EDRi assessment of the European Child Sexual Abuse Legislation Advocacy Group's (ECLAG) fact-checking of top 9 claims made on the CSA Regulation

A recent fact-checking paper by ECLAG claims that there are "a lot of misconceptions out there about the proposed Regulation to fight Child Sexual Abuse and the technology that is deployed to detect child sexual abuse material (CSAM)."

This is EDRi's analysis of the claims made by ECLAG under the guise of fact-checking misconceptions about the CSA Regulation.

Claim #1 Detection technology won't be effective in stopping the spread of CSAM.

ECLAG: *Detection can help dramatically stop the spread of child sexual abuse material online, as part of a toolbox of solutions needed to tackle this complex crisis. In 2021, detection efforts slowed down as the legal basis for detection expired. As a result, the number of incidents of reporting went down, despite the fact that data shows the volume of abusive material increased. Increasing detection efforts means more CSAM being found, removed and reported, which ensures dignity for victims and survivors and increases children's online safety.*

EDRi assessment: the number of NCMEC reports is not a reliable indicator of the spread of CSAM online. NCMEC reports depend on the voluntary scanning practices of service providers, with 85% of reports coming from Meta (in 2022). There are many false positives (e.g. naked children on a beach or voluntary "sexting" between teenagers in private messages) in the NCMEC reports. Of NCMEC's 4,192 referrals to the Irish police in 2020, 11% were clearly not CSAM, while less than 10% were "actionable". An analysis by Meta for October and November 2020 showed that 90% of the detected content was visually similar to previous reports, and that just **six videos were responsible for half of the child exploitive content reported by Meta** in that period. In a sample of 150 accounts reported to NCMEC, Meta estimates that 75% did not intend to harm a child, but shared the material for other reasons, such as outrage or poor humour.

Claim #2 This legislation is establishing general and ungrounded mass surveillance. The EU wants to "open every letter" and read each and every private message.

ECLAG: *[The claim that the CSA Regulation will lead to mass surveillance] builds on unfounded fears and on a misunderstanding of the technology at hand. Detection technology doesn't "read" messages. It either compares digital fingerprints of images via hash-matching to a database of known and verified CSAM - or it uses a classifier to flag content that is suspected to be CSAM which then undergoes a multi-step process to get verified as CSAM including human review. What's more, the legislation sets out a safeguard usage process through triangulation of private companies, public authorities* *and the courts. Additionally, the legislation builds in safeguards to ensure transparency in the use of detection tools, including the input of independent bodies or courts and of the data protection authorities.*

EDRi assessment: the proposed detection measures in the CSA Regulation are **general and indiscriminate.** Service providers must access every private message in order to compare its content against a database of hash values and AI classifiers. This is "reading" by automated means. The **detection obligations also apply to end-to-end encrypted (E2EE) communications services**, that is services that are specifically designed to prevent private messages from being read by anyone other than the sender or recipient, including by the service provider. **Detection in E2EE services is only technically possibly by deliberately undermining the security design of the communications service.**

While detection of known and unknown CSAM uses image and video attachments, detection of grooming (solicitation of children) requires an analysis of the text content of the message (with AI classifiers), and possibly parts of the message history between the sender and recipient since grooming detection often depends critically on context. If there is a match, including falsepositives, the message content along with other information about the user profile will be reported to the EU Centre, where only manifestly unfounded reports are filtered out before forwarding private communications to law enforcement. With the high error rates of especially grooming detection, a large part of the EU population risks being wrongly implicated in child abuse and reported to law enforcement.

The fact that the CSA Regulation provides a transparent process for issuing detection orders does not make the detection technology and deployment practices transparent from the view point of the individuals who believe they are communicating privately online even though their communication is, in fact, subjected to mass surveillance.

Claim #3 CSAM detection tools are easy to reengineer for other purposes.

ECLAG: *The tools to detect CSAM are highly targeted at finding CSAM – and only that. If someone wanted to detect other content, they would need to design entirely different tools and resources. Under the legislation, the EU Centre will provide access to accredited state of the art tech which is designed to only detect CSAM. It will also periodically review their effectiveness. What's more, the use of these technologies will only be permitted on a case by case basis under the review of public authorities.*

EDRi assessment: while the proposed CSA Regulation only provides a legal basis for detection of child sexual abuse, the envisaged detection technologies, whether perceptual hashing or AI classifiers, are general technologies. They can easily be repurposed for detection of other material. It's not a matter of developing new tools, as ECLAG claims, but only about adding digital signatures for detection of other material. **This type of mission creep can happen unofficially, by silently (and unlawfully) adding a low number of non-CSA images to the hash database, e.g. to prevent the online spread of a video documenting police brutality or finding whistleblowers leaking secret documents to journalists, or officially ("prescribed by the law") by introducing new legislation that require scanning of private messages to combat other serious crimes than child sexual abuse.** Independent auditing of the digital signatures used for detection and accreditation by the EU Centre can only address the former problem, not the latter.

Once the mass surveillance (detection) infrastructure intended for CSA has been widely deployed, there will inevitably be a temptation among some lawmakers to use it for other purposes. This argument is by no means a theoretical one. **Spain, the current holder of the Council Presidency, sees the CSA Regulation as an opportunity for state access to encrypted communications more generally or even banning end-to-end encryption entirely.** Widespread deployment of client-side scanning would be tantamount to banning end-to-end encryption because the security guarantees become broken by design with the mandatory spyware on user devices.

Claim #4 CSAM detection tools have a high false positive rate which leads to innocent people getting prosecuted.

ECLAG: *False positive rates are a trade-off between precision rates (how much of the flagged content is CSAM) and recall rates (how much of the CSAM on a platform is being detected). In practice, detection methods are tuned to have extremely high precision rates. For all unknown CSAM, there is a multi-step system in place ensuring only CSAM gets flagged as CSAM: First, service providers should conduct human review of newly flagged CSAM. Secondly, the US-American Center for Missing and Exploited Children (NCMEC) or in future the EU Centre - with analysts trained to identify illegal content - get to review to ensure the material flagged is actually CSAM. It is highly unlikely that these two instances misinterpret an image.*

EDRi assessment: claims about precision rates for detection of known CSAM, unknown CSAM and grooming have never been independently verified. A freedom of information request in 2022 by Felix Reda confirms that **the precision rates reported in the Commission's Impact Assessment are industry figures without any independent validation.**

For known CSAM, where perceptual hashing techniques are used, claims of extremely high accuracy rates are frequently reported, notably the the *expected* error rate of 1 in 50 billion for PhotoDNA, which comes from the PhotoDNA inventor Hany Farid. This claim is highly questionable in light of the vulnerabilities of perceptual hashing against adversarial attacks and other research findings on the limitations of PhotoDNA. The use of PhotoDNA by LinkedIn showed that in 2021, **only 41% of what PhotoDNA flagged as CSAM actually constituted illegal material in the EU.** This shows that discussions of accuracy can be something of a red herring, because much of the material that has actually been put into hash databases does not constitute CSAM.

For unknown ("new") CSAM and grooming (solicitation of children) the error rates are much higher since detection is based on classification by AI systems rather than distance metrics (matching of hash values) from images that have previously been classified as CSAM by human experts (e.g. the EU Centre in the CSA Regulation proposal). Precision rates for true positives of *up to* 99.9% for unknown CSAM must be viewed with substantial scepticism since there is a trade-off between false-positive and false-negative error rates..The false-positive error rate in real-world deployments of the detection technologies will most likely be higher than the theoretically attainable 0.1%. For grooming, the Commission Impact Assessment reports an accuracy of 88%. The source is Microsoft which has publicly challenged the Commission's claims in the public consultation on the CSA Regulation.

As outlined in the analysis of claim #9 below, **even a high precision rate still means that millions of lawful private messages and social media content will be wrongly flagged as CSA.**

The supplementary Impact Assessment requested by the LIBE Committee concludes that technologies to detect new child sexual abuse material and grooming are of substantially lower accuracy than technologies for known CSAM. Similarly, at a hearing in the German Bundestag on 1 March 2023, Markus Hartman, senior public prosecutor from North Rhine-Westphalia cautioned against using AI to detect unknown criminal content due to the high number of false positive detections whereby innocent members of the public will come under suspicion and investigation by law enforcement.

The risks for innocent people are genuine, as demonstrated by a request from the Irish Council for Civil Liberties (ICCL) to the Irish police. In Ireland, which has been receiving NCMEC referrals directly since 2015, and indirectly since 2010, it is the protocol of the Irish authorities to retain the personal data of *all* NCMEC referrals, even the referrals that the Irish authorities themselves are satisfied are not CSAM, including innocuous images of naked children on a beach or in a bath, or consensual sexting images shared between adults. This means innocent people are being kept in a net of surveillance and suspicion with no cause.

Claim #5 CSAM detection tools wrongly flag consensually shared imagery or pictures of kids in bathtubs.

ECLAG: *CSAM detection tools are specifically trained to not find "kid in the bathtub" type innocent images. These tools are trained on known CSAM, adult pornography and benign images particularly to tell the difference between them and to keep benign imagery from being misinterpreted as CSAM.*

EDRi assessment: The evidence speaks to the contrary. When asked about the general nature of the non-illegal content which triggers a false positive referral from NCMEC, the Irish police confirmed: "OnCE [Online Child Exploitation Unit] will not action a referral further for a number of reasons on the basis of its content, the following are examples: Children playing on a beach, topless content, nudist, adult content, etc." When asked how many referrals contained non-illegal content per year, the Irish police confirmed: "OnCE doesn't use a specific categorisation of non-illegal. A total of 471 were marked as being not Child Abuse Material in 2020 from a total of 4,192. This is the focus of the OnCE unit. 506 referrals were marked as being age undetermined. 940 referrals included IP addresses which could not be progressed further. 852 referrals were marked as Child Abuse Material. 606 were marked as below the threshold. 75 were self-generated. 333 were marked as viral. 51 were adult."

Claim #6 CSAM is hosted primarily on the dark web, therefore it is not helpful to detect and fight CSAM on the open web.

ECLAG: *Tens of millions of pieces of CSAM are distributed on the open web, which for example the NCMEC report numbers show. The dark web is similar to the internet from the early nineties before search engines - it is slow! This means that uploading and downloading content like CSAM, videos in particular, takes very long. One of the most common use cases for the dark web is therefore to share links to CSAM that is hosted on the open web. Some sites intentionally mix*
CSAM with benign imagery to hide CSAM in plain sight. Fighting grooming and sextortion also makes most sense on the open web since this is where most kids are and where they are approached by perpetrators.

EDRi assessment: NCMEC statistics cover service providers that report instances of child sexual abuse online to NCMEC in accordance with their obligations under US law. As noted elsewhere in this analysis, **the NCMEC numbers are inflated by false-positive detections such as innocent pictures of naked children and consensual sharing of images**. There are no comparable statistics for other distribution channels, such as the anonymous Tor network ('dark web') and private virtual networks run by organised criminals that distribute CSAM and real-time online child sexual abuse in exchange for money. While Tor is slower than the ordinary internet due to routing traffic through multiple layers (onion routing), the speed of the Tor network has improved considerably in recent years along with the overall internet infrastructure. Comparisons to the internet in the early 1990s are clearly misleading.

Claim #7 There is no detection possible for E2EE environments, so we should carve out E2EE from this legislation.

ECLAG: *Detecting for CSAM within end-to-end encrypted environments can be done in a privacy-forward way through homomorphic encryption, multi-party computation, secure enclaves, or client-side-scanning (or a combination of these) with client-side-scanning as the most feasible option currently. And, there may be more ways we have yet to discover: it will take a multitude of solutions from industry to tackle the problem so that they can be used by a variety of companies of different sizes and scales. Knowing how fast technology evolves, it would be fatal to exclude any technology from the scope of this legislation. To incentivize innovation, this legislation must stay tech-neutral.*

EDRi assessment: all of the possible "solutions" mentioned by ECLAG break the security design of E2EE since the private messages can be accessed (read) by other parties than the sender and recipient. **In addition to interfering with the confidentiality of communications and the end-to-end protection principle, measures that circumvent encryption undermine security of information systems and create new vulnerabilities that can be abused by malicious actors** to gain access to user content, further violating their privacy, or target individuals with specially crafted innocent-looking content that triggers detection. Such attacks have been demonstrated for all known perceptual-hash algorithms, as pointed out in the joint statement of scientists and researchers on the EU's proposed Child Sexual Abuse Regulation on 4 July 2023.

Client-side scanning has been labelled "bugs in our pockets" in a critical analysis by security experts because the surveillance (detection) is moved from a central server to the user's device. This could directly facilitate function creep if the client-side scanning technology is tweaked to gain access to other information on the user's device than the private communications it is supposed to scan before they are encrypted. Indeed, client-side scanning arguably meets the formal definition of spyware in other EU legislative proposals. Client-side scanning is called the most feasible option currently by ECLAG, but it's noteworthy that there are no real-world deployments of client-side scanning for detection of child sexual abuse. The plans outlined by Apple in August 2021 were quickly put on hold after substantial public protest from security experts, and they were definitively abandoned by Apple in December 2022.

With homomorphic encryption, detection using encrypted content (without decrypting) is theoretically possible, but the key word here is *theoretical*. In Annex 9 of the Commission Impact Assessment, on-device homomorphic encryption with server-side hashing and matching has "LOW" technical feasibility. Moreover, the alternatives to client-side scanning (homomorphic encryption, secure enclaves and multi-party computation) are not considered in Annex 9 because of their additional privacy protection, but mainly as a security measure to prevent the hash list from being stored and processed on the user's device (where the hash list along with the detection algorithm can be extracted or reserve engineered).

From the viewpoint of the individual expecting private communications without any third-party access, client-side scanning vs. secure enclaves is simply a matter of substituting one encryption backdoor for another. None of the methods deserve to be called privacy-preserving because they all break the end-to-end security and confidentiality of communications technically guaranteed by E2EE.

Claim #8 The algorithms of a classifier are a black box and cannot be trusted. Also, it is biased.

ECLAG: We rely on algorithms already in numerous ways in our daily life. They are by no means a black box: The algorithms deployed to detect CSAM are carefully chosen by engineers. In this particular use case, algorithms cannot be made public to avoid that perpetrators use this knowledge to circumvent them. We support that the EU Centre provides exemplary sets of data on which these algorithms must meet certain benchmarks to be deployed in the EU. Any bias in such algorithms would simply be a lack of diverse enough data. The more data available for training, the better.

EDRi assessment: the algorithms are a black box from the viewpoint the individuals who rely on private messaging services and social media on a daily basis because, as ECLAG itself states, the detection algorithms cannot [will not] be made public. The problems with AI systems go far beyond debiasing, as detailed in this academic report commissioned by EDRi. Simply collecting more data, as ECLAG suggests, will not fundamentally address the problems with using AI-based detection systems.

Claim #9 This legislation would further overburden law enforcement agencies due to the rising number of reports. All we need is more resources for law enforcement.

ECLAG: *As for every crime, rising numbers are a reason to intensify efforts and not to look away from the evidence. This said, law enforcement officers involved in the fight against CSAM are dependent on more information and details to quickly identify and rescue children at immediate risk. Sufficient funding for law enforcement goes hand in hand with technologies that support their*
efforts. Beyond law enforcement investigation, the detection of CSAM plays a crucial role in reducing the further spread of CSAM online, which ensures dignity for victims and survivors, and reduces access to CSAM. It requires the whole child protection ecosystem ranging from hotlines to survivor support to research to technology and more to effectively fight child sexual abuse.

EDRi assessment: As noted in the research paper *Chat Control or Child Protection?* by professor Ross Anderson, **modern messaging systems operate at such a scale that filters need a false-positive error rate of 0.01% to be deployable, and 0.001% to be effective. Given the 10 billion messages sent and received every day in the EU, even error rates of 0.001% would still mean 100,000 messages sent for moderation every day.** The CSA Regulation envisages that the newly formed EU Centre will deal with this traffic. Ian Levy (NCSC) and Crispin Robinson (GCHQ) note in a paper that the UK National Crime Agency triages 100,000 alerts a year from NCMEC, and that this takes 200 staff. With 5% false positives, the task would not be feasible at all. Indeed, the Commission assumed in an internal discussion document presented to Council in June 2022 that they might get 10% false positives but then calculated that these would be 10% of the 1,000,000 true positives. However, the Commission got their arithmetic wrong. **The false alarms would be 10% of all the private messages processed, or a billion alarms a day.**