FAQ: Client-Side-Scanning im Kontext des Verordnungsentwurfs zur Prävention und Bekämpfung des sexuellen Missbrauchs von Kindern

In diesen FAQ fassen wir die wichtigsten Punkte eines offenen Briefs zusammen, der von mehr als 700 Fachwissenschaftler:innen aus aller Welt unterzeichnet wurde und sich mit den Problemen des Kompromissvorschlags der dänischen Ratspräsidentschaft für die Verordnung zur Prävention und Bekämpfung von sexuellem Kindesmissbrauch vom Juli 2025 befasst.

Prof. Anja Lehmann (HPI and University of Potsdam, Germany), Dr. Anne Canteaut (Inria, France), Prof. Aurélien Francillon (EURECOM, France), Prof. Bart Preneel (KU Leuven, Belgium), Prof. Carmela Troncoso (Max Planck Institute for Security and Privacy, Germany), Prof. Cas Cremers (CISPA Helmholtz Center for Information Security, Germany), Prof. Christian Makenz (Technical University of Denmark, Denmark), Dr. Gaëtan Leurent (Inria, France) Prof Thorsten Holz (Max Planck Institute for Security and Privacy, Germany)

Was ist Client-Side-Scanning im Kontext der Verordnung zur Prävention und Bekämpfung von sexuellem Kindesmissbrauch?

Client-Side-Scanning ist ein Verfahren, das Nachrichten (Text, Bilder, Video, Audio, Links) auf Endgeräten von Nutzer:innen nach schädlichen Inhalten durchsucht und überprüft. Im Kontext der Verordnung zur Prävention und Bekämpfung von sexuellem Kindesmissbrauch (CSA-Verordnung) handelt es sich bei diesen schädlichen Inhalten um Material im Zusammenhang mit sexuellem Kindesmissbrauch. Im Falle einer Einstufung von Inhalten als Darstellungen sexualisierter Gewalt gegen Kinder sieht der Verordnungsentwurf bzw. der dänische Vorschlag, die Benachrichtigung relevanter Behörden und die entsprechende Ausleitung des Materials vor.

Wenn Nachrichten über einen verschlüsselten Messenger-Dienst (wie WhatsApp, iCloud oder Signal) gesendet werden, analysiert das clientseitige Scannen die Nachrichten, bevor sie verschlüsselt werden. Wird der Inhalt der Nachricht als schädlich eingestuft, wird die Nachricht weitergeleitet. Ist dies nicht der Fall, wird die Nachricht verschlüsselt und an den Empfänger gesendet.

Im Laufe der Zeit hat sich der Geltungsbereich der Verordnung hinsichtlich der verschiedenen Arten von zu scannenden Inhalten geändert, wobei sich die neueste Version vom Juli 2025 ausschließlich auf Links und Bilder konzentriert. Diese FAQ gilt für alle Arten von Inhalten.

Wie funktionieren CSAM-Detektoren? Sind sie zuverlässig?

Technische Verfahren, um Darstellungen sexualisierter Gewalt gegen Kinder (child sexual abuse material, CSAM) zu erkennen, werden je nach dem Material, auf das sie abzielen, auf eine von drei Arten implementiert.

Erkennung bekannter Versionen von bekannten Missbrauchsdarstellungen. Detektoren können exakte Kopien von Bildern identifizieren, die bereits zuvor gemeldet als CSAM bestätigt wurden. Beispiele sind Kopien von Missbrauchsdarstellungen, die ohne Änderungen vorzunehmen, geteilt werden, oder ein Link zu CSAM-Material.

Um diese Identifizierung durchzuführen, nutzen Detektoren sogenannte kryptografische Hash-Funktionen. Eine Hash-Funktion bildet eine Eingabemenge (z. B. ein Bild oder einen Link) als kleinere Zielmenge, einer Zeichenfolge ab. Eine solche Zeichenfolge funktioniert ähnlich wie ein Fingerabdruck und wird Hashwert genannt. Die Hash-Funktion stellt sicher, dass die Zuordnung nahezu eins zu eins ist (d. h., keine zwei Bilder werden demselben Hashwert zugeordnet) und dass man anhand des Hashwerts die ursprüngliche Eingabe (das Bild oder den Link) nicht wiederherstellen kann.

Um Kopien bekannter CSAM anhand von Hashwerten zu erkennen, müssten Detektoren über eine Liste von Hashwerten verfügen, die den bekannten Missbrauchsdarstellungen entsprechen. Detektoren würden den Hashwert einer Nachricht berechnen und prüfen, ob dieser Hashwert in der Liste bekannter Missbrauchsdarstellungen enthalten ist. Ist dies der Fall, kann der Detektor sicher sein, dass es sich bei der Nachricht um eine Kopie von bekanntem CSAM handelt. Die Weitergabe dieses Hashwerts an Strafverfolgungsbehörden bedeutet, dass ein Dritter den Inhalt der Nachricht kennt, wodurch die Ende-zu-Ende-Verschlüsselung des Messengers außer Kraft gesetzt wird.

Solche Detektoren können jedoch nur exakte Kopien erkennen. Jede Änderung des Inhalts, z. B. Zuschneiden, Drehen, Größenänderung, Hinzufügen einer Linie, Ändern der Farbe eines Pixels oder die von allen kommerziellen Messengern zur Bandbreiteneinsparung verwendete Komprimierung, würde die Erkennung fehlschlagen lassen. Daher ist dies im Zusammenhang mit der Erkennung von Missbrauchsdarstellungen ein sehr unzuverlässiger Mechanismus.

Erkennung modifizierter Versionen bekannter Missbrauchsdarstellungen. Hierbei handelt es sich um Detektoren, die darauf abzielen, Variationen von bekannten Missbrauchsdarstellungen zu identifizieren. Ihr Ziel ist es, die Einschränkungen von Detektoren zu überwinden, die auf kryptografischen Hash-Funktionen hinsichtlich der Modifikation der Bilder basieren.

Diese Detektoren verwenden eine Technologie namens „Perceptual Hashing”, die Eingaben ebenfalls eine kurze Zeichenfolge (Hashwert) zuordnet, sodass bei zwei visuell ähnlichen Eingaben (z. B. einem gedrehten Bild oder einem Bild, bei dem einige Pixel gelöscht wurden) die resultierenden Hashwerte identisch oder ähnlich sind. Diese Detektoren verfügen über eine

Liste von Perceptual-Hash-Werten, die bekannten Missbrauchsdarstellungen entsprechen. Um zu entscheiden, ob ein Inhalt in der Liste der Perceptual-Hash-Werte enthalten ist, berechnet der Detektor die Perceptual-Hash-Werte der Eingabe und prüft, ob diese ausreichend nahe an einem Perceptual-Hash-Wert der Liste liegen. Wenn ja, wird der Inhalt als Missbrauchsdarstellung gekennzeichnet und wie im vorherigen Fall gemeldet.

Je nachdem, wie „ausreichend nahe“ definiert ist, kann es vorkommen, dass dieser Detektor viele bekannte Missbrauchsdarstellungen nicht erkennt (wenn „nahe“ sehr restriktiv definiert ist) oder sie zwar erkennt, aber auch viele Inhalte, die keine Missbrauchsdarstellungen zeigen, als solche markiert (wenn „nahe“ sehr locker definiert ist). Darüber hinaus hat die Forschung mehrere gravierende Mängel in allen Entwürfen von Perceptual-Hash-Funktionen festgestellt: Es ist leicht, Erkennung von Missbrauchsdarstellungen zu umgehen, indem man für das menschliche Auge unsichtbare Änderungen vornimmt, es ist möglich, unproblematische Inhalte zu generieren, die als Missbrauchsdarstellungen klassifiziert würden, um die Arbeitsbelastung von Strafverfolgungsbehörden zu erhöhen. Zudem ist die Falsch-Positiv-Rate von Perceptual-Hash-Funktionen so hoch, dass angesichts der Menge an Material, das sie analysieren müssen, Millionen von Bildern fälschlicherweise als CSAM identifiziert würden.

Daher ist ein auf Perceptual-Hash-Funktionen basierender Detektor nicht ausreichend zuverlässig, und es bestehen ernsthafte Zweifel daran, dass es technisch möglich ist, all diese Mängel gleichzeitig zu vermeiden. Darüber hinaus kann es bei einigen Perceptual-Hash-Funktionen möglich sein, aus dem Hashwert bestimmte Informationen über das Bild (z. B. die Kontur einer Person) abzuleiten, was dazu führen kann, dass diese Hash-Funktionen Informationen über gehashte Inhalte preisgeben.

Erkennung unbekannter Missbrauchsdarstellungen. Diese Detektoren zielen darauf ab, unbekannte Missbrauchsdarstellungen zu erkennen. Dazu stützen sie sich auf die Erkennungen von bestimmten Merkmale in Darstellungen sexualisierter Gewalt gegen Kinder.

Um Muster zu erkennen, stützen sich diese Detektoren auf Algorithmen des maschinellen Lernens, die anhand bekannter Missbrauchsdarstellungen (Bilder, Audio, Grooming-Texte) trainiert werden. Um den Algorithmus zu trainieren, muss das Material in Merkmale umgewandelt werden, die es repräsentieren. Um zu entscheiden, ob ein Inhalt CSAM darstellt, wird die Eingabe in Merkmale umgewandelt und diese Merkmale werden in den Algorithmus eingespeist, der eine Entscheidung ausgibt. Da der Begriff CSAM komplex und von Natur aus kontextabhängig ist, weisen solche Algorithmen in der Regel eine geringe Genauigkeit auf, das heißt, dass ihre Einschätzungen oft falsch sind. Ohne Kenntnis des Kontexts, in dem ein Inhalt entstanden ist, ist es sowohl für Menschen als auch Maschinen sehr schwierig einzuschätzen, ob ein Bild von Kindern neben einem Swimmingpool, ein von einem Elternteil aus medizinischen Gründen aufgenommenes Bild oder eine Sexting-Nachricht zwischen einwilligenden Teenagern als CSAM gemeldet werden sollten.

Ähnlich wie bei Detektoren, die auf Perceptual-Hash-Funktionen beruhen, ist es einfach, Eingaben so zu manipulieren (z. B. durch Ändern der Pixelwerte in Bildern), sodass sich die

Muster bei der Umwandlung der Eingabe in Merkmale so stark verändern, dass der Detektor das Bild nicht markiert. Daher sind auch diese Detektoren ungeeignet, um Darstellungen von Missbrauchsdarstellungen zuverlässig zu erkennen.

Wie unterscheidet sich Client-Side-Scanning von Malware-Filtern?

Auf den ersten Blick mag die Erkennung von Missbrauchsmaterialien durch Clien-Side-Scanning vergleichbar mit Spam- und Malware-Filtern durch Antivirensoftware erscheinen. Tatsächlich unterscheiden sich Client-Side-Scanning und Malware-Filter grundltzeregend. Wenn Antivirensoftware potenzielle Malware auf einem Endgerät findet, informiert es den Nutzer, der gebeten wird, eine Entscheidung über die Meldung zu treffen. Das heißt, dass Malware-Scanning freiwillig, transparent und nicht mit Hintertüren für Strafverfolgungsbehörden verknüpft ist. Dies ist beim Client-Side-Scanning nicht der Fall. Nutzer:innen haben keine Wahl, ob sie Clien-Side-Scanning auf ihren Geräten implementieren möchten, und haben keinerlei Kontrolle oder Korrekturmöglichkeiten von automatisierten Meldung von angeblichen Missbrauchsdarstellungen. Daher ist Client-Side-Scanning von Natur aus invasiver als jeder Malware-Schutz, unzuverlässiger und anfälliger für Missbrauch, und unterminiert immer die Vertraulichkeit privater Kommunikation.

Verletzt Client-Side-Scanning die Vertraulichkeit privater Kommunikation, die durch die Ende-zu-Ende-Verschlüsselung gewährleistet wird?

Verschlüsselung ist das Werkzeug, mit dem wir die Vertraulichkeit von Kommunikation gewährleisten. Bei der Verschlüsselung werden Nachrichten in unverständliche, verschlüsselte Daten umgewandelt, sodass nur diejenigen, die den Entschlüsselungscode kennen, den Inhalt der Nachrichten erkennen können. Ende-zu-Ende-Verschlüsselung bedeutet zusätzlich, dass niemand außer dem Absender und dem vorgesehenen Empfänger der Kommunikation den Inhalt der Nachrichten entschlüsseln kann.

Der Kompromissvorschlag der dänischen Ratspräsidentschaft sieht Scannen jeglicher privater und verschlüsselter Kommunikation, und das Melden von potenziellen Missbrauchsdarstellungen vor. Dies verletzt die Vertraulichkeitsbedingung der Ende-zu-Ende-Verschlüsselung: Ein Dritter, der weder Absender noch Empfänger ist – in diesem Fall Strafverfolgungsbeamte – erfährt, was in der Nachricht steht. Auch ohne dass Inhalte an Behörden ausgeleitet werden, verletzt das grundsätzliche Scannen jeglicher privaten Kommunikation ihre Vertraulichkeit.

Dieser Eingriff in Ende-zu-Ende-Verschlüsselung und die Vertraulichkeit privater Kommunikation ist unabhängig von der Art des verwendeten Detektors. Selbst wenn nur kryptografische Hash-Werte verwendet werden, werden alle Nachrichten aller Nutzer:innen gescannt. Dies ist besonders relevant, wenn es keine Möglichkeit gibt, die Inhalte, nach denen die privaten Nachrichten der Benutzer gescannt werden, technisch zu beschränken oder zu überprüfen.

Somit bricht Client-Side-Scanning, das Dritte (z. B. Strafverfolgungsbehörden) über den Inhalt von Nachrichten informieren kann, die Garantien der End-to-End-Verschlüsselung vollständig.

Kann Client-Side-Scanning so eingeschränkt werden, dass nur CSAM erkannt wird?

Clientseitige Detektoren identifizieren Material, das entweder in einer Liste enthalten ist (für kryptografische und Perceptual-Hash-Werte) oder dessen Muster mit denen in Trainingsdatensets übereinstimmen. Die Erkennungsalgorithmen sind unabhängig davon, welches Material in der Liste enthalten ist oder zum Trainieren des Algorithmus verwendet wurde. Wenn die Liste oder der Trainingsdatensatz Inhalte enthält, die mit Terrorismus, regierungsfeindlichen Positionen oder Protestorganisationen in Verbindung stehen, würde der Erkennungsmechanismus die gleichen Maßnahmen ergreifen.

Sobald clientseitige Detektoren eingesetzt werden, gibt es daher keine technischen Möglichkeiten, ihre Verwendung technisch einzuschränken, da dies nur von der Konfiguration abhängt, mit der sie ausgestattet sind, die sich jedoch leicht über ein Remote-Update ändern lässt. Dabei kann es sich um legitime Updates handeln, die von denjenigen bereitgestellt werden, die Client-Side-Scanning einsetzen, oder um illegitime Updates von Akteuren, die eine Schwachstelle des Systems ausnutzen. Angesichts der Sensibilität von CSAM dürfen die Erkennungstechnologien keine Rekonstruktion der von ihnen überprüften Inhalte zulassen, was es den Benutzern unmöglich macht, zu überprüfen, ob die Detektoren nur nach CSAM suchen und nicht nach mehr.