Hintergrundpapier zu den grundlegenden Unterschieden zwischen Spam-/Malware-Filtern und dem Scannen nach CSAM in privater Kommunikation[1]

September 2025

Bündnis „Chatkontrolle STOPPEN!“

Spam-/Malware-Filter werden manchmal mit den Scanningtechnologien verglichen, die in der von der EU vorgeschlagenen Verordnung zur Festlegung von Vorschriften zur Prävention und Bekämpfung des sexuellen Missbrauchs von Kindern (CSA-Verordnung) vorgesehen sind, insbesondere von Befürwortern des Vorschlags wie der ehemaligen EU-Kommissarin für Inneres, Johansson. Zwar gibt es in beiden Fällen eine automatisierte Erkennung bekannter und unbekannter Inhalte, doch gibt es eine Reihe entscheidender Unterschiede, die einen allgemeinen Vergleich zwischen CSA-Erkennung und Spam-/Malware-Filtern höchst irreführend machen.

Der Zweck dieses Hintergrundpapiers ist es, diese Unterschiede aufzuzeigen.

Spam-/Malware-Filter sind nicht gesetzlich vorgeschrieben

Spam-/Malware-Filter sind nicht gesetzlich vorgeschrieben. Sie werden freiwillig von Systemadministratorinnen oder Nutzern eingesetzt, um die technische Infrastruktur vor Bedrohungen durch Malware und Spam zu schützen. Diese Technologien zielen aber nicht auf das Berichten über gesendete oder empfangene Inhalte ab, sondern auf zum Beispiel das Erkennen und unmittelbare Löschen von Malware , während E-Mail-Spam in einem Spamfilter zur möglichen Überprüfung durch den Empfänger abgelegt wird, da Spam in erster Linie ein Ärgernis darstellt. Der Benutzer behält hier die Kontrolle über das Ergebnis des Erkennungsprozesses, und manchmal können Spamfilter vom Benutzer vollständig deaktiviert werden.

Im Gegensatz dazu sind die Aufforderungen zur Ermittlung im Vorschlag für eine CSA-Verordnung für die Diensteanbieter verbindlich. Der Zweck besteht in der Ermittlung und Meldung potenziell rechtswidriger Inhalte (CSA) an das EU-Zentrum und die Strafverfolgungsbehörden. Da die obligatorische Erkennungsmaßnahme allgemein und unterschiedslos auf alle Nutzer angewendet wird, ist der Juristische Dienst des Rates zu dem Schluss gekommen, dass Erkennungsanordnungen einen besonders schwerwiegenden Eingriff in die Grundrechte darstellen, der wahrscheinlich den Kern des Grundrechts auf Privatsphäre verletzt und mit dem Grundsatz der Verhältnismäßigkeit gemäß Artikel 52 Absatz 1 der Charta unvereinbar ist. Der Wissenschaftliche Dienst des Deutschen Bundestages kam in seinem Rechtsgutachten zu dem Vorschlag zu einem ähnlichen Ergebnis.

Die Verarbeitung der Inhalte ist sehr unterschiedlich

Spam-/Malware-Filter melden die erkannten Inhalte ausschließlich dem Nutzer, was nicht als gleichwertiger Eingriff in die Privatsphäre angesehen werden kann. Die erkannten Inhalte werden in einem Spam-Ordner abgelegt, wo der Nutzer die Inhalte überprüfen und gegebenenfalls die automatische Klassifizierung ändern kann. Spam-Filter haben inhärente Probleme mit falsch-positiven Erkennungen, und ohne manuelle Überprüfung können wichtige Nachrichten verloren gehen.

In der CSA-Verordnung werden jedoch alle Inhalte, die als CSA erkannt werden, automatisch an das EU-Zentrum gemeldet, das verpflichtet ist, die Meldung an die Strafverfolgungsbehörden weiterzuleiten, es sei denn, das EU-Zentrum hält die Meldung für „offensichtlich unbegründet“. Auch falsch-positive Erkennungen werden dem EU-Zentrum und in vielen Fällen den

1 Grundlage dieses Papiers ist eine englischsprachige Fassung von European Digital Rights (EDRi): https://edri.org/wp-content/uploads/2025/09/Briefing-on-the-fundamental-differences-between-spam-malware-filters-and-CSA-detection-in-private-communications.pdf.

Strafverfolgungsbehörden gemeldet (aufgrund der Schwelle der offensichtlichen Unbegründetheit, die verlangt, dass selbst höchstwahrscheinlich rechtmäßige Inhalte vorsichtshalber weitergeleitet werden müssen). Diese automatisierte Überwachung privater Nachrichten stellt einen massiven Eingriff in die Privatsphäre dar.

Transparenz hinsichtlich der Erkennungsregeln

Bei Spam-/Malware-Filtern ist vollständige Transparenz hinsichtlich der Erkennungsregeln möglich, da die erkannten Inhalte nicht illegal sind. Einige Spam-/Malware-Filterlösungen basieren auf Open-Source-Software und öffentlich zugänglichen Listen mit Erkennungsindikatoren für bösartige Inhalte (Spam und Malware).

Die automatisierte Erkennung in der CSA-Verordnung wird für die Nutzer jedoch völlig undurchsichtig sein, da die Liste der Hash-Werte (für bekannte Inhalte) und KI-Klassifikatoren (für unbekannte Inhalte) geheim gehalten wird. Diese Geheimhaltung ist zwar aus Sicht der Strafverfolgung verständlich (z. B. um laufende Ermittlungen nicht zu gefährden) und weil Untersuchungen gezeigt haben, dass Menschen das Missbrauchsmaterial durch die Filter selbst rekonstruieren können, bedeutet sie aber auch einen gravierenden Mangel an Transparenz, der für das Vertrauen der Öffentlichkeit und den Schutz vor Missbrauchsrisiken von entscheidender Bedeutung ist. Das bedeutet, dass selbst unabhängige Technologieexperten die Technologie nicht überprüfen können.

Im Zusammenhang mit der (mangelnden) Transparenz ist anzumerken, dass es keine einheitliche Definition von CSAM als rechtswidrigen Inhalten gibt. Laut einer Studie von INHOPE gibt es eine Reihe von Unterschieden zwischen den nationalen Gesetzen zur Einstufung von CSAM, selbst innerhalb der Europäischen Union, wo alle Mitgliedstaaten die CSA-Richtlinie umgesetzt haben.

Im Wesentlichen muss die Öffentlichkeit blind darauf vertrauen, dass nur CSA-Inhalte von der „Black Box“-Software erkannt werden und dass es keine Ausweitung auf andere Inhalte gibt, weder auf andere illegale Inhalte als CSA noch auf Inhalte, die von den etablierten politischen Interessen als schädlich angesehen werden. Zwar ist eine unabhängige Überprüfung zumindest konzeptionell möglich, doch wird die Geheimhaltung des Erkennungsrahmens die Zahl der Personen begrenzen, die für eine solche Überprüfung zur Verfügung stehen (es ist illegal, CSAM anzusehen, was bedeutet, dass unabhängige Experten die Technologie in der Regel nicht überprüfen können). Die Überprüfung der Funktionsweise der automatisierten Erkennung – normalerweise eine gute Sicherheitsvorkehrung für solche Systeme – kann rechtlich und praktisch nicht unter Einbindung der Öffentlichkeit stattfinden.

Erkennung in Ende-zu-Ende-verschlüsselten Diensten

Der Umgang mit Ende-zu-Ende-verschlüsselten (E2EE) Kommunikationsdiensten ist wohl der größte Unterschied zwischen der CSA-Verordnung und Spam-/Malware-Filtern. Spam-/Malware-Filter werden in der Regel auf zentralen Servern für unverschlüsselte Kommunikation, z. B. E-Mails, eingesetzt. Bei E2EE-Diensten kann die Erkennung nur auf dem Gerät des Benutzers erfolgen, da die Kommunikation zwischen dem Versenden und dem Empfangen vollständig verschlüsselt ist. Die Erkennung von Spam/Malware bei E2EE-Diensten wird auf das technisch Machbare reduziert, ohne die kritischen Sicherheits- und Datenschutzgarantien von E2EE zu verletzen, nämlich dass die verschlüsselte Kommunikation nur für den Absender und den/die vorgesehenen Empfänger zugänglich ist. **Dienstanbieter unternehmen große Anstrengungen, um Nutzer zu versichern, dass keine Informationen das Gerät verlassen und dass die Privatsphäre der Kommunikation nicht beeinträchtigt wird (d. h. dass niemand außer dem Absender oder Empfänger den Inhalt der Kommunikation sehen oder darauf zugreifen kann).**

Dies wird anhand von zwei Fällen veranschaulicht. Der erste wird häufig von Befürwortern der CSA-Verordnung angeführt, da WhatsApp Ende-zu-Ende-verschlüsselt ist:

1. WhatsApp warnt den Nutzer vor verdächtigen URLs. Diese Klassifizierung erfolgt vollständig auf dem Gerät des Nutzers, und es findet keine externe Kommunikation zum Abrufen von URL-Datenbanken statt. Stattdessen wendet WhatsApp eine Erkennungsregel an, die nach Zeichen sucht, die typischerweise nicht Teil echter URLs sind.

2. Google bietet in der Android-App „Google Messages“ eine KI-basierte Spam-Erkennung an (die vom Nutzer in den App-Einstellungen deaktiviert werden kann). Diese App verarbeitet sowohl unverschlüsselte Nachrichten (SMS) als auch verschlüsselte Nachrichten (RCS). Laut Googles eigener Beschreibung können unverschlüsselte Nachrichten vorübergehend auf einem Server verarbeitet werden, wenn die Analyse auf dem Gerät nicht unterstützt wird (z. B. aufgrund von CPU- und Speicheranforderungen für die KI-Analyse auf dem Gerät). **Google weist jedoch ausdrücklich darauf hin, dass dies nur für unverschlüsselte Nachrichten geschieht und dass Chats zwischen Nutzern bei aktiviertem RCS immer Ende-zu-Ende-verschlüsselt sind**.
   Während die Verarbeitung personenbezogener Daten zur Spam-Erkennung in Google Messages wegen möglicher Verstöße gegen die EU-Datenschutzvorschriften durchaus kritisiert werden kann, respektiert Google hier zumindest das wichtige Prinzip von E2EE, dass die Kommunikation nur für den Absender und Empfänger zugänglich ist.

Für die CSA-Verordnung ist derselbe Umfang der Erkennung für unverschlüsselte und verschlüsselte Kommunikationsdienste vorgeschrieben („eine Ergebnisverpflichtung (keine Mittelverpflichtung)“, wie es die Kommission formuliert). Das „Client-Side-Scanning“, d. h. die Erkennung auf dem Gerät des Nutzers, ist in Artikel 10 Absatz 1 des dänischen Kompromisstextes vom Juli 2025 ausdrücklich vorgeschrieben. Während der Inhalt der Nachricht auf dem Gerät des Nutzers analysiert wird, muss die Erkennungssoftware in der Regel mit externen Servern kommunizieren, z. B. um Hash-Listen und KI-Klassifikatoren abzurufen und möglicherweise Fragmente des Inhalts zur Analyse auf zentrale Server hochzuladen. Die technischen Gründe dafür sind in Anhang 9 der Folgenabschätzung für die CSA-Verordnung dargelegt (ein Grund ist die oben erwähnte Geheimhaltung rund um Hashes und Klassifikatoren).

Darüber hinaus wird die Nachricht im Falle eines Erkennungsereignisses, unabhängig davon, ob es sich um einen echten oder einen falsch-positiven Treffer handelt, sofort an das EU-Zentrum und die Strafverfolgungsbehörden weitergeleitet. Dies verstößt gegen die kritischen Datenschutz- und Sicherheitsgarantien von E2EE, da die Nachricht anderen Stellen als dem Absender und den vorgesehenen Empfängern zugänglich gemacht wird. Die automatisierte Erkennung auf dem Gerät und die damit verbundene Interaktion mit externen Servern führen zu neuen Cybersicherheitsrisiken, die realistisch gesehen nicht angemessen gemindert werden können.[2]

Zusammenfassend lässt sich sagen: die Tatsache, dass ein Unternehmen wie WhatsApp eine sehr rudimentäre Überprüfung von Links durchführt, ohne dass Informationen das Gerät verlassen (oder in das Gerät gelangen), kein Beweis dafür ist, dass die im Rahmen der CSA-Verordnung vorgeschlagene Art der Überwachung funktionieren würde. Im Gegenteil, die komplexe CSAM-Erkennung steht, wie bereits erläutert, vor vielen technischen Hürden und Einschränkungen und ist darauf angewiesen, Informationen von außerhalb des Geräts einzubeziehen und diese dann auch außerhalb des Geräts zu melden. Zusammengenommen bedeuten diese wesentlichen technischen, betrieblichen und rechtlichen Unterschiede: ein Vergleich der Praktiken ist nicht möglich.

2 Vgl. Anderson et al, „Bugs in our pockets: the risk of client-side scanning”: https://academic.oup.com/cybersecurity/article/10/1/tyad020/7590463.