6 FEB · 8 MIN

EU-backed surveillance software for detecting child abuse has serious flaws, manufacturer admits

ALEXANDER FANTA

Controversial non-profit Thorn touts its software as a privacy-friendly way to detect child abuse content. Experts have criticised such tools as potentially leading to mass surveillance – and the technology, while praised by the European Commission, is less reliable than it purports to be, new documents obtained by Follow the Money show.

Software designed to fight the spread of child sexual abuse might be less effective than the European Union presented it to be, lobby papers obtained by Follow the Money show.

Thorn, a US-based foundation set up by actors Demi Moore and Ashton Kutcher, claims its technology deploys a way to find child sexual abuse material (CSAM) in personal user messages while respecting privacy of other users that don't share such content. Thorn has lobbied the Commission over the past years to create legal obligations for online platforms such as Facebook or TikTok to deploy software like the one it has developed, a previous investigation by Follow the Money shows.

For ordinary users, this could mean every single photo or video they send via a platform such as TikTok or X, formerly Twitter, including private messages, gets scanned. If the software recognises a match a child abuse image, it will be forwarded to law enforcement agencies in the country of the user – at least in theory.

The most common form of scanning for CSAM compares the pictures' digital fingerprints, so-called hashes, with those of content already known to authorities as containing child abusive content. If there is a match, the image gets flagged.

Facebook already uses such technology on a voluntary basis. If the EU law passes, it could mean that the obligation is extended to end-to-end encrypted services such as WhatsApp or Signal.

The most common form to find child abuse online compares the pictures' digital fingerprints with those of content already known to authorities

The Commission's Impact Assessment report accompanying its proposal mentions Thorn's software as "one example of industry's ability to detect child sexual abuse material."

But negotiations over the proposal amongst member states have stalled over concerns that mandatory scanning of private user messages would be unduly intrusive and, if deployed against millions of users without prior suspicion, amounts to mass surveillance. Digital rights NGOs also warn that erroneously flagged images could lead to false accusations of CSAM possession and overload law enforcement with bogus leads.

The Commission has pushed back against the criticism, arguing that its proposal requires companies to use technologies that are "the least privacy-intrusive."

Automated detection software is meant to scan messages and only report them if they match a known child abuse image, or if machine-learning software finds a new image closely resembles CSAM. The Commission insists that technologies are extremely accurate in detecting the content, and that "no false positives" – images wrongly tagged as child abuse material – would reach law enforcement.

But that claim doesn't reflect reality, documents show.

Documents reveal what the Commission is trying to hide

The Commission didn't supply all requested documents (see box). However, Follow the Money obtained unredacted versions of the policy papers through an access request with the Swedish government, which was also lobbied by Thorn. The documents cast doubt on technical capabilities of Thorn's software.

One of the papers, "False Positive Mitigation", is sparse on details, but hints at the technical challenges of automated CSAM detection.

The company does insist that "no technology is completely infallible, but as we innovate these proven tools only improve." However, the document also makes clear that – far from being a technical panacea – Thorn's software will wrongly identify images as child abusive content. This, in turn, means that humans – either corporate content moderators

or law enforcement – will have to review flagged images.

GETTING ACCESS TO DOCUMENTS:

When asked to provide technical descriptions of the detection software provided by Thorn, Safer, the Commission refused to provide the relevant documents, citing the need to protect the non-profit foundation's commercial interests. The EU executive only provided – after consultation with Thorn – versions that were redacted at key parts of two policy papers and the minutes, and withheld the third document entitled "False Positive Mitigation".

But that doesn't comply with EU transparency rules, the European Ombudsman, the EU's watchdog, found after FTM lodged a complaint. Access to the documents should "enable the public to participate more effectively in a decision-making process that will very likely directly affect citizen's day-to-day life by limiting their right to privacy", Ombudsman Emily O'Reilly concluded. "Stakeholders who actively provide input should not be allowed to do so behind closed doors."

Follow the Money's complaint concerns three one-page policy papers provided by Thorn, as well as minutes to a meeting between Commission officials and the non-profit.

The Ombudsman's verdict is not legally binding on the Commission. The Commission has yet to reply to the Ombudsman's recommendation to release the papers.

In deploying algorithms which match the characteristics of an image with those of known CSAM, "companies [that use the technology] get to make decisions on how accurate they want their systems to be," Thorn's policy document states.

According to a Commission report, Thorn's software is about 99.9 per cent accurate for detecting known child abuse content. At this rate, one in a thousand images would turn out to be false positives – images that don't contain CSAM but were flagged by the programme as problematic.

"One in 1,000 images being falsely flagged might seem like a small number, but it isn't," said Kris Shrishak of the NGO Irish Civil Liberties Council.

European users exchange millions of pictures every day; even if only a small fraction of these are flagged, and of those 0.1 percent are false positives, that could produce tens of thousands of false positives a month, Shrishak said.

Furthermore, this would imply that ordinary users could have their photos sifted through and analysed by people working for companies that use the technology – even if they shared no problematic content.

Potentially similarly problematic, set at this precision rate, the algorithm would fail to identify 20 percent of the CSAM images in the flagged data, according to tests cited by the Commission.

This means that if companies want to limit the number of false accusations that Thorn's algorithm sends for further review by human content moderators, it will need to accept a large amount of actual CSAM will never be flagged by the software.

"There is a trade off in how much illegal material is found", Thorn admitted in one of the unredacted versions of the documents, dated March 2022.

Thorn admits difficulties with scanning technology

The Swedish documents also contain one that Thorn did not send to the Commission, but had shared with Sweden's Permanent Representation to the European Union.

The two-page paper from February 2022, drafted by Thorn lobbyist

Emily Slifer and entitled *Encryption*, deals with a particularly controversial aspect of the Commission's proposal: the possibility to oblige social media companies, for example, to use detection software that circumvents end-to-end encryption by scanning images directly on the user's device.

While negotiations between member states over the draft child abuse law are still ongoing, the European Parliament has made clear that it doesn't want obligations to scan for CSAM to apply to end-to-end encrypted communications.

Thorn's *Encryption* paper shows that what the non-profit publicly says about the capability of software to detect CSAM and what is actually possible don't match.

"Current end-to-end encryption technology makes it impossible to detect, remove, and report illegal CSAM," Thorn wrote in its policy paper.

Thorn introduces two techniques to detect such content.

One, called "on-device hashing", creates a digital fingerprint of pictures and video files stored on the device. It then sends this fingerprint to a server where it is compared with those of known CSAM images, such as photos or videos that local police had identified as containing prohibited content.

According to Thorn's paper, this technique reliably identifies child abuse images that are already known to law enforcement.

Yet, that only relies on previously identified photos and videos. The technology "does not comprehensively detect new child abuse material," the non-profit acknowledges.

Kutcher's false appeal

Another method, called "homomorphic encryption", would detect both new and known child sexual abuse material in a "privacy centric" way, according to Thorn's paper.

Homomorphic encryption is a method that is meant to allow querying data, such as finding certain patterns in an image, even when encrypted. In theory, this could allow the detection of CSAM in photos or videos without sharing the actual image with a central database.

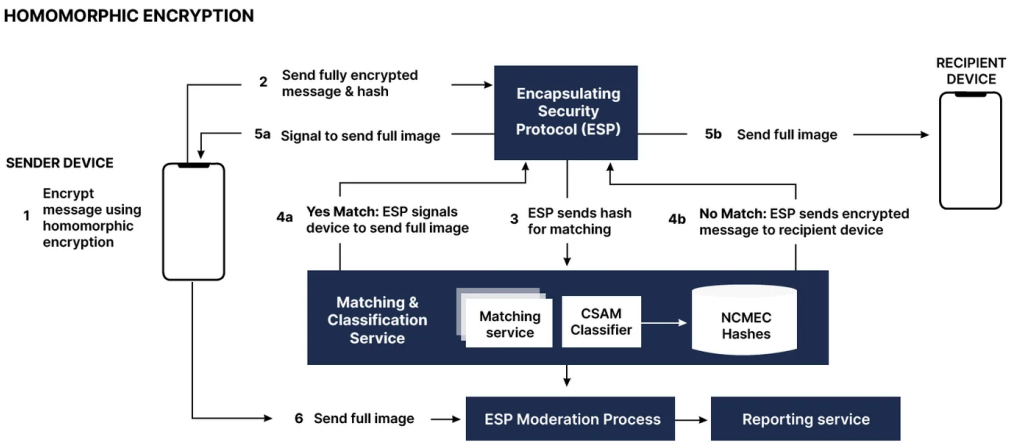
Thorn founder Ashton Kutcher, who recently left the non-profit over accusations of having shielded a fellow actor from rape charges, has touted the virtues of homomorphic encryption.

In a tearful, dramatic presentation in the European Parliament, Kutcher made his case for using the technology. "We can also use it to detect child sexual exploitation without ever looking at the image," he said.

However, according to Thorn's policy document sent months before Kutcher's presentation, "devices currently available in the market today do not have the processing capacity for this solution to be deployed immediately at scale. More technological progress and research is needed to fully vet this opportunity," the document said.

Homomorphic encryption requires a significant amount of computing power. This means that its use for on-device scanning on phones is "unrealistic" for the foreseeable future, said Anja Lehman, a cryptography researcher at the University of Potsdam. She said that currently no privacy-friendly technology exists that can be used to scan for CSAM in encrypted environments.

Lehmann pointed out another apparent contradictory statement in the presentation. A graphic included in Thorn's paper purports to show how detection software works in practice.



Lehmann said that, if the graphic is taken at face value, homomorphic encryption would be used only on the server. This means that, to detect a CSAM image, the server must have access to the entire image.

“In simple terms, this would render any privacy claims absurd, as this would allow an authority controlling the server to access any user image sent for detection”, the cryptography researcher said.

This raises the question whether Thorn publicly promoted a technology that it has privately acknowledged to not be technically feasible, raising the possibility that it wants to use it as mere window dressing.

The non-profit’s Brussels lobbyist Emily Slifer admitted that Thorn’s software currently cannot scan images on-device. She clarified in an e-mailed reply that “Thorn’s technology does not work in encrypted spaces.”

> **“Thorn’s technology does not work in encrypted spaces” - Thorn lobbyist Emily Slifer**

“While effective solutions for detecting CSAM in encrypted

environments do not currently exist, we believe from our deep experience in the tech space that such solutions will one day exist," Slifer said.

Asked why Thorn promoted homomorphic encryption despite privately saying it could not be deployed to detect child abuse images, Slifer acknowledged that "the physical chips for the computing power required to use it at scale to detect CSAM on encrypted platforms are in development.

"Therefore, we must have more conversations and brainstorming as a broader child safety ecosystem to determine how we can scale the tech, bring it up to speed, and best use it to defend children from sexual abuse," Slifer added. She did not address the question why Kutcher promoted a technology that Thorn knew to be not feasible at the moment.

Asked why Thorn had tried to prevent the release of the documents in question, Slifer replied that "these actions were not aimed at hiding information but were necessary because some of them contain sensitive details about the operational aspects and methodologies of our software, which could potentially undermine its effectiveness and inadvertently aid those intent on circumventing these protective measures."

Patrick Breyer, a German lawmaker who is part of the Parliament's Green group, said that the documents revealed the Commission's plans to scan for CSAM were based on flawed technical assumptions.

"We can tell from the latest disclosures that Thorn's lobby documents don't concern its intellectual property at all," he said, "but are withheld because their content is politically inopportune to the Commission's chat control plans."

Author: **Alexander Fanta**

Covers technology and tech policy-making in the EU, and likes to uncover lobbying with Freedom of Information requests.