen hebben het moeilijk om de aard van een foto te bepalen. Het Nederlands Forensisch Instituut weigert bijvoorbeeld leeftijdsbepaling te doen aan de hand van een foto.

De vraag is nu hoe goed de classifiers zullen zijn. De EU maakt zich er vanaf door in artikel 10.3 lid d te stellen dat deze “voldoende betrouwbaar zijn”, en “het foutenpercentage zo veel mogelijk beperken”. Maar hoop is geen plan.

Ik ga hier niet in detail treden, maar er worden nogal wat foto’s en gesprekken met volledig ‘consent’ online uitgewisseld tussen oudere kinderen en (jong-)volwassenen, waarvan je kunt vermoeden dat het algoritme in de classifier in de war raakt. En dat leidt dan tot automatische meldingen.

Het is ook goed om te weten dat veel telefoons van kinderen en jong-volwassenen geregistreerd staan op de naam van de ouders. Een chatbericht kan daarom lijken te komen van een man van 45, terwijl het is ingetypt door een meisje van 13. Is een heel andere context.

Het welzijn van het kind

In de gesprekken vallen we vaak terug op “denk aan de kinderen”, dus laten we dat doen. Correct gedetecteerde verspreiding van seksueel materiaal van kinderen, of grooming kan inderdaad een bijdrage zijn aan kinderwelzijn. Mogelijk kunnen vreselijke dingen voorkomen worden. Laten we dat nooit vergeten.

Maar een *onterechte* melding is niet schadeloos, en wel om twee redenen.

De onterechte melding leidt mogelijk tot veel ellende bij betrokkenen. Er komen onderzoeken, telefoons worden leeggetrokken, ouders worden verdachten, iedereen wordt met de nek aangekeken. Veilig Thuis komt over de vloer om onderzoek te doen. Dit ontwricht hele gezinnen, mogelijk met thuissituaties die dit er niet bij kunnen hebben. Als we geluk hebben concludeert men al binnen een paar maanden (!) dat de onterechte melding echt onterecht was.

Daarnaast is helaas veelvuldig gebleken dat ook onterechte meldingen kunnen leiden tot veroordelingen, die soms pas na vele jaren hoger beroep teruggedraaid worden. Betrokken mensen kunnen de schijn tegen zich hebben, of al meer problemen hebben waarbij dit de druppel is.

Je brandt overigens zo al je “zeven vinkjes” op aan dit soort onderzoeken. Het risico op een onterechte uitslag is veel hoger als je geen traditioneel wit welvarend gezin bent.

Ook is het zo dat zelfs afgehandelde meldingen leiden tot blijvende sporen in databases. Een andere vreselijke uitkomst is dat er geen officieel oordeel wordt geveld en iemand eindeloos blijft hangen in een schaduwtoestand ‘niet bewezen/niet weerlegd’. Dit alleen al kan je leven ruïneren.

Dit alles doet helemaal niets goeds voor het welzijn van kinderen.

We moeten daarom goed kijken naar de som - hoeveel leed wordt voorkomen door al onze communicatie te scannen, en hoeveel leed wordt veroorzaakt door onterechte detecties, of het permanent ‘vlaggen’ van onbewezen gevallen.

Actief misbruik

De classifier codes worden gedwongen geïnstalleerd op onze telefoons en computers. Dit betekent dat kwaadwillenden kennis kunnen nemen van de gebruikte algoritmes. En met die kennis kunnen onschuldige materialen gemaakt worden die toch worden aangewezen als gevaarlijk.

Het wordt zo mogelijk om mensen leuke kattenfoto’s te sturen die toch leiden tot meldingen en onderzoek. Nou zal dit meestal wel goed aflopen, maar het is toch een manier om kwetsbare mensen onder de aandacht van de politie te krijgen.

Ook is het zo dat Europol en het EU-centrum dus een steeds grotere berg foto’s opbouwen, inclusief uiteindelijk niet fout bovenden materiaal. Iedere EU instantie is de afgelopen tien jaar weleens gehacked. Die database en de foto’s komen dus ooit op straat te liggen, met de bijbehorende de namen en locaties van iedereen.

De classifier software zelf

Op al onze telefoons komen verplicht classifiers, maar waar komen die vandaan? De EU zelf zegt deze beschikbaar te gaan stellen. Berichten in de pers lijken erop te wijzen dat deze software via de EU geleverd kan worden door een bedrijf van de acteur Ashton Kutcher, tevens oprichter van een non-profit startup genaamd Thorn, die zich inzet voor de bescherming voor kinderen. Follow the Money publiceert nuttig onderzoek over dit thema.

Ook zouden Meta, Apple, Snap etc zelf met software kunnen komen.

Recent zijn de Apple en Google ecosystemen slachtoffer geworden van een lek in een stuk software gebruikt om plaatjes en foto’s mee te verwerken. Hierdoor moesten alle Apple en Android telefoons geüpdate worden, om te voorkomen dat ze gehacked zouden worden.

Deze zelfde software zal ook door classifiers gebruikt worden. Dit geeft aan dat deze classifiers zelf ook een risico vormen voor de veiligheid van gebruikers: het is weer meer code met kans op kwetsbaarheden. Juist omdat deze software al onze berichten, foto’s en filmpjes verplicht moet scannen is dit een groot risico.

De balans

De overgrote meerderheid van onze communicatie is niet bedreigend voor kinderen. De allerbeste classifiers die we kennen zitten er toch makkelijk 0,1% van de gevallen naast. Dit betekent dat het EU-centrum, Europol en lokale politiebureaus een stortvloed aan onjuiste meldingen zullen ontvangen. Want 0,1% is veel meer dan hoeveel seksueel materiaal over kinderen er verspreid wordt.

De medewerkers op kantoor bij het EU-centrum en Europol zitten dan voor hun werk de hele dag foto’s en filmpjes van blote mensen te bekijken en oordelen te vellen, en we moeten maar hopen dat ze dat goed doen. Het lijkt me overigens ook een zeer belastende baan, en ik heb rechercheurs meegemaakt die helemaal opgebrand zijn door dit soort werk.

Maar, alles hangt er dus vanaf hoe goed de classifiers zijn, en wat de kwaliteit is van de afhandeling van onterechte meldingen.

Op beide fronten zijn de tekenen niet goed. Cruciaal, beide onderwerpen worden in het voorstel nauwelijks concreet besproken, behalve door vast te leggen dat de technologie gewoon geen fouten moet maken. Maar in eerdere gevallen hebben we gezien dat overheidsalgoritmes er wel degelijk naast kunnen zitten, met alle gevolgen (en parlementaire enquêtes) van dien.

Ook is het zorgelijk dat de EU een enorme database opbouwt van foto’s, filmpjes, chatgesprekken en locaties, of die nou zorgwekkend waren of niet. Iedere EU instantie is al eens gehacked.

Afsluitend is er nog het cybersecurity risico - alle software zal kwetsbaarder worden voor aanvallen, en er is ook nog de mogelijkheid dat subtiel bewerkte onschuldige content zal leiden tot meldingen.

Wat moet er dan wel?

Alles heeft voor- en nadelen, maar Apple kent bijvoorbeeld een ‘familieprogramma’, waarbij telefoons van kinderen onderdeel zijn van je Applefamilie. De telefoons van kinderen kunnen zo al beperkt toegang krijgen tot internet, of de gebruiker beschermen tegen naaktfoto’s.

Zolang we nog geloven in gezinnen zou het geen kwaad idee zijn om telefoons dan zo in te richten dat ze gevaarlijk gedrag/ervaringen van kinderen melden bij hun eigen ouders in plaats van bij Europol.

Afsluitend - de EU heeft gelijk dat het Internet geen vrijplaats moet zijn voor seksueel misbruik van kinderen. Met voldoende nadenken en specifiekere regelgeving zou het mogelijk moeten zijn dit te bereiken zonder al onze prive-communicatie via nog onbewezen software direct naar overheden te sturen.