Briefing on the fundamental differences between spam/malware filters and CSA detection in private communications

September 2025

Spam/malware filters are sometimes compared to the detection technologies envisaged by the EU's proposed Regulation laying down rules to prevent and combat child sexual abuse (CSA Regulation), notably by proponents of the proposal such as former EU Home Affairs Commissioner Johansson. Whilst there is automated detection of known and unknown content in both cases, there are a number of critical differences that make the overall comparison of CSA detection and spam/malware filters highly misleading. The purpose of this briefing is to outline these differences.

Spam/malware filters are not mandated by law

Spam/malware filters are not mandated by law. They are deployed voluntarily by systems administrators to protect the technical infrastructure against threats from malware and spam. The purpose of the detection (filter) is not to create reports about specific users and the content they send or receive. In most cases, detected malware is deleted (because of the potentially serious threat), whereas email spam is placed in a spam filter for the recipient's possible review, as spam is mainly an annoyance. The user retains considerable control over the outcome of the detection process, and sometimes spam filters can be switched off entirely by the user.

By contrast, Detection Orders in the CSA Regulation proposal are mandatory for the service provider. The sole purpose is detection and reporting of potentially unlawful content (CSA) to the EU Centre and law enforcement. Since the mandatory detection measure is applied to all users in a general and indiscriminate manner, the Council Legal Service has concluded that detection orders entail a particularly serious interference with fundamental rights which is likely to violate the essence of the fundamental right to privacy and not be compliant with the proportionality requirement in Article 52(1) of the Charter. The Research Service of the German Parliament reached an equivalent conclusion in its legal opinion on the proposal.

Handling of detected content is very different

Spam/malware filters report the detected content solely to the user, which cannot be regarded as an equivalent intrusion into the private sphere (moreover, it is not mandated by law). The detected spam content is placed in a spam folder where the user can review the content and change the automated classification, if needed. Spam filters have inherent problems with false-positive detections, and without manual review important messages can get lost.

In the CSA Regulation, however, any content which is detected as CSA will be automatically reported to the EU Centre, which is required to forward the report to law enforcement unless the EU Centre considers the report "manifestly unfounded". False-positive detections will also be reported to the EU Centre and, in many cases, law enforcement (due to the manifestly unfounded threshold, which requires that even most-likely lawful content should still be forwarded just in case). This automated surveillance of private messages is a massive intrusion into the private sphere, cf. the legal analysis from the Council Legal Service.

Transparency about detection rules

For spam/malware filters, full transparency about the detection rules is possible, since the detected content is not illegal. Some spam/malware filter solutions are based on open-source software and publicly available lists of fingerprints for the malicious content (spam and malware).

The automated detection in the CSA Regulation, however, will be completely opaque to the users, because the list of hashes (for known content) and AI classifiers (for unknown content) is kept secret. Whilst this secrecy is understandable from a law enforcement perspective (e.g. not jeopardising ongoing investigations) and because research has shown that people can reconstruct the abuse material through the filters themselves, it also means that there will be a serious lack of transparency which is critical for public trust and safeguarding against the risk of abuse. This means that even independent technological experts cannot vet the technology.

In connection with (lack of) transparency, it is pertinent to note that there is no unique definition of CSAM as unlawful content. According to a study by INHOPE, there are a number of differences across national laws for the classification of CSAM, even within the European Union where all Member States have implemented the CSA Directive.

In essence, the public must blindly trust that only CSA content will be detected by the “black box” software, and that there is no mission creep to other content, either unlawful content other than CSA or content that is regarded as harmful by the incumbent political interests. Whilst independent review is a possibility, at least conceptually, the secrecy of the detection framework will limit the number of persons available for conducting such review (it is unlawful to view CSAM, meaning independent experts usually cannot vet the technology). The review of how the automated detection works – normally a good safeguard for such systems – legally and practically cannot be crowdsourced to the public.

Detection in end-to-end encrypted communications

The handling of end-to-end encrypted (E2EE) communications services is, arguably, the biggest difference between the CSA Regulation and spam/malware filters.

Spam/malware filters are generally deployed on central servers for unencrypted communications, e.g. email. With E2EE services, the detection can only be done on the user’s device, since the communication is fully encrypted in transit between the sender and recipient(s).The detection for spam/malware on E2EE services is scaled down to what is technically feasible without breaking the critical security and privacy guarantees of E2EE, namely that the encrypted communication is only accessible for the sender and intended recipient(s). **Service providers go to great length to assure the user that no information leaves their device, and that the privacy of communications is not compromised (meaning that no one but the sender or recipient can see or access the content of the communications).**

This will be illustrated with two cases. The first one is commonly cited by proponents of the CSA Regulation because WhatsApp is E2EE:

1. WhatsApp alerts the user to suspicious URLs. This classification is done entirely on the user’s device, and there is no external communication to retrieve databases of URLs. Instead, WhatsApp applies a detection rule which looks for characters that are typically not part of genuine URLs.

2. Google offers AI-based spam detection in the Google Messages Android app (can be disabled by the user in the app settings). This app handles unencrypted messages (SMS) as well as encrypted messages (RCS). According to Google’s own description, unencrypted messages may be temporarily processed on a server if on-device analysis is not supported by the device (e.g. CPU and memory requirements for on-device AI analysis). **However, Google explicitly points out that this is only done for unencrypted messages, and that chatting between users is always end-to-end encrypted when RCS is enabled.**

   Whilst the processing of personal data for spam detection in Google Message can definitely be criticised for possible non-compliance with EU data protection rules, at least Google respects the critical principle of E2EE that the communication is only accessible for the sender and intended recipient(s).

For the CSA Regulation, the same scope of detection is prescribed for unencrypted and encrypted communications services ("an obligation of result not of means", as the Commission puts it). Client-side scanning, that is detection on the user's device, is explicitly mandated by Article 10(1) of the July 2025 Danish compromise text. Whilst the message content is analysed on the user's device, it will generally be necessary for the detection software to communicate with external servers, e.g. to retrieve hash lists and AI classifiers, and possibly upload fragments of the content for analysis on central servers. The technical reasons for this are outlined in Annex 9 of the Impact Assessment for the CSA Regulation (one reason is the secrecy around hashes and classifiers noted above).

Moreover, in case of a detection event, whether genuine or false-positive, the message is immediately forwarded to the EU Centre and law enforcement. This breaks the critical privacy and security guarantees of E2EE, as the message is made available to other entities than the sender and intended recipients(s)**.** The on-device automated detection and its associated interaction with external servers introduce new cybersecurity risks that realistically cannot be properly mitigated, cf. the academic article "Bugs in our pockets: the risk of client-side scanning".

To summarise the critical differences, the fact that a company like WhatsApp performs a very rudimentary check of links, without any information leaving (or entering) the device, is not evidence that the sort of scanning proposed under the CSA Regulation would work. To the contrary, complex CSAM detection faces many technical hurdles and limitations as already explained, and relies on bringing in information from outside the device, as well as then reporting it outside the device. Together, these key technical, operational and legal differences mean that it is not possible to compare the practices.