Deutscher Bundestag
Ausschuss für Digitales

Ausschussdrucksache
zu 20(23)133

17.05.2023

ATHENE
National Research Center
for Applied Cybersecurity

Fraunhofer
SIT

Statement on the hearing "Chat control" in the Committee for Digital Affairs in the German Bundestag on 1 March 2023

February 23, 2023

Prof. Dr.-Ing. Martin Steinebach
martin.steinebach@sit.fraunhofer.de

Fraunhofer Institute for Secure Information Technology SIT, Darmstadt

Head of Department Multimedia Security and IT Forensics Fraunhofer SIT

Honorary Professor Multimedia Security and IT Forensics
Darmstadt University of Technology

Coordinator of the ATHENE research areas
Security and Privacy in Artificial Intelligence (SenPAI) and
Reliable and Verifiable Information through Secure Media (REVISE).

Thank you for inviting me to the Public Hearing on "Chat Control" in the Committee for Digital Affairs in the German Bundestag as an expert. My answers address the technical aspects in the questions to which I can contribute from my experience in developing recognition algorithms. The objective here is to lay a foundation for decisions from the perspective of technical feasibility. I skip questions, which I cannot contribute to from my expertise in this statement.
The basis for my assessments is our study[1] on youth protection, which focusses on sexting and cyber grooming. An updated summary of the study is provided in the essay "Maschinelles Lernen im Jugendschutz"[2] (Machine Learning in Youth Protection). A detailed lecture (in German) on the topic can be found on YouTube[3]. The issue with error rates, which is addressed in the following, is covered in more detail in an essay on upload filters[4].

1) The EU Commission's proposal for the CSA regulation, also known as chat control, has caused much discussion since its publication in May 2022. Please explain the technical, legal, fundamental rights, data protection, social and/or societal implications of the proposal.

The proposal discusses very different measures, some of which pose considerable technical challenges. The proposed measures for detecting evidence of child abuse or its initiation can be used in numerous scenarios. A significant challenge, however, is posed by the required detection mechanisms and the assessment of their practical usability.

The proposal identifies three areas of required detection:
- Detection of **known depictions** of child abuse
- Detection of **new and thus unknown depictions** of child abuse
- Detection of **grooming**, i.e. contact and potential initiation of child abuse.

In practice, these three areas present significantly different technical challenges, which lead to great differences in the expected detection rates. Also, the maturity of the technologies used is not comparable; while the recognition of visual content is now a widespread standard technology, the detection of grooming can still be considered a research topic.

Robust hashes[5] are commonly used when it comes to detecting **known visual content** as they can recognise images even after various operations/modifications such as scaling, adjustments in colour and lossy compression. Compared to conventional cryptographic hashing methods, robust hashing methods clearly outperform in regard to recognising content that is potentially altered easily. Police and internet services use them successfully when adding images to black or search lists. A compact uniform code, the robust hash, is created from an image and stored in a database. This code is based on the content's visual characteristics and not on

[1] https://www.sit.fraunhofer.de/jugendschutz/

[2] https://fsf.de/publikationen/medienarchiv/beitrag/heft/maschinelles-lernen-im-jugendschutz-beitrag-1024/

[3] https://www.youtube.com/watch?v=5ZygUm_KT2k&t=1713s

[4] https://link.springer.com/chapter/10.1007/978-3-658-33306-5_20

[5] Martin Steinebach, Huajian Liu, and York Yannikos. Forbild: Efficient robust image hashing. In *Media Watermarking, Security, and Forensics 2012*, volume 8303, page 83030O. International Society for Optics and Photonics, 2012 oder die Grundlage für Microsoft PhotoDNA, Hany Farid. Reining in online abuses. Technology & Innovation, 19(3):593– 599, 2018.

its digital representation. A block hash for images, for example, calculates whether each of the 256 areas is greater than, equal to or less than the brightness median for a representation of the image, which is reduced to 16x16 pixels, thus generating a 256-bit long binary sequence. Other methods calculate the number of edges in individual sub-areas of an image or use machine learning to determine a hash. If it is then to be determined for an image to be analyzed whether it exists in the database, its code is calculated with the same procedure and searched for in the database. Here, individual errors in the binary representation of the code are accepted, so that the code must only be similar, but not identical. Thus, the tolerance against changes arises, which is not found in a cryptographic hash.

Detecting **unknown visual content**[6] is more challenging. Software must determine whether, for example, an image depicts elements of child abuse. This is typically achieved using supervised machine learning, in which the software is shown numerous examples of child abuse during a training phase and subsequently learns to identify the typical features. Although such methods can still recognise relevant content even when massive changes have been made, errors occur much more frequently than when recognising known content. Recognition can fail if the training data does not contain new variations of content to be recognised. The more diverse the content and characteristics to be recognised, the more difficult it becomes for the software to reliably avoid a false alert, as an ever-increasing variety of image content can be interpreted as an indication of child abuse.

Grooming poses a great challenge, as the texts to be evaluated by the software in order to determine if they contain an indication of grooming[7] are often short. Examples include a request to send a nude photo of oneself or an attempt to arrange a personal meeting. Machine learning, especially Natural Language Processing (NLP), can be used to analyse language and infer meanings of texts or relationships between writers. However, such estimations are difficult to implement due to the diversity of expressions, language abilities and conversational topics, especially if the perpetrator's approach is not straightforward. Some approaches aim to deduce the writer's age from the texts to identify whether a person is lying about their age. Some simple methods only search for precise key words or phrases, but this can easily be circumvented by avoiding these phrases.

From a technical point of view, given the detection methods' very different levels of maturity, it is important to consider them separately. The detection rates and behaviour of the methods, especially the delimitation of similar cases which are not relevant, are not comparable. Robust hash methods can distinguish very well between visually similar but not identical image content. Procedures for evaluating unknown content have greater difficulties in this respect for technical reasons.

From an IT security perspective, it should also be noted that the recognition procedures are usually not "secure" as defined by IT security research. This means that they are not designed to provide security against attacks. Methods for recognising known and unknown representations originate from signal processing and are evaluated in terms of their recognition performance. This evaluation is carried out using typical content examples, in many cases with standardised data

[6] Abhishek Gangwar, Eduardo Fidalgo, Enrique Alegre, and Víctor González-Castro. Pornography and child sexual abuse detection in in image and video: A comparative evaluation. 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017)

[7] Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online working notes/labs/workshop)*, Volume 30, 2012

sets. In IT security, the procedures would also be examined in terms of how they behave in the event of targeted attacks. For example, with cryptographic hashes, a key property is how difficult it is to generate data with a given hash.

However, in the context of a security application such as this one, it can be assumed that content is deliberately altered by attackers in such a way that it shows relevant material, but that this is not recognised by an automated system. Such attacks are known for both robust hash methods and machine learning. Attacks in the opposite direction can be just as relevant: Content is altered in such a way that it shows apparently harmless content but is mistakenly considered relevant by a system. False negative and false positive classifications are thus brought about by attacks. This in turn allows content that should be recognised to be disseminated unnoticed, or for content extractions to be forced.

2) The Commission's proposal foresees that detection orders will be issued, which will lead to providers of communication services or devices having to covertly leak information if there is a suspicion that abusive material is being exchanged via these services or devices or that grooming is taking place on them. In your view, which services and devices are potentially affected by this and to what extent, and what impact will this have on their users?

The proposal requires the integration of detectors in the service providers' apps on all devices with which the services can be used – primarily **smartphones, tablets and computers**. Communication services, such as Messenger, are affected. The reach extends to all users of hosting services and interpersonal communication services within the proposal's scope if content is to be reliably detected as described in the proposal.

Since messengers, for example, are mostly proprietary and detection is built into the app, every participant in the communication will automatically be subject to the investigation. And since it cannot be assumed that individual channels can be effectively classified as unsuspicious, all communication must be examined. This means that every message sent, or every sequence of messages must be examined for grooming. For every image sent or received via a service, it is assessed whether it is a known image depicting child abuse or whether the image contains signs of child abuse.

The guidelines' impact depends on how the decision about a suspicion is made. If a single incident is enough to trigger an information deflection, it follows that the deflection is directly dependent on the recognition rates of the detection procedures. A single false positive alert for an individual image leads to the image being routed out to the EU centre (see article 40). A high volume of communication from a user statistically increases the risk of being subjected to an erroneous deflection. The problem can be mitigated by a "sluggish" behaviour of the deflection. There would be no immediate reaction, but a counter would be incremented as soon as a detector classifies a content as relevant. Only when the counter reaches a predefined threshold content will be redirected. This counter should be sufficiently high and consider the detector's false positive rate (FPR). Assuming an FPR of one per thousand and 10,000 messages received, the probability of starting a false rejection at a threshold of 10 is over 50%, i.e. triggering at least 10 false alarms among the 10,000.

4) How do you assess the risk of innocent citizens coming under suspicion through false positive automated detection, and what would be the impact of such false positives on both suspects and investigating authorities?

It must be assumed that citizens will at least be subject to data leakage from their devices to the EU centre, especially in the case of detecting previously unknown content and grooming. The likelihood of misinterpretation depends on the error rates of the system used and the amount of content examined. Frequent users of services are more likely to be subject to leakage than those who receive or share content sporadically. When using methods to detect previously unknown content, it can be assumed that the probability of a false positive alert increases with the proximity to content that is to be detected. With an age limit of 18 (under Article 2i), a consumer of legal erotic content featuring young performers is more likely to be the subject of a false positive report than a person who frequently views landscape images.

A manual screening is carried out at the EU centre, which in the case of a false positive report should cease suspicion. However, it is difficult to estimate how reliable this screening can be. Depending on the error rates of the detection systems used in combination with the large number of messages being searched, weariness can cause another error source. A double misjudgment by the system and the centre results in a referral to the investigating authority. The extent to which this then causes a sharp increase in the workload of the investigating authorities depends on the probability of both misjudgments occurring together.

Furthermore, it should be noted that regardless of the averted suspicion in the case of a false positive report, a breach of privacy occurs, since content that is transmitted confidentially and protected is viewed in the centre after it has been leaked. Furthermore, it is important how providers of hosting services and providers of interpersonal communication services proceed with their users after the report, i.e. whether they continue to offer or block services until the allegations have been reviewed. In the event of a block, even if only temporary, false positive reports could have a significant impact on the affected users.

5) Hosting service providers and interpersonal communication service providers who have received a discovery order are required by Article 10 CSAM-E to install and operate technologies that detect contact with children with intent to abuse ("grooming"). Are you aware of technologies that can reliably distinguish between innocuous, sexually or romantically charged communication and grooming?

There are several scientific papers in the field of NLP (natural language processing) that address the detection of grooming[8]. The recognition rates here range from 80 to 90 percent. As stated in the answer to question 2, this can lead to a great number of false positive reports. The problem here is, among other things, to obtain a data basis for training a network and evaluating its reliability that contains corresponding conversations in a large number and covers many cases and procedures. The risk with a small data base, which may also come from a specific content context, is that the results are not transferable to other cases and scenarios. A forum in which examples have been collected and which is about sports may

[8] Fabián Muñoz, Gustavo Isaza, and Luis Castillo. Smartsec4cop: smart cyber-grooming detection using natural language processing and convolutional neural networks. In International Symposium on Distributed Computing and Artificial Intelligence, pages 11–20. Springer, 2020.

provide data by interspersing statements about sports, which later provide for strongly diverging results in a forum about animal welfare. For example, if in the sports forum the perpetrator lures with having their own training facility to attract children, then this approach will not occur in the animal welfare forum and will potentially lead to errors. It should also be noted that the chances of success of NLP solutions are language specific. Communication in a widely used language can most likely be investigated more reliably than in a little-used language.

In addition to detecting grooming via content, it is also feasible to detect discrepancies between a given and an inferred age[9]. This is done by profiling an author's written messages to infer his or her age (also by NLP) and then comparing it with the given age. On average, these methods can determine an age accurately to within a few years, but they also generate errors. In our own experiments, deviations of up to 28 years were detected with a median of 4.6 years.
A weaker form of recognition is one using keywords and the recognition of links. Here, no network is trained using examples, but experts draw up lists of terms or phrases that are typical of grooming and can be an indicator, especially if they occur frequently. A disadvantage is that these terms can be specifically avoided if they become known. It is also possible to recognise embedded links within communication that lead to private channels.

6) What technical approaches do you consider to be effective alternatives to the measures envisaged in the draft regulation that do not raise fundamental rights concerns?

Without being able to give an assessment of the legal implications, **detecting and blocking content at the transmitter** in the case of interpersonal communication or at the process of uploading in the case of hosters is a technically feasible alternative that can generally be based on the same detection methods. Blocking would prevent or at least make it more difficult to disseminate problematic content and grooming without necessarily having to report the finding to third parties. False positives then solely prompt a refusal of dissemination by the system in question that is not traceable to the user. The technical disadvantage of this approach is that it creates a so-called "oracle" with which an attacker can learn to circumvent the detection.

10) In your view, which political package of measures is holistically promising in order to take action against sexualised violence against children in an efficient, effective and fundamental rights-compliant manner - where is there potential for follow-up and improvement in the area of prevention and in combating sexualised violence and its portrayal on the internet?

On the one hand, a basis must be created to efficiently share findings about known content with all parties involved (at least providers of communication services, hosters, police authorities). Any reliably detected content that can be recognised makes the attempt to recognise an unknown content unnecessary and thus reduces error rates. According to the state of the art, this can only be achieved by a central collection of all recognised content, from which robust hashes (or indicators in one of their possible manifestations) can then be calculated. This can be done at the EU centre to be established as proposed. Since all parties involved are free to decide on the technical implementation according to Article 10 in the proposal, it must be

[9] https://www.sit.fraunhofer.de/jugendschutz/

possible at this centre to apply all detection methods to the central collection, for example to calculate the respective variants of the robust hashes. It is not possible to convert the hashes or indicators of the different methods. This means that from a hash or indicator calculated by one entity using a given method, another entity cannot determine its own hash or indicator using its method. The hashes or indicators must be recalculated from the collected raw data for each method.

On the other hand, it should be noted that the recognition of unknown content and grooming still requires research. Research in this context is difficult and requires the development of an appropriate strategy on how to access sensitive content without further burdening victims and researchers. A central operation of a system in which researchers can train and evaluate their procedures without access to the data is also feasible.

In general, it can be observed that there is still a great need to harmonise the level of knowledge of all involved regarding the technical feasibility of the different detection variants. Here, politics can help initiate an exchange. For one, the scientific disciplines involved must jointly develop a state of the art and the requirements, and for another, this state of the art must also be communicated to participants such as those mentioned above.
Since the assessment of feasibility and impact depends heavily on error rates, especially the false positive rates of the methods, it is necessary to develop as standardised an approach as possible for obtaining the key figures. This must be based on realistic data and take into account which content is usually disseminated via the channels under investigation and then compare these in their behaviour with the content to be detected. The expected number of contents investigated and to be detected must also be considered. The different methods must be fully represented in terms of their detection rates. This requires a complete confusion matrix with a gauge of correct and incorrect assignments of positive or negative cases (i.e. false positive, false negative, true positive and true negative). Only then can a reliable estimate be made on the behaviour in practice.