CSAM?

Could a Snitch Analyse our Messages?

Jaap-Henk Hoepman

Privacy & Identity Lab
iHub
Radboud University
Karlstad University
University of Groningen

✉ *jhh@cs.ru.nl // www.cs.ru.nl/~jhh // blog.xot.nl // @xotoxot*

EUROPEAN
COMMISSION

Brussels, 11.5.2022
COM(2022) 209 final

2022/0155 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

laying down rules to prevent and combat child sexual abuse

(Text with EEA relevance)

{SEC(2022) 209 final} - {SWD(2022) 209 final} - {SWD(2022) 210 final}

CSAM: Child Sexual Abuse Material

(c) 'child pornography' means:

(i) any material that visually depicts a child engaged in real or simulated sexually explicit conduct;

(ii) any depiction of the sexual organs of a child for primarily sexual purposes;

(iii) any material that visually depicts any person appearing to be a child engaged in real or simulated sexually explicit conduct or any depiction of the sexual organs of any person appearing to be a child, for primarily sexual purposes; or

(iv) realistic images of a child engaged in sexually explicit conduct or realistic images of the sexual organs of a child, for primarily sexual purposes;

[Art. 2 Directive 2011/93/EU]

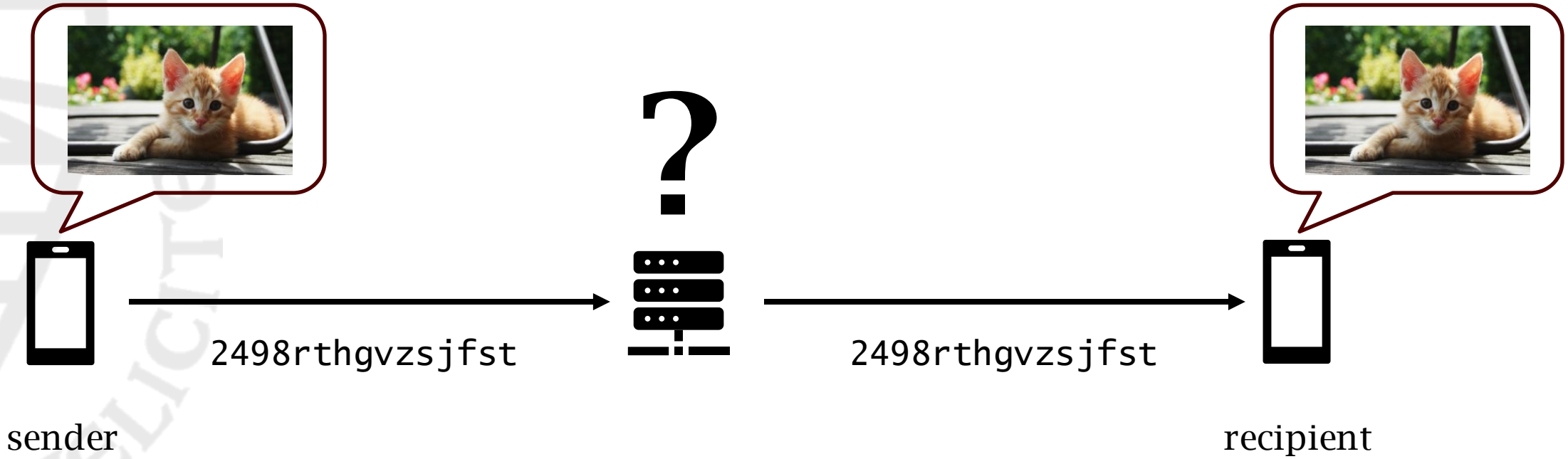
The regulation in a nutshell

- **Replaces temporary derogation to scan CSAM voluntarily**
  - Now extended to 2026
- **Scope**
  - hosting providers
  - providers of interpersonal communication services (Art 2.5 Dir. EU 2018/1972)
    - *Including services where this is only a minor ancilliary feature*
    - *WhatsApp, Zoom, email,…*
- **Risk assessment**
- **Risk mitigation**
  - Content moderation
  - Age verification/assessment
- **Risk reporting**
  - To national authority
- **Blocking orders**
  - Issued by national authority
  - Internet access services
  - URL based
    - *NL law proposal: also DNS block*

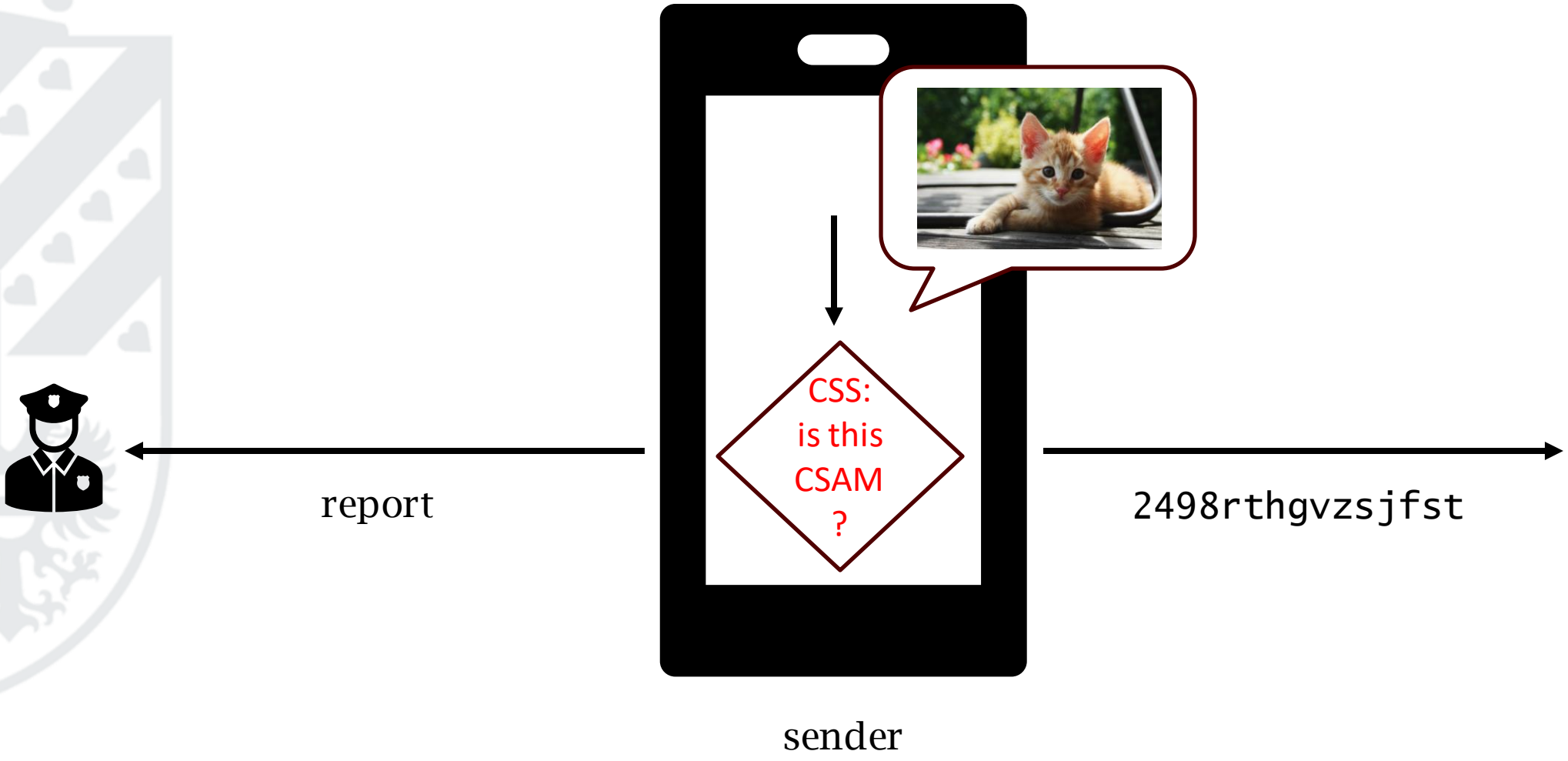
The regulation in a nutshell (continued)

- **Detection orders**
  - Issued by national authority
  - In case of significant risk, outweighing breach of fundamental rights
- **Detection orders**
  - Known CSAM
  - Unknown CSAM
  - Sollicitation of children ('grooming')
  - No longer than 24 m. (for CSAM)
  - Supposedly 'targeted' and 'limited'
  - Based on indicators and technology provided bt EU Centre
- **Indicators (images + URLs)**
  - Maintianed by new EU 'CSAM' Centre
  - Submittted by national Authorities
- **Reports of CSAM**
  - Filed to EU Centre
  - Users only notified after EU Centre decision (or 3 month delay)
  - Anything not manifestly unfounded forwarded to law enforcement
  - Europol has access
  - No strict retention limit
  - All reports (including manifestly unfounded) are kept

The elephant in the room: end-to-end encryption



The “solution”: Client side scanning (CSS)



Different targets and techniques

Target	Technique	Primary issue
Known CSAM	Perceptual hashing	Negligible risk of false positives
Unknown CSAM	AI	Risk of false positives
Grooming (chat)	AI	Significant risk of false positives; age verification required

AI based techniques?

- **Unknown CSAM**
  - 0.1% FAR would be super good (but not yet on the market)
  - WhatsApp: 140 billion msg/day
  - If 1 in 100 tested, still 1.4 million false positives/day
  - E.g. family pictures of grandchildren on the beach, or pictures of
  - *Thorn claims current smartphone hardware cannot even run the detector for all images being sent!*
- **Grooming**
  - Even more contextual
  - Age verificiation required

Perceptual hashing

- **Map image to a short numerical digest/fingerprint**
  - **One-way**: image cannot be reconstructed.
  - **Perceptual equality:** essentially similar images map to the same fingerprint

→ 174960701456

- **Products**
  - PhotoDNA, PDQ, NeuralHash
- **Proprietary**

“Properly” implementing CSS fro known CSAM?

The Apple PSI System

Abhishek Bhowmick, Apple Inc.
Dan Boneh, Stanford University
Steve Myers, Apple Inc.
Kunal Talwar, Apple Inc.
Karl Tarbe, Apple Inc.

July 29, 2021

Abstract

This document describes the constraints that drove the design of the Apple *private set intersection* (PSI) protocol. Apple PSI makes use of a variant of PSI we call *private set intersection with associated data* (PSI-AD), and an extension called *threshold private set intersection with associated data* (tPSI-AD). We describe a protocol that satisfies the constraints, and analyze its security. The context and motivation for the Apple PSI system are described on the main project site.

Cryptographically

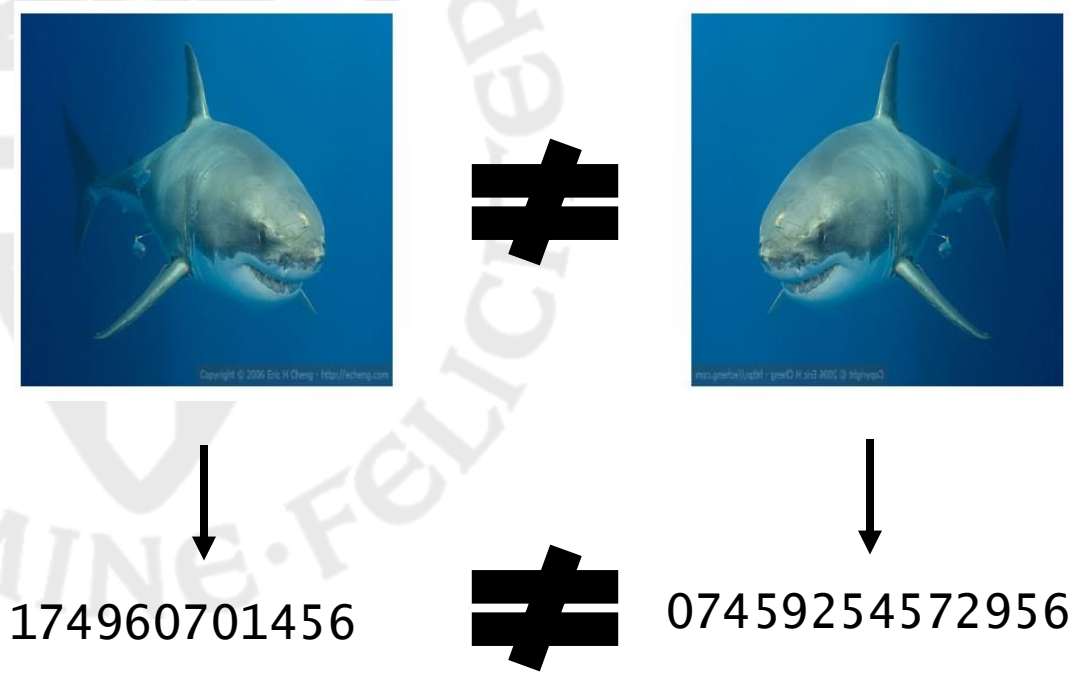
- **Private Set Intersection (PSI)**
  - User has stream of fingerprints that must be matched privately against server database
- **PSI w. associated data (AD)**
  - Server learns data associated for matched fingerprints
- **Threshold PSI-AD**
  - Server only learns associated data when $> t$ fingerprints match
- **Local matching**
  - Blinded copy of server database stored on user device.
  - Only ‘coupons’ sent to server.

General issues with detecting known CSAM

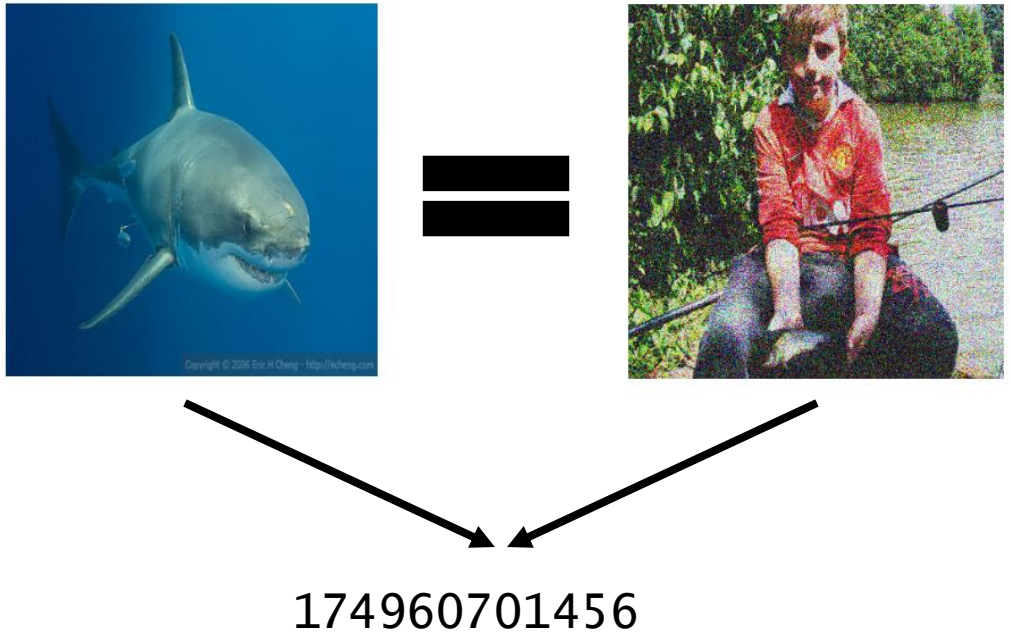
- **CSS has to be installed on all EU phones**
  - Not targeted/limited
  - Perhaps remotely activated
- **Opaque system**
  - Proprietary hashing algorithms (though some revere engineered).
  - Even fingerprints of known CSAM are secret.
  - Independent verification of *what* is scanned is impossible.
  - Service providers have to trust what is given to them.

But also perceptual hashing problems

- **Easy to evade**
  - Rotation, mirroring



- **False positives can be constructed**



Prokos et. al. 2023

So more issues with detecting known CSAM

- **Risk of direct function creep**
  - Terrorist, antisemitic, … pictures can be added to the database by the authorities
  - Can be “prevented” with proper oversight
- **Tainting the database (surreptitious function creep)**
  - Add images that look like CSAM but actually match terrorist or antisemitic material.
    - *Target image A, with fingerprint f(A)*
    - *Generate convincingly looking synthetic CSAM image B*
    - *Tweak B using techniques of Prokos et. al. to generate B’ with f(A)=f(B’)*
    - *Submit B’ to the database*
  - Tainting detected when non-CSAM images get reported
    - Requires proper oversight at EC Centre

Even more issues with detecting known CSAM

- **People can be falsely reported**
  - Attacker somehow obtains fingerprint f(A) of known CSAM
  - Attacker creates cute innocent image B, tweaks it using Prokos techniques so that f(B')=f(A)
  - Sends B' to victim
  - If curte enough, victim forwards it
  - This causes victim to be reported (as fingerprint matches)
  - (And later cleared; but what happens in the meantime?)

- **DDoS on the EC Agency**
  - Activists somehow obtain fingerprint f(A) of known CSAM
  - Create and tweak cute images B such that f(B)=f(A)
  - Send each other these cute images
  - Get reported as fingerprint matches
  - Clogging the reporting pipeline at the EC Agency with false positives

On the reporting pipeline

- **Potentially dealing with**
  - very disturbing images
  - suspects of serious crime
- **Therefore**
  - Tightly secured, with
  - Specially trained personnel

- **Already now unable to keep up with incoming reports**

More fundamental objections

- **End-to-end encryption is a means to an end**
  - CSS breaks confidentiality of correspondence
  - A snitch is watchign while we put our messages in the digital envelope
- **Smartphone = ‘digital home’**
  - CSS creates first law enforcement foothold inside
  - Would we be OK with a webcam in every home to figth domestic violence
- **Mandatory age verification**
  - Restricted access to services; see also eIDAS update.
- **Fighting symptoms instead of actual abuse**
- **Even manifestly unfounded reports are kept**
  - “no smoke without fire”

Fundamental rights assessment of the framework for detection orders under the CSAM proposal

April 2023

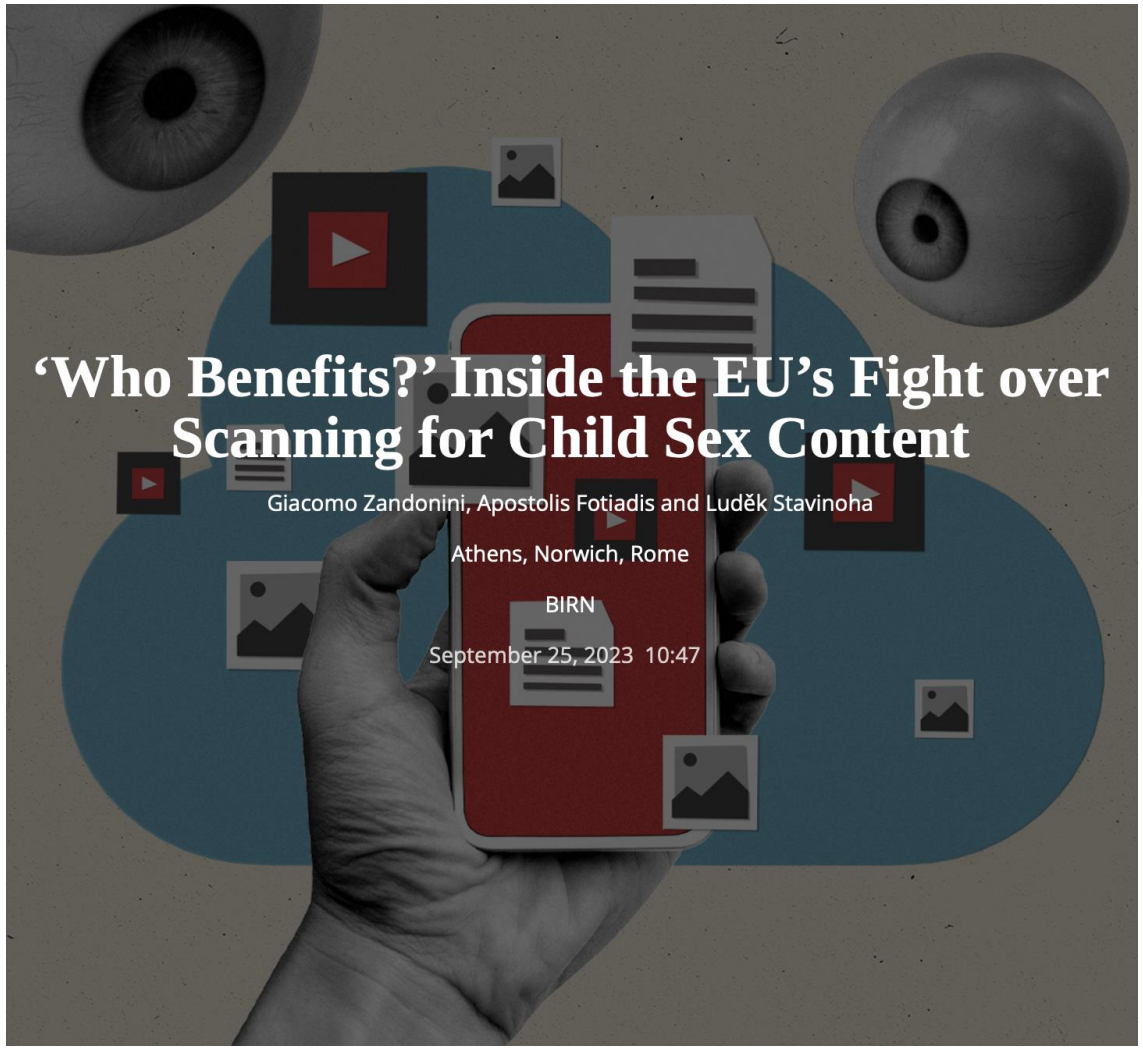
[https://www.ivir.nl/publicaties/download/CSAMreport.pdf]

IViR report (Ot van Daalen) conclusions

- **Detection orders**
  - affect the rights to privacy, data protection and communications freedom under the Charter.
- **Any measure affecting these rights**
  - always must respect the essence of these rights, and be proportionate to the aim of the measure.
  - (unless for national security purposes).
- **Case law Court of Justice of the EU (e.g. re. data retention)**
  - Indiscriminate analysis of confidential communications affects the essence of these rights.
- **Detection orders not proportionate to the aim**
  - Aimed at *services*, not individual people

History and current status

- **Timeline**
  - 11-5-2022: Proposal for regulation
  - 22-3-2023: Motion v. Ginneken
  - 4-7-2023: 1st open letter
  - 26-10-2023: counterproposal EP
  - 27-3-2024: New proposals EU
  - 7-5-2024: 2nd open letter
  - …
- **Dutch government position (until now):**
  - Only known CSAM



[https://balkaninsight.com/2023/09/25/who-benefits-inside-the-eus-fight-over-scanning-for-child-sex-content/]