This project has received funding from the European Union’s Horizon 2020 - Research and Innovation Framework Programme, H2020 SU-FCT-2019, under grant agreement no 883341.

Global Response Against Child Exploitation

Instrument:	Research and Innovation Action proposal
Thematic Priority:	FCT-02-2019

Review Mechanism and Procedure

Deliverable number	D9.9	
Version:	1.0	
Delivery date:	July 2023	
Dissemination level:	PU	
Classification level:	Non classified	
Status	Final	
Nature:	Report	
Main author(s):	Christiana Marcou Thalia Prastitou Merdi	EUC
Contributor(s):		

DOCUMENT CONTROL

Version	Date	Author(s)	Change(s)
0.5	06/06/2023	Christiana Markou Thalia Prastitou Merdi	Submission of v0.5 of this document:
0.7	28/06/2023	Christiana Markou Thalia Prastitou Merdi	Submission of v0.7 of this document:
0.9	17/07/2023	Christiana Markou Thalia Prastitou Merdi	Submission of v0.9 of this document:
0.9	19/07/2023	Gary Ellis (SAB)	SAB assessment with no classification required
1.0	28/07/2023	Peter Leskovsky	Final check before submission.

DISCLAIMER

Every effort has been made to ensure that all statements and information contained herein are accurate; however, the Partners accept no liability for any error or omission in the same.

This document reflects only the view of its authors and the European Commission is not responsible for any use that may be made of the information it contains.

Table of Contents

1. Introduction 4
1.1. Overview 4
1.2. Approach 4
1.3. Relation to other deliverables 4
1.4. Structure of the deliverable 5
2. Checklist 7
2.1. Access control 7
2.2. Audit trail 8
2.3. Data Protection 9
2.4. Database Search 11
2.5. Human Oversight 13
2.6. Measures Against Re-Victimization and Over-Exposure 14
2.7. Technical Robustness and Safety 15
2.8. Transparency 15
2.8.1. *Traceability* 15
2.8.2. *Explicability* 16
2.8.3. *Open Communication* 17
2.9. Unfair biases 18
2.10. Electronic Evidence 19
2.11. Use of Crawlers 21
3. Conclusion 24
3.1. Summary 24
3.2. Evaluation 24

1. Introduction

1.1. Overview

The DoA describes this deliverable as:

D9.9 – Review Mechanism and Procedure. This deliverable provides the results of the process of developing the review mechanism as well as the final review procedure and mechanism consisting of the check list, guidelines and instructions. Related task(s): T9.5. (Month 38).

The description of the related Task T9.5 provides the following details:

T9.5 - A review procedure will be developed with the intention to confirm that future national operation of the GRACE capability will not be circumventing the implemented safeguards and that it will be in full compliance with the legal and ethical rules and recommendations analysed in previous tasks. More specifically, the work carried out under T9.2, T9.3 and T9.4 will be subjected to analysis intended to give rise to the basic rules and principles, compliance with which national LE authorities will have periodically to check, verify and confirm during the operation cycle of GRACE. The relevant procedure and mechanism will consist of a specifically-designed check list which will have to be gone through by the reviewer accompanied by a set of guidelines and instructions guiding the relevant party responsible for the review as how when and how they should go through it as well as how to deal with any arising issues, thus ensuring that GRACE operation is in accord with all rules, regulations and ethical principles. This review procedure and mechanism will be constructed towards the end of the project.

The main objective of this Deliverable D9.9 is to provide to LEA officers, working in the area of child sexual exploitation, that will be using the GRACE system, a review procedure, compliance with which, national LEAs will have periodically to check, verify and confirm during the operation cycle of GRACE. Apart from the correct and legitimate operation of the GRACE system this review procedure will guide LEAs as to how to deal with any arising issues, thus ensuring that the GRACE operation is in accordance with all legal rules, regulations and ethical principles found in this area of law.

1.2. Approach

In practice, the review mechanism is divided in eleven thematic sections each consisting of two main parts, a) a specifically-designed check list, which has to be gone through by the reviewer and b) a set of guidelines and instructions, which actually precedes each check list, guiding the relevant party responsible for the review as to why and how these issues are important and needed to be assessed. These eleven sections concern the issues of access control, audit trail, data protection, database search, human oversight, measures against re-victimization and over-exposure, technical robustness and safety, transparency (traceability, explicability, open communication), unfair bias, electronic evidence and the use of crawlers.

1.3. Relation to other deliverables

This deliverable is related to the following other GRACE deliverables:

Receives inputs from:

Deliv. #	Deliverable title	How the two deliverables are related
Deliv. #	D9.1 Ethical report v1	D9.9 employs information found in D9.1 in order to develop its checklist and related instructions
Deliv. #	D9.3 Legal report v1	D9.9 employs information found in D9.3 in order to develop its checklist and related instructions
Deliv. #	D9.5 Overall legal and ethical framework v1	D9.9 employs information found in D9.5 in order to develop its checklist and related instructions
Deliv. #	D9.7 Architecture for technical safeguards – "security and privacy by design" v1	D9.9 employs information found in D9.7 in order to develop its checklist and related instructions

Table 1 – Relation to other deliverables – receives inputs from

1.4. Structure of the deliverable

This document includes the following sections / the information in this document is structured as follows:

- Section 2: In this section the review mechanism is found. In practice, this section is divided in eleven thematic subsections each consisting of two main parts, a) a specifically-designed check list, which has to be gone through by the reviewer and b) a set of guidelines and instructions, which actually precedes each check list, guiding the relevant party responsible for the review as to why and how these issues are important and needed to be assessed. These eleven subsections concern the following issues

    - Subsection 2.1 focuses on guidance, instructions and questions for LEAs regarding the issue of access control. Access control is a method of allowing access to a platform's, a database's or a system's, sensitive data so that solely authorized users are allowed access to such data and, furthermore that such access is restricted for unauthorized users.

    - Subsection 2.2 focuses on guidance, instructions and questions for LEAs regarding the issue of audit trail. Auditing can be characterised as the main mechanism with which compliance is monitored. In practice, there are two main objects that are needed, in order to perform reliable and meaningful auditing: there are audit logs and audit trails.

    - Subsection 2.3 focuses on guidance, instructions and questions for LEAs regarding the issue of data protection. In practice, in the EU, data protection, in the police and criminal justice sector is regulated in the context of both national and cross-border processing by police and criminal justice authorities of the Member States and EU actors.

    - Subsection 2.4 focuses on guidance, instructions and questions for LEAs regarding the issue of database search. At this point, important information is provided regarding access to specific data and allocated privileges within the GRACE system.

    - Subsection 2.5 focuses on guidance, instructions and questions for LEAs regarding the issue of human oversight. Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects.

- Subsection 2.6 focuses on guidance, instructions and questions for LEAs regarding measures against re-victimization and over-exposure, provided that the safety of both the victims and themselves is an issue of fundamental importance.
- Subsection 2.7 focuses on guidance, instructions and questions for LEAs regarding the issue of technical robustness and safety. Technical robustness requires AI systems to be developed with a preventative approach to risks and in a manner such that it reliably behaves as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm.
- Subsection 2.8 focuses on guidance, instructions and questions for LEAs regarding the issue of transparency and focusing on the topics of traceability, explicability and open communication.
- Subsection 2.9 focuses on guidance, instructions and questions for LEAs regarding the issue of unfair bias. AI, or algorithmic, bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as favouring one arbitrary group of users over others.
- Subsection 2.10 focuses on guidance, instructions and questions for LEAs regarding the issue of electronic evidence. This is another issue of fundamental importance for LEAs as electronic data could potentially be useful for law enforcement in crime investigation.
- Subsection 2.11 focuses on guidance, instructions and questions for LEAs regarding the use of crawlers. In practice, crawling is the process of exploring web applications automatically. The web crawler aims at discovering the web pages of a web application by navigating through the application.

2. Checklist

2.1. Access control

2.1.1. Guidance and Instructions

Access control is a method of allowing access to a platform's, a database's or a system's, sensitive data[1] so that solely authorized users are allowed access to such data and, furthermore that such access is restricted for unauthorized users. In practice, there are various methods of controlling and restricting the access of users to a database which are considered best practices. These are, *inter alia,* implementing proper authentication and authorization mechanisms, account management and session management. In detail, account management and user authentication, can be seen as the two main components of a database solution, and they, therefore, need to follow security best practices.

Initially, access to GRACE must be available only after authentication. Furthermore, application level authorisation must be applied to ensure GRACE functionality and data access shall be restricted in line with the authorisation rules that will be agreed. In addition, there must be a clear process where new user accounts will be created. All users will, then need to make use of these personal accounts in order to access the GRACE platform. All unnecessary user accounts shall be removed. A password policy must be additionally applied. In detail, authentication credentials should be sufficient to withstand attacks that are typical of the threats in the deployed environment (for example requiring the use of alphabetic -at least one upper-case letter and one lower-case letter - as well as numeric and/or special characters). Furthermore, the allocation of privileges to users and applications to that required to perform the business function shall be managed and restricted accordingly. These users' allocated privileges shall be reviewed periodically and, in case some users' privileges are no longer needed they shall be removed. Lastly, proper configuration of session parameters across all components (such as time-out, session id generation, auditing) shall be safeguarded.

2.1.2. Questions

- Is access to GRACE components possible only after authentication?
- Are there in place specified authorization rules governing access to GRACE functionality and data?
- Are these authorization rules enforced?
- Is there a clear process where new user accounts are created?
- Is there a rule in place and enforced dictating that all users must only access GRACE through their own personal user account?
- Are unnecessary user accounts swiftly removed?
- Is there in place and technically enforced a password policy sufficient to withstand attacks that

[1] The following personal data is considered 'sensitive' and is subject to specific processing conditions: personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs; trade-union membership; genetic data, biometric data processed solely to identify a human being; health-related data; data concerning a person's sex life or sexual orientation.

are typical of the threats in the deployed environment (e.g., requiring the use of alphabetic as well as numeric and/or special characters)?

- Are authorized users technically able to see and do strictly only what is necessary to perform their tasks (restrictions on allocated privileges)?
- Are allocated user privileges periodically reviewed so that privileges no longer needed are revoked?
- Is proper configuration of session parameters across all components (timeout, session id generation, auditing, etc) ensured?

2.2. Audit trail

2.2.1. Guidance and Instructions

Auditing can be characterised as the main mechanism with which compliance is monitored. In practice, there are two main objects that are needed, in order to perform reliable and meaningful auditing: there are *audit logs* and *audit trails*.[2]

In detail, all user actions need to be audited. *Audit logs* are a chronological record of activities performed on a specific technical application implementing the legal concept laid down in Regulation (EU) 2016/794[3] and its national implementing legislation. Any application used to process personal data shall log activities related to the access of data that it controls. All the accesses and the operations that can be done over stored resources in GRACE need to be tracked.

An *audit trail* is a chronological record of technical components allowing the reconstruction and examination of the sequence of activities surrounding or leading to a specific operation, procedure or event in a transaction from inception to final result. In other words, its the process of taking, seizure, access, processing, transport and storage of evidence that must be recorded. If done right, an independent third party shall be able to examine those processes and achieve the same result.

More specifically, it shall be possible to ascertain from *audit logs* specific minimum information, including, the identification of the user, the date and time of the event and its outcome as well as log-on and log-off attempts to the application and their outcome. Similarly, it must be possible to ascertain from the *audit trails*, additional minimum information, including, the network equipment used to transmit data (such as proxies, routers, firewalls, etc), the user name or unique identifier and the date and time of event.

2.2.2. Questions

- Are all user actions on GRACE logged and audited?

[2] Based on Europol's Policy on the Control of Retrievals (EDOC #893185), which sets out the requirements on how every user action which accesses personal data shall be logged and audited.

[3] Regulation (EU) 2016/794 of the European Parliament and of the Council of 11 May 2016 on the European Union Agency for Law Enforcement Cooperation (Europol) and replacing and repealing Council Decisions 2009/371/JHA, 2009/934/JHA, 2009/935/JHA, 2009/936/JHA and 2009/968/JHA, Official Journal of the EU, 24 May 2016, L 135/53, as amended by Regulation (EU) 2022/991 of the European Parliament and of the Council of 8 June 2022, as regards Europol's cooperation with private parties, the processing of personal data by Europol in support of criminal investigations, and Europol's role in research and innovation.

- Are all the accesses and the operations that can be done over stored resources in GRACE tracked?
- If yes, is it possible to ascertain from *audit logs* the following information:
    - A unique reference number related to the retrieval or the attempted retrieval;
    - Which of the components of the information processing activities referred to in Chapter IV of Regulation (EU) 2016/794 as amended are accessed or consulted;
    - The identification of the user, such as User Name or Unique Identifier;
    - The date and time of the event and its outcome (retrieval, consultation, attempted retrievals, modification, attempted modification, deletion, attempted deletion, etc);
    - Object content being accessed, including the identity of the person or persons concerning whom data were queried or accessed and displayed or the identification of the record retrieved;
    - Trace of changes performed in the accessed object;
    - Device address or other logical location indicator of the source of the request;
    - Log-on and log-off attempts to the application and their outcome

Are the following services, actions or information also logged?

- Authentication services used to access the system
- Network equipment used to transmit data (such as proxies, routers, firewalls)

  For any technical components participating in a transaction linked to retrieval of information, where possible:
- Username or Unique Identifier (in case the object logged is the Unique Identifier, it shall be possible to determine the identity of the User in a simple and reasonable way)
- Date and time of event;
- Device address or other logical location indicator of the source of the request and the final destination of the request (including port and protocol if relevant);
- The specific request of the user;
- Any actions taken on the request;
- Any replies provided to the user;

- Changes to the user accounts allowing access to the technical component and its configuration files;
- Changes to files or directory permissions on the technical component;
- Changes on the configuration files of the technical component;
- Logon and logoff attempts to the management console or application used to manage the technical component and their outcome.

2.3. Data Protection

2.3.1. Guidance and Instructions

Within the EU, data protection in the police and criminal justice sector is regulated in the context of both national and cross-border processing by police and criminal justice authorities of the Member States and EU actors. The central instrument at EU level, is Directive (EU) 2016/680[4] which aims to protect personal data collected and processed for criminal justice purposes including prevention, investigation, detection or prosecution of criminal offences.

In detail, the end user should record the processing activities that process personal data, keeping information like who processed what, when and why. When the end user requires the use of personal data for law enforcement purposes, what needs to be recorded and documented is the following: the origin of the personal data, the method of acquisition, as well as the fact that they can be processed for law enforcement purposes only. Only authorised personnel for authorized purposes should be able to process data which are intended for law enforcement purposes only. Processing special categories of personal data (revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, data concerning health or data concerning a person's sex life or sexual orientation) should take place only when strictly necessary. Furthermore, according to the *data protection by default principle* the end user, should ensure that only data strictly necessary for each specific purpose of the processing are processed by default (without the intervention of the user). Each personal data piece should be linked to the data subject it belongs to and it should be trackable, regardless of how many databases/file systems it has been stored in. The end user should have the capability to trace, extract and delete all personal data belonging to a data subject, in case such requests come from the data subject itself, from supervisory authorities, or emanate from data retention obligations. Any action which processes personal data should be logged. Log information should include the user or service which processed the data, the purpose, as well as the date and time. Every data piece should be accompanied by a “retention evaluation” or “retention expiration” date, and the GRACE platform should be able to inform users of upcoming data retention timelines in due time. In cases where personal data are used for law enforcement purposes, data subjects should clearly be classified as victims, suspects, informants, etc. The GRACE platform should make it clear to users as to which pieces of information have been inferred by automated processing activities, and which of those pieces have been confirmed by users. Any data breach should be communicated to the competent authority and every end user should be aware of the authority to which he/she has to notify such a potential data breach. A process should be additionally created for such notification to be provided to the competent authority in a timely and concise manner. One should also consider, if the type of processing activities he/she is going to perform guarantees the use of a Data Protection Impact Assessment (DPIA), before this processing action takes place. Lastly, before the end user processes personal data for any purpose, the applicability of a DPIA or prior consultation should be clear, and the relevant communication and decisions should be documented.

2.3.2. Questions

- Do you maintain a record of processing activities?

[4] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ L 119.

- If yes, is it periodically reviewed and when necessary updated?
- Are (i) the origin of the personal data, (ii) the method of acquisition and (iii) the fact that said personal data can be processed for law enforcement purposes recorded and documented?
- Is personal data only and strictly accessible to authorised personnel for law enforcement purposes only?
- Is personal data, especially personal data of special categories (e.g. religion, sexual orientation) processed only if and when strictly necessary for a law enforcement purpose?
- Do you ensure that only data strictly necessary for the specific purpose is processed by default, i.e., without user intervention?
- Is each personal data piece linked to the data subject it belongs to and is trackable, regardless of how many databases/file systems it has been stored in?
- Is any action which processes personal data logged?
- If yes, does log information include the user or service which processed the data, the purpose, as well as the date and time?
- Is every data piece accompanied by a “retention evaluation” or “retention expiration” date?
- Do you ensure that notifications of upcoming data retention timelines communicated by the system are acted upon on time and relevant data is deleted or rendered anonymous?
- Is personal data used for law enforcement purposes classified by reference to the data subject category to which it relates, i.e., victims, suspects, informants, etc?
- Is data generated by inferences of AI tools distinguishable from data which have been triaged and confirmed by humans?
- Are you fully aware of the data protection authority to which data breaches, if any, should be notified?
- Is a sufficiently clear and detailed process in place in accordance with which data breach notifications shall be made in a timely and concise manner?
- Have you ensured that a Data Protection Impact Assessment (DPIA) has been performed for the processing to take place through the GRACE system?
- Prior to start using the GRACE system, have you considered, with the assistance of your DPO, whether prior consultation with the Data Protection Authority is merited with regard to the data processing to be performed through the GRACE system?
- Have any communications and decisions related to prior consultation been documented?

2.4. Database Search

2.4.1. Guidance and Instructions

Information Security is a principle which addresses the security of information available in software solutions, by, among others, mitigating risks such as unauthorized access or use of data, introducing *security-by-design principles* in development or hosting, and addressing security incidents, when they arise. Although Information Security addresses more than just database solutions, there are nevertheless many useful recommendations which are applicable to database solutions. While, generic procedural issues regarding granting access to specific users within the GRACE system have been dealt with above,[5] specific guidelines regarding access to specific data / allocated privileges within the GRACE system/database shall be given here. Specifically, end users should only have access to data which they need to know, regardless of their role in their security clearance. The allocation of privileges must be constrained to users and applications to that required to perform the specific business function. End users must review the allocated privileges periodically and revoke privileges which are no longer needed. The environment on which GRACE will be deployed should follow proper security principles (including secure firewall configuration, network segmentation guidelines, etc.). GRACE's deployment should be accompanied by a) a process of responding to security incidents (data breaches, data leaks, etc.), b) process of recovering from disasters (hardware failure, software failure, loss of data, etc.). Encrypting a database, means that data are encrypted when they are stored in the database, and need to be decrypted when an authorized user or service wants to access them. This is particularly important for the GRACE system, considering the sensitivity and nature of the data that will be stored. Creating database backups is a common task when working with databases, for reasons ranging from testing to disaster recovery. Any database backup (same as the operational database itself) and every connection to a database shall be encrypted. Furthermore, monitoring and analysing the activity of the GRACE system is a good practice which provides operations personnel with valuable performance and security information of a database. Database hardening and database access auditing are another two issues that should be taken into account. In detail, the first, relates to the process of mitigating security risks by analysing the security vulnerabilities of a database and implementing security best practices and processes, while the second one is a security best practice, which refers to the monitoring of data which have been accessed by which user/service. Lastly, robust change management processes should be implemented to handle GRACE changes and the GRACE software should be patched regularly.

2.4.2. Questions

- Do you ensure that every user, regardless of role or security clearance, only has access to the database connected with the GRACE system and/or any parts thereof, only if and when strictly necessary for the performance of his/her duties?
- Has it been ensured that every user can only perform on the system the operations (such as storing, retrieval and editing) which are strictly necessary for the performance of his/her duties?
- Are allocated user privileges periodically reviewed so that privileges no longer needed are revoked?
- Has it been ensured that the environment on which the database is deployed follows basic security principles, such as secure firewall configuration and network segmentation guidelines?
- Is there in a place a process of responding to security incidents, namely data breaches?
- Is there in place a process of recovering from disasters (hardware failure, software failure, loss of data, etc) documented in a relevant disaster discovery plan?
- Has it been ensured that data in the database are encrypted when they are stored in the database, and decrypted when an authorized user or service wants to access them? (Database encryption)
- Has it been ensured that all database backups are encrypted? (Backup encryption)

[5] See section 2.1. Access Control

- Has it been ensured that every connection to the GRACE database is encrypted? (Secure connections - TLS)
- Are the stored procedures and database views constantly monitored and revised? (Usage of stored procedures or database views)
- Do you engage in systematic monitoring and analysing of the activity of the GRACE database so as to receive valuable performance and security information of said database? (Database Activity Monitoring solutions (DAM))
- Do you implement security best practices and processes based on any security vulnerabilities that may become evident from said monitoring and analysis? (Hardening)
- Do you ensure that you monitor which data have been accessed by which user/service? (Database access auditing)
- Are there in place change management processes to handle database changes? (Change management)
- Is the database software regularly patched? (regular patching)

2.5. Human Oversight

2.5.1. Guidance and Instructions

Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system.[6]

The purpose of the GRACE system is to serve flagging CSEM reports according to their priority and leaving the ultimate decision about which CSEM report is investigated to be made by the end-user. The explicability of the GRACE system hinges on the ability to explain to the end-user both the technical processes of the GRACE tools and platform and the reasoning behind the decisions or predictions that the GRACE system suggests. The end-user, can only build and maintain trust in the GRACE system, if he/she understands the AI driven decisions of the GRACE tools and platform. This understanding needs to put him/her in the position to contest as well as to identify ethically unacceptable considerations used by an AI-generated decision. Furthermore, the end users need to be continuously surveyed as to whether and how they understand the decision(s) of the GRACE tools and platform. Lastly, comprehensive documentation of the education and specific training required for anyone involved in a governance mechanism (HITL, HOTL, HIC) concerning a GRACE tool, the GRACE platform or the GRACE system as whole, shall be provided so that any end-user can exercise meaningful oversight.

[6] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

2.5.2. Questions

- Can all users understand in non-technical terms all algorithmic decisions of the GRACE system?
- Are all users able to understand what elements used in the (machine) learning model were responsible for each specific outcome?
- Are users continuously surveyed whether and how they understand the decision(s) of the GRACE tools and platform?
- Have users received appropriate training on the GRACE tools, the GRACE platform and the GRACE system as a whole so that they can exercise *meaningful* oversight of the operations and outputs of the system and/or tools?

2.6. Measures Against Re-Victimization and Over-Exposure

2.6.1. Guidance and Instructions

The safety of both the victims as well as LEAs members, being the end users, is an important issue that shall be taken, additionally, intro consideration. The GRACE platform supports and harmonises the necessary evaluation of CSEM reports for police investigations. Yet, the final evaluation of a CSEM report's content data has to be carried out by the individual end-users. Therefore, the GRACE system has the potential to equally affect the end-user's physical and mental well-being by exposing the human end-user to the content data of CSEM reports. The development and integration of mechanisms which ensure that the GRACE system monitors and regulates the number of times the content data of a CSEM report are accessed by a human LEA officer should have been considered so far. In terms of staff members working with CSEM, sufficient psychological relief and support, should be available, both at a collegial and professional level. As such the GRACE system should keep track of the number of images or videos that a certain user is exposed too, as well as their duration. A suitable detection and response mechanisms for undesirable adverse effects of the GRACE system for the end-user should have been so far both developed and established. Such mechanisms should include something like a 'stop button' or a procedure to safely abort an operation when needed. Furthermore, comprehensive documentation of the education and specific training required for any human end-user involved in a governance mechanism (HITL, HOTL, HIC – explained above) concerning a GRACE tool, the GRACE platform or the GRACE system as whole, should have additionally been established so far, so that human end users can now easily exercise meaningful oversight. Any accidental or purposely generated access or data leak should be avoided at all cost. As such, the use of independent overall system and administration monitoring or oversight should have been also considered. Lastly, it is indisputable that, the final evaluation of an CSEM report and its content data has to be carried out by a human officer, being the individual end-user.

2.6.2. Questions

- Are there measures in place, such as a mechanism that monitors and regulates the number of times an CSEM report is accessed by a user, to safeguard individual users' physical and mental well-being against the risk inherent in the exposure to the content data of CSEM reports?
- Is psychological relief and support organized at a collegial and professional level provided to

users?

- Do you ensure that the detection and response mechanisms for undesirable adverse effects of the GRACE system for the user, such as a 'stop button' or a procedure to safely abort an operation when needed, are always enabled?
- Are LEAs sufficiently trained and encouraged to use the aforementioned detection and response mechanisms?
- Are all information security requirements complied with to avoid accidental or purposely generated access or data leaks relating to victims, suspects or perpetrators?
- Is the final evaluation of a CSEM report's content data always carried out by a human LEA officer?

2.7. Technical Robustness and Safety

2.7.1. Guidance and Instructions

Technical robustness requires AI systems to be developed with a preventative approach to risks and in a manner such that it reliably behaves as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm.[7] One of the measures that should have been developed and adopted in practice, in order to ensure various aspects related to tool robustness and safety, is a user accessible tool for signalling issues encountered with the GRACE tool or system during processing, which could thus trigger a more detailed manual review process during real-life use.

2.7.2. Questions

- Is there a tool or procedure in place through which users can signal issues encountered with the tool or system during processing when it comes to its robustness and safety?
- Is this tool used in practice?
- When relevant issues have been signalled, does a manual review process follow to improve tool robustness and safety?

2.8. Transparency

2.8.1. Traceability

2.8.1.1. Guidance and Instructions

For transparency within the law enforcement ecosystem, auditability of the GRACE system should be ensured by providing traceability mechanisms which document the methods used for its development. The auditability of the GRACE system requires documentation of testing methods especially for explicability, privacy, fairness, performance, safety and security. Transparency is closely linked to the principle of explicability which requires that all algorithmic decisions of an AI system can be understood by end-users in non-technical terms outlining what elements used in the (machine)

[7] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 16, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

learning model were responsible for each specific outcome. Focusing transparency on the question how an AI system arrives at a certain outcome requires predominantly technical properties of the system itself including the sourcing, the usage of training data as well as the processes of development and implementation. Especially for law enforcement, transparency is an essential component in figuring out who or what is accountable for potential problems with the use of AI-powered systems. The incorporation of sufficient traceability mechanisms will build end-user's trust in the GRACE system and ultimately society's trust in their use by LEAs. For proper traceability, there should be measures in place to continuously assess the quality of the input data to the AI system. This could take the form of a standard automated quality assessment of data input: quantifying missing values, gaps in the data; exploring breaks in the data supply; detecting when data is insufficient for a task; detecting when the input data is erroneous, incorrect, inaccurate or mismatched in format. The GRACE tools and platform should incorporate mechanisms for tracing back not only which data was used by the GRACE system to make a certain decision(s) or recommendation(s), but also which AI model or rules led to the decision(s) or recommendation(s) of the GRACE system. Furthermore, adequate logging practices should be put in place to record the decision(s) or recommendation(s) of the GRACE tool, platform and system. Lastly, measures to continuously assess the quality of the output(s) of the GRACE system which could take the form of a standard automated quality assessment of AI output (e.g. predictions scores are within expected ranges; anomaly detection in output and reassign input data leading to the anomaly detected) should have been so far developed, established and enabled.

2.8.1.2. Questions

- Is the quality of the input data to the Grace system continuously assessed?
- Is it possible to track back which data was used by the GRACE system to make a certain decision or recommendation as well as which AI model or rules led to that decision or recommendation?
- Are all of the decision(s) or recommendation(s) of the GRACE tool, platform and system recorded?
- Is a standard automated quality assessment of AI output in place?
- If yes, is it enabled?

2.8.2. *Explicability*

2.8.2.1. Guidance and Instructions

Ultimately, transparency concerning the reasons for AI-generated decisions amounts to explicability and primarily serves to maintain meaningful human oversight over the decisions made by an algorithm. Such meaningful human control is necessary to trace moral accountability for the outcomes of machine learning algorithms back to human beings. Transparency, is an essential component in figuring out who or what is accountable for potential problems with the use of AI-powered systems, especially in law enforcement. In this respect, transparency and explicability are a gradual matter catering to the level of understanding needed by the group it is provided for. This has led to the need to consider producing various degrees of explanations so that a sufficient understanding can be gained by all groups potentially in touch with the GRACE tools and platform and/or affected by its use:

The explicability of the GRACE system hinges on the ability to explain to the end-user both the technical processes of the GRACE tools and platform and the reasoning behind the decisions or predictions that the GRACE system suggests. The end-user can only build and maintain trust in the GRACE system, if the end-user understands the AI driven decisions of the GRACE tools and platform. This understanding needs to put the end-user in the position to contest as well as to identify ethically unacceptable considerations used by an AI-generated decision. Thus, specific mechanisms should have been, so far, incorporated in the GRACE system which will continuously survey the end-users whether and how they understand the decision(s) of the GRACE tools and platform. Lastly, explanations providing sufficient insight into the use of the GRACE system according to the level of confidentiality along the chain of authorisation should have been additionally prepared.

2.8.2.2. Questions

- Are users systematically surveyed on whether and how they understand the decision(s) of the GRACE tools and platform?
- Do users sufficiently respond to the surveys on whether and how they understand the decision(s) of the GRACE tools and platform?
- Do users consult explanations and material providing insight into the use of the GRACE system according to the level of confidentiality along the chain of authorisation?

2.8.3. *Open Communication*

2.8.3.1. Guidance and Instructions

The ethical dimension of transparency encompasses additionally open communication. This component, requires to communicate appropriately an AI system's capabilities and limitations to the end-users. In this respect, it is important for the end user, to be informed about the applied methodologies, technologies and protocols as well as the reason for choosing them, on the one hand, and the design decisions which create the GRACE tools and platform and for what purpose, on the other. However, it is vital that the accuracy of each tool and functionality is explained to the end users and furthermore how can they appropriately use it. Furthermore, an overview of all decisions automated by the GRACE tools and platform needs to be provided to the end users as well as an understanding of how they come about and on which criteria they are based on. This is quite important, since the GRACE system is intended to help and suggest a prioritisation of the enriched CSEM reports which immediately affect the use of law enforcement resources. Sound mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the GRACE tools and platform, should have therefore been already developed and incorporated. Such mechanisms need to have thoroughly communicated not only the benefits of the GRACE tools and platform to end-users, but also their technical limitations and potential risks to end-users, especially their level of accuracy and/or their error rates. Based on the above, it seems indispensable that the end users, are provided with appropriate training material and disclaimers on how to adequately use the GRACE system.

2.8.3.2. Questions

- Have authorized users received appropriate training about the purpose, criteria and limitations of the decision(s) generated by the GRACE tools and platform?

- Does this training communicate to them not only the benefits of the GRACE tools and platform but also their technical limitations and potential risks, especially their level of accuracy and/or their error rates?
- Has the functionality of each GRACE tool been explained to users?
- Have they been trained on how to use each of said tools?
- Have the applied methodologies, technologies and protocols as well as the reason for choosing them, and the design decisions which create the GRACE tools and platform been explained to users?

2.9. Unfair biases

2.9.1. Guidance and Instructions

AI bias: AI (or algorithmic) bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as favouring one arbitrary group of users over others. Bias can emerge due to many factors, including but not limited to the design of the algorithm or the unintended or unanticipated use or decisions relating to the way data is coded, collected, selected or used to train the algorithm. Bias can enter into algorithmic systems as a result of pre-existing cultural, social, or institutional expectations; because of technical limitations of their design; or by being used in unanticipated contexts or by audiences who are not considered in the software's initial design. AI bias is found across platforms, including but not limited to search engine results and social media platforms, and can have impacts ranging from inadvertent privacy violations to reinforcing social biases of race, gender, sexuality, and ethnicity.[8]

The development of a set of procedures in order to avoid creating or reinforcing unfair bias in the GRACE system was indispensable. More specifically, the GRACE system should contain processes to, inter alia, a) address and rectify for potential harm caused to children and b) to test and monitor for potential harm to children during the use phase of the GRACE system. It furthermore, needs to encompass a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance. Lastly, it should provide clear steps and ways of communicating (on how and to whom) issues related to bias, discrimination and poor performance should be addressed.

2.9.2. Questions

- Is a mechanism or procedure to test and monitor for potential harm to children victims or potential victims as a result of the use of the GRACE system systematically used?
- Are the system’s behaviour and results monitored closely for potential discrimination or unfairness following the input of real CSEM report content data?
- Is a mechanism that allows for the flagging of issues related to bias, discrimination or poor

[8] High-Level Expert Group on Artificial Intelligence (AI HLEG), The Assessment List for Trustworthy Artificial Intelligence (ALTAI), Available at: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

performance of the GRACE system in place and used?

- Have clear steps and ways of communicating such issues (eg. how and to whom to do so) been communicated to LEA users of the GRACE system?

2.10. Electronic Evidence

2.10.1. Guidance and Instructions

A digital footprint is the trail of data that one leaves when using the internet, otherwise known as electronic data. This electronic data could potentially be useful for law enforcement in crime investigation. Yet, at the same time, it needs to meet certain criteria in order to be accepted as judicial evidence. National laws on criminal procedure provide the requirements for data to be admissible as evidence in court. While the precise requirements of data admissibility might differ from one Member State to another, in general, they can be divided into two categories: legal requirements and technical requirements. Failing to comply with the former requirements, means that the electronic data will not be accepted by the court as evidence, while failing to comply with the latter, means that the electronic data might be questioned in court.

Starting off, national legislation on criminal procedure, usually regulates the issue of legitimate data collection, authorization procedure and requirements for evidence. In some cases, these laws establish different requirements for electronic evidence. Furthermore, the authorising body for electronic data collection can be also found in the national legislation. It might be a court, a public prosecutor or the head of a law enforcement institution. The collection of electronic evidence has to be documented in accordance to the national law requirements. Usually, it is established in forensic science methodologies or recommendations for law enforcement. Any activity relating to the search, seizure, access, storage or transfer of electronic evidence must be fully documented, preserved and available for review, in order to establish the authenticity of the data and initiate the chain of custody. According to the ISO/IEC 27037 standard, the documented chain of custody should consist of:

1) unique evidence identifier;

2) information on who accessed the evidence and the time and location it took place;

3) information on who checked the evidence in and out from the evidence preservation facility and when it happened;

4) information why the evidence was checked out (which case and the purpose) and the relevant authority, if applicable;

5) information if any unavoidable changes to the potential digital evidence, as well as the name of the individual responsible therefore and the justification for the introduction of the change.

Additionally, according to the authenticity requirement for electronic evidence, evidence must establish authentic facts in a way that authenticity thereof could not be disputed and is representative of its original state.[9] One should note that, documenting a chain of custody helps to prove authenticity and integrity of the data. Cyclic Redundancy Check (CRC 16, CRC 18) or one-way hash algorithm

[9] Council of Europe, Electronic Evidence Guide, p. 13.

(ex. MD2, MD4, MD5, SHA-1, SHA-2) with time could be used at each stage to prove integrity[10] of the electronic evidence and check for any errors in the evidence file. This way it would be possible, first of all, for the end users to identify changes if they occurred to digital evidence at any point of an investigation.[11] Secondly, it helps to (im)prove its value as evidence in court if a dispute occurs.

In general, for the prevention of any tampering with data contained in the GRACE system a technical solution should be implemented to prevent accidental or intentional modifications of documents. The system should use hash values (block-chain like) and auditing trails to prevent unrecognised interactions.

Applicable standards are the following:

1. ISO/IEC 27037 – guidelines for identification, collection, acquisition and preservation of digital evidence[12].
2. ISO/IEC 27041 – guidelines on assuring suitability and adequacy of incident investigation method[13].
3. ISO/IEC 27042 – guidelines for analysis and interpretation of digital evidence[14].
4. ISO/IEC 27043 – guidelines for incident investigation principles and processes[15].

2.10.2. Questions

- Have you considered whether under your national legal system, legal authorization (for example, by the court) has to be secured for the collection of personal data through the use of the GRACE systems and/or tools?
- Have you ensured that all law enforcement activities, namely, search, seizure, access, storage or transfer of electronic evidence are fully documented, preserved and available for review in order to establish the authenticity of the data and initiate the chain of custody?
- If yes, does such documentation includes the following:
  1 unique evidence identifier;

  2 information on who accessed the evidence and the time and location it took place;

  3 information on who checked the evidence in and out from the evidence preservation facility and when it happened;

  4 information why the evidence was checked out (which case and the purpose) and the relevant authority, if applicable;

  5 information if any unavoidable changes to the potential digital evidence, as well as the name of the individual responsible therefore and the justification for the introduction of the

[10] Hosmer, Chet, Proving the Integrity of Digital Evidence with Time, International Journal of Digital Evidence, 2002.

[11] Schmitt, Veronica & Jordaan, Jason. Establishing the Validity of Md5 and Sha-1 Hashing in Digital Forensic Practice in Light of Recent Research Demonstrating Cryptographic Weaknesses in These Algorithms. International Journal of Computer Applications. 2013. 68. 40-43. 10.5120/11723-7433.

[12] https://www.iso.org/standard/44381.html

[13] https://www.iso.org/standard/44405.html

[14] https://www.iso.org/obp/ui/#iso:std:iso-iec:27042:ed-1:v1:en

[15] https://www.iso.org/standard/44407.html

change

- Do you apply mechanisms to ensure the authenticity of the evidence securing that evidence is representative of its original state and has not undergone or suffered any changes during the investigation and/or before it presented to the court?
- Is the accidental or intentional modifications of the documents and data in GRACE technically prevented?
- Are relevant technical solutions preventing unrecognised interactions, such as hash values (block-chain like) and auditing trails in place and enabled when the GRACE system is used?
- Do you adhere to any of the following standards?
  1. ISO/IEC 27037 – guidelines for identification, collection, acquisition and preservation of digital evidence.
  2. ISO/IEC 27041 – guidelines on assuring suitability and adequacy of incident investigation method.
  3. ISO/IEC 27042 – guidelines for analysis and interpretation of digital evidence.
  4. ISO/IEC 27043 – guidelines for incident investigation principles and processes.

2.11. Use of Crawlers

2.11.1. Guidance and Instructions

“Crawling is the process of exploring web applications automatically. The web crawler aims at discovering the web pages of a web application by navigating through the application. This is usually done by simulating the possible user interactions considering just the client-side of the application. As the amount of information on the web has been increasing drastically, web users increasingly rely on search engines to find desired data. In order for search engines to learn about the new data as it becomes available on the web, the web crawler has to constantly crawl and update the search engine database”.[16]

Although not originally planned, due to the investigative need to verify and update the data contained in a CSEM report at some stage, on the one hand, and because of potential synergy effects with the results of the EU-funded AviaTor project which developed a Targeted Online Research as optional functionality for the AviaTor solution, on the other hand, the integration of a search tool in the GRACE system was suggested to the GRACE Consortium in May 2021.[17]

Unlike traditional web crawlers that create an index of available content, the tool utilized within the GRACE solution focuses on enriching existing data sets with additional information. Furthermore, this crawler is, in practice activated only once a LEA officer has initiated it.

In using the crawler, along/in combination with the GRACE system/and or tools, a LEA officer, being

[16] Seyed M. Mirtaheri, Mustafa Emre Dincturk, Salman Hooshmand, Gregor V. Bochmann, Guy-Vincent Jourdan, *A brief history of web crawlers, CASCON '13: Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research*

[17] See “T3.1 Memo - targeted crawling of open source information”, 19 May 2021.

the end user, needs to have in mind the information below:

Initially, while a CSEM report presents sufficient evidence for a LEA officer to start a proper investigation, the crawler's element of automation reduces the amount of human agency. In contrast, a LEA officer, has better common-sense reasoning, and as a result, a better chance to recognise the bigger picture and unusual context. Therefore, by evaluating the search results of such a crawler, before a LEA officer proceeds to take any decision or action based on those results, he/she might decide to select only specific parts of the evidence for investigation, which might also have to take place in a strategic sequence.

Furthermore, the crawler would filter the gathered information automatically on the basis of selectors.[18] Any potential search term consists of specific selectors. Such selectors are difficult to establish in an ethically acceptable manner, because they have to be reasonable, evidence-based and non-discriminatory.

As an example, one possibility, for the crawler, is to choose particular keywords as selectors. An ideal selector keyword in this regard would be a word that is known to be used exclusively in the context of CSE activities, perhaps something like a secret codeword. The next best would be words providing reasons for suspicion based strongly on evidence. Other possible keywords appear to be rather terms used by large numbers of people for almost any reason. A keyword may not be discriminatory because it has to be indicative only of suspicious CSE activity. A single keyword appears difficult, if not impossible, to define in this regard, but also a set of several keywords may not contain a discriminatory keyword because the use of certain words is not only ambivalent but also more likely to be used by certain cultural groups and in that respect discriminatory. It follows from all this, as it has been aforementioned, that a keyword used as a selector for an automated search tool would have to be reasonable, evidence-based and non-discriminatory. Based on the above, there should be mechanisms and procedures in place guiding the chosen selectors.

In addition, it has to be born in mind that suitable selectors have to be flexible enough to respond to suspects who are forensically aware and aim to avoid the use of incriminating language, for example, by avoiding direct reference to suspicious substances, items, groups, or people.

Additionally, when using the crawler, the end user should not rely blindly on the crawler's search results but continue to follow procedures that ensure that the observation will be stopped as soon as it becomes clear that insufficient evidence exists for continued suspicion so that the measure could be defended as legally proportionate and ethically legitimate.

Moreover, the quality of evidence produced by the crawler would probably suffer from a risk of inconclusiveness, a risk of the algorithm's inscrutability and potential bias. As a minimum safeguard, the search results of such a crawler would have to be evaluated by the end user, before he/she takes any decision or action based on evidence produced by the crawler. Therefore, the LEA officer, evaluating the evidence presented by the crawler, would have to be trained sufficiently regarding the following aspects: distinguishing between correlation and causality; awareness of false or too simplified models; awareness of the ethical risks and dangers involved with the inscrutability of algorithms; awareness that meaning is not self-evident in statistical models and that the explanation

[18] What are selectors? Selectors are a "mechanism for extracting data. They're called selectors because they "select" certain parts of the HTML document specified either by XPath or CSS expressions. XPath is a language for selecting nodes in XML documents, which can also be used with HTML. CSS is a language for applying styles to HTML documents. It defines selectors to associate those styles with specific HTML elements.https:// docs.scrapy.org/en/latest/topics/selectors.html

of any correlation requires additional justification; and awareness of the risk of unfair discrimination by or based on the profiling by the crawler.

Lastly, although the potential harm of a CSE act is, in general, very high, at the time he/she uses the automated search tool neither the probability of this harm is established to be high nor a very likely source for it might be established. As a consequence, the effectiveness of using the automated tool appears in doubt while the probability of intruding deeply and unjustifiably into the lives of individuals who are not involved in terrorism seems rather high. Having this in mind, it is preferable that LEAs, using such an automated search tool, will have to monitor the tool's effectiveness so that they have an evidence base, from which to draw for future decisions about its use in operations.

2.11.2. Questions:

- If a crawler is used in combination with the GRACE system and/or tools, are the search results of such a crawler always evaluated by a human LEA officer before they may trigger any decision or action based on them?
- Are there mechanisms and procedures in place guiding the choice of selectors for an automated search?
- Can the mechanisms and procedures referred to in the previous question ensure that the chosen selectors are reasonable, evidence-based and non-discriminatory and that they do not contain incriminating language?
- Are there mechanisms in place preventing users from relying blindly on the crawler's search results?
- If yes, have users received training regarding the following aspects:
    - distinguishing between correlation and causality;
    - awareness of false or too simplified models;
    - awareness of the ethical risks and dangers involved with the inscrutability of algorithms;
    - awareness that meaning is not self-evident in statistical models and that the explanation of any correlation requires additional justification;
    - awareness of the risk of unfair discrimination by or based on the profiling by the crawler.
- Do users have access to mechanisms enabling them to monitor the crawler's effectiveness so that they can have an evidence base from which to draw for future decisions about its use in operations?

3. Conclusion

3.1. Summary

This Deliverable D9.9 has provided the review procedure, that LEA officers, working in the area of child sexual exploitation, will need to follow in order to confirm that the operation of the GRACE system, at national level, will not be circumventing the implemented safeguards and that it will be in full compliance with the relevant legislative and ethical framework analysed in previous tasks.

In section 2 eleven thematic subsections have been developed, receiving input from previous deliverables, each consisting of two main parts, a) a specifically-designed check list, which has to be gone through by the reviewer and b) a set of guidelines and instructions, which actually precedes each check list, guiding the relevant party responsible for the review as to why and how these issues are important and needed to be assessed. These eleven sections concern fundamental legal and other issues that need to be examined by LEA officers, and more specifically cover the topics of access control, audit trail, data protection, database search, human oversight, measures against re-victimization and over-exposure, technical robustness and safety, transparency (traceability, explicability, open communication), unfair bias, electronic evidence and the use of crawlers.

3.2. Evaluation

This Deliverable D9.9 constitutes the last deliverable of Work Package 9 (and the sole deliverable of T9.5) and can be argued to be a deliverable of fundamental importance as it brings together in the form of a review mechanism – in detail, a checklist, accompanied by guidelines and instructions - a great part of the significant and vital work carried out under other deliverables and tasks of the same work package (D9.1, D9.3, D9.5, D9.7). More importantly it includes in its content, questions, divided in eleven thematic sections, that will be vital for the correct, ethical and legitimate use of the GRACE system and tools, in practice, by LEA officers. Therefore, the realization of this review process can be seen as a fundamental step forward in the successful use of the GRACE system and tools in its roll-out phase.