Note technique relative à la résolution sur les abus sexuels d'enfants (CSA- Child Sexual Abuse Act)

La présente note technique décrit les technologies de détection d'images connues et de nouvelles images d'abus sexuels commis sur des enfants (CSAM ou Child Sexual Abuse Material) et de "grooming", y compris dans des environnements cryptés ou chiffrés. Dans le cadre des discussions actuelles sur la **proposition de résolution de l'UE** visant à prévenir et à combattre les abus sexuels en ligne, cette note vise à clarifier les dispositions relatives au champ d'application.

Il incombe à l'industrie de déployer les technologies existantes pour empêcher la propagation des abus sexuels en ligne. Un cadre législatif est nécessaire pour encadrer les technologies actuelles et en développer de nouvelles tout en respectant la vie privée de chacun.

Technologies de scanning en fonction du type d'images

Nous distinguons ci-dessous trois types d'images :

1. CSAM connu
2. Nouveau CSAM
3. Processus de manipulation psychologique (grooming)

Une technologie de scanning spécifique est appliquée à chacun de ces trois types d'images. Si les images se trouvent dans un environnement crypté ou chiffré, des outils techniques supplémentaires sont nécessaires pour scanner les images.

1. CSAM connu

Pourquoi est-il important de détecter les images de CSAM connus ?

La majorité des images signalées sont des images déjà connues des instances judicières, mais où la victime n'est pas toujours identifiée. Sur les 49,4 millions d'images signalées au centre américain NCMEC (le centre d'expertise pour le traitement du CSAM), 18,8 millions d'images étaient nouvelles. Pour une victime, l'image en elle-même est aussi traumatisante que l'abus sexuel lui-même, c'est pourquoi il est important de retirer ces images le plus rapidement possible. Le volume est tel que les simples signalements des citoyens ou la détection humaine ne suffisent pas. La technologie de scanning applique un premier filtre, ce qui rend le volume gérable pour une analyse complémentaire manuelle et permet ainsi une utilisation plus efficace des ressources judiciaires.

Technologie de scanning pour les images CSAM connues ?

Le File hashing est une technologie qui permet de créer une impression numérique d'un fichier en le dépouillant de ses couleurs et de sa forme. Cela permet à la technologie de détecter tous les fichiers identiques.

Le Perceptual hashing va encore plus loin. Il s'agit d'une technologie dans laquelle les images sont converties en une grille, chaque case dans la grille étant un calcul numérique, ce qui donne lieu à plusieurs impressions numériques. Ces empreintes sont comparées à celles d'autres images.

Contrairement au *File hashing*, cette technique permet de détecter par exemple des images recadrées.

Dans chacune des deux technologies de hashing il existe différents types.

Quelle est l'efficacité de ces technologies ?
Les deux technologies sont sur le marché depuis 15 ans et sont très efficaces. Child Focus travaille avec photoDna, un type *perceptual hahing*, dans le cadre du projet Arachnid. Le projet Arachnid traque les CSAM connus sur Internet à l'aide de *crawlers*. Les robots Arachnid ont déjà été en mesure de détecter 160 milliards d'images sur le net. Les images légèrement modifiées par rapport à l'image originale sont vérifiées par des analystes. Cela permet de limiter le nombre de faux positifs.

Qu'est-ce qu'un faux positif ?
On parle de faux positif lorsque la technologie de scanning qualifie à tort une image de CSAM. Comparons cette situation à celle d'une caméra de surveillance de la circulation. Si la caméra flashe un conducteur roulant à la vitesse autorisée, il s'agit d'un faux positif.

Qu'entend-on par un nombre acceptable de faux positifs ?
Dans le meilleur des cas, une technologie fiable génère le moins de faux positifs possible. L'objectif ultime est "0", mais il est impossible à atteindre, même pour les scanners de virus et les filtres anti-spam. Lorsque la fiabilité est fixée à 99,9 %, 1 image sur 1 000 sera retenue à tort.
Cela signifie que le taux de détection du CSAM réel sera inférieur (recall). Dans le cas d'une fiabilité de 99,9 %, le taux de recall est de 84 %. Il est également possible d'abaisser la fiabilité à 99 %, ce qui permet d'obtenir à la fois un nombre acceptable de faux positifs et un taux de *recall* de 94 %.

Comment un plate-forme peut-elle accroître la fiabilité ?
La fiabilité d'un outil de hashing peut être améliorée par les moyens suivants :

- une modération adéquate du contenu
- un hashing de qualité
- adapter la fiabilité en fonction du type de plateforme
- ajustement régulier de l'outil de hashing

Quelle peut être la contribution du cadre juridique ?
Dans la présente proposition de résolution visant à prévenir et à combattre le CSAM, le Centre d'Expertise de l'UE sera chargé d'identifier les technologies de pointe appropriées. À ce titre, il peut lui-même fixer les critères de fiabilité de la technologie de scanning. La création d'un comité technique est également recommandée, qui testera la technologie de scanning sur des ensembles de données (datasets) afin de déterminer le niveau de fiabilité et de recall dans différents environnements.

2. Nouveau CSAM

Pourquoi est-il important de détecter les nouvelles images d'abus ?

Il est important de détecter le plus rapidement possible le nouveau CSAM afin d'identifier et de protéger le plus tôt possible les victimes qui sont gravement menacées, de traduire les auteurs en justice et de mettre un terme à la diffusion exponentielle de ces images.

Technologies de scanning pour des nouvelles images de CSAM?

Les classificateurs d'images (*Image classifiers*) sont des algorithmes qui classent automatiquement les données sur base de machine learning.

Quelle est l'efficacité de cette technologie ?

La technologie existe depuis environ 5 ans. L'apprentissage de la technologie de scanning est basé sur des ensembles de données d'images innocentes, de pornographie adulte et d'images connues d'abus sexuels.

Le classificateur ne peut pas analyser ou reconnaître les images. Il fonctionne comme un arbre de décision qui classe les images en fonction de la présence de critères définis.

3. Processus de manipulation psychologique (grooming)

Pourquoi est-il important de détecter les schémas de manipulation psychologique ?

De plus en plus d'enfants sont victimes de "grooming" en ligne. Le "grooming" est le processus par lequel un adulte approche et manipule délibérément des mineurs à des fins sexuelles. Dans le cas du "grooming" axé sur le contenu, le "groomer" cherche à obtenir des images intimes du mineur. En 2022, 4 000 cas de "grooming" ont été signalés dans l'Union européenne.

Technologies en vue de scanner des processus de grooming ?

Regular expression rules et l'une des méthodes les plus courantes utilisées par le secteur. Ces lignes de texte ou expressions, semblables à des mots-clés, sont prédéterminées par des humains et introduites dans un programme afin que l'ordinateur apprenne à les reconnaître automatiquement.

Grooming classifier est une technologie plus avancée, qui utilise des modèles de langage et un classificateur de texte.
Les techniques d'apprentissage en profondeur permettent de prédire le comportement de grooming sans disposer d'une description exacte du comportement dans un contexte spécifique.

Quelle est l'efficacité de cette technologie ?

Le classificateur de toilettage est actuellement en phase de recherche et de développement, mais les premiers retours d'expérience des ONG sont positifs.

Technologies de scanning dans des environnements cryptés

Qu'est-ce que le cryptage ?

Le cryptage ou chiffrement est la codification d'une transaction. Il faut une clé pour déchiffrer le code. De nombreuses applications utilisent aujourd'hui une forme ou une autre de cryptage : transactions bancaires, échanges de courriers électroniques, chats,... La norme la plus élevée dans le monde du cryptage est le cryptage E2EE ou cryptage de bout en bout, car les messages et les appels envoyés sont sécurisés de l'expéditeur au destinataire, empêchant ainsi les tiers d'accéder au contenu.

Quelle est l'importance du cryptage ?

Le cryptage empêche le vol ou la manipulation des données, avec comme principal avantage la protection de notre vie privée.

Que signifie l'E2EE en termes de lutte contre les abus sexuels en ligne ?

L'E2EE rend la détection de CSAM connus et nouveaux difficile. Contrairement à ce que prétendent les défenseurs de la vie privée, les solutions déjà en place ne vont pas à l'encontre la vie privée de chacun.

Que peuvent faire les plateformes numériques ?

Les plateformes numériques peuvent appliquer et innover les solutions existantes pour lutter contre la propagation des CSAM tout en garantissant le respect de la vie privée.

Quelles sont les solutions existantes aujourd'hui ?

Il n'existe pas de solution unique pour détecter les images d'abus sexuels dans les environnements cryptés. Les plateformes en ligne peuvent combiner différentes méthodes :

1. Analyse sur l'appareil ou analyse côté client avant le cryptage (*client side scanning*)
   Cette méthode permet de détecter les images d'abus sexuels connus avant même le cryptage. Cette solution est respectueuse de la vie privée et n'affecte pas les performances de l'appareil.

2. *Secure enclave* ou encryptage accessible à l'entreprise
   La technologie d'analyse est déployée pendant la transmission du contenu. Cela nécessite un décryptage temporaire des messages. Avec cette solution, on attend des entreprises qu'elles analysent elles-mêmes leurs environnements cryptés à la recherche d'images d'abus sexuels.

3. *Chiffrement homomorphe*
   Le chiffrement homomorphe permet d'effectuer des opérations mathématiques complexes sur des données chiffrées sans affecter le chiffrement. Cette solution peut être utilisée pour les CSAM connus et nouveaux. L'implémentation de cette technologie à grande échelle doit être élaborée, mais avec les budgets des grandes entreprises technologiques, cela ne devrait pas poser de problème.

Comment le cadre juridique peut-il contribuer à trouver des solutions ?
En réglementant la détection des CSAM connus et nouveaux, nous encourageons la poursuite de la recherche et du progrès technologiques. Une politique cadrée est la garantie que ces méthodes de scanning intègrent les garanties nécessaires pour protéger la vie privée de chacun.

Contacts et sources

- Dr. Hany Farid (Professeur à l'Université de Berkeley, Californie et développeur de fotoDna)
    - Video by Dr Hany Farid sur le fonctionnement des technologies de recherche d'images d'abus sexuels, y compris dans des environnements cryptés
- Emily Slifer, Director of Policy at Thorn (https://www.thorn.org/)
- Arachnid (https://www.projectarachnid.ca/en/)