This project that 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

Architecture for technical safeguards – "security and privacy by design" v2

Deliverable number	D9.8	
Version:	2.0	
Delivery date:	July 2023	
Dissemination level:	PU	
Classification level:	Non classified	
Status	Final	
Nature:	Report	
Main author(s):	Ulrich Gasper, Marco Gercke	CRI
Contributor(s):	Ulrich Gasper Prof. Dr. Marco Gercke	CRI CRI

DOCUMENT CONTROL

Version	Date	Author(s)	Change(s)
1.0	30/06/2021	Ulrich Gasper, Marco Gercke	Submission of v1 of this document: D9.7 – Architecture for technical safeguards – "security and privacy by design" v1
1.1	08/07/2023	Marco Gercke	Considerations for not updating D9.7 in chapter 1.
2.0	28/07/2023	Peter Leskovsky	Deliverable submission

DISCLAIMER

Table of Contents

1. Introduction

Considerations related to the decision not to update D9.7:

Based upon the Grant Agreement, the Ethical Report (D9.1), the Legal Report (D9.3), the Overall Legal and Ethical Framework (D9.5) and the Architecture for Technical Safeguards (D9.7) are to be updated towards the end of the GRACE project. With regard to the Ethical Report (D9.1) and the Legal Report (D9.3) updates have been useful as especially in the legal field significant developments have taken place (i.e. proposal and debate of draft legislation such as the Draft AI Act). Therefore, WP9 has invested significant time and effort in continuously updating D9.1 and D9.3 the results of which are documented in D9.2 and D9.4. The aim was to ensure that at each stage of the GRACE project and especially towards its end, the impact of any latest development was and is monitored, properly examined and included in these reports in order to provide accurate information for future readers of D9.2 and D9.4.

With regard to the Architecture for Technical Safeguards (D9.7) the situation is different. The Architecture contains important technical solutions for the developers to consider when developing the GRACE solution. The partners of the GRACE Consortium participating in WP9 continuously monitored potentially relevant developments related to the range of content covered in D9.7. If such relevant developments had been identified, a process to update D9.7 would have been implemented immediately - prior to the due date of D9.8. This monitoring and update process was established to ensure that the input to the Architecture was not outdated during the lifespan of the GRACE project. However, no development requiring the need to initiate an update process has been identified. Therefore, the decision was taken to leave the content of D9.7 unchanged and only add this explanation in D9.8.

1.1. Overview

The DoA describes this deliverable as:

> *Architecture for the development of technical safeguards to prevent violations of ethical and legal standards – "security and privacy by design" has to be implemented. Related task(s): T9.4. [M 13]*

The main objective of this Deliverable D9.7 is to provide an overview of reliable ways for implementing technical safeguards by design that significantly mitigate the risk of the GRACE tools and platform being (mis)used in an illegal and/or ethically unacceptable manner. In order to further strengthen the acceptance of the GRACE solution, the Consortium Members involved in Task T9.4 together and in close cooperation with the technical partners responsible for the development of the GRACE architecture looked into ways to implement "technical safeguards by design" that address the issues and thereby significantly limit the possibility that the tools can be used in legally or ethically challenging ways. With regard to all five scenarios listed below, technical solutions were not only developed as theoretical concept but will be implemented into the GRACE solution.

Technical Environments

Two environments have been defined to carry out the GRACE project activities and host the various platforms constituting the GRACE solution:

- *Development/Integration Environment:*

  In this environment, not only the merged work can be built, together, as a combined system, but also basic tests of one system's integration points with other upstream or downstream systems can be performed.

  The infrastructures of environment are and will be deployed mainly in the facilities of the technical partner Vicomtech and other technical partners such as CERTH.

- *Staging/Pre-Production Environment:*

  In this environment, the Consortium partners EUROPOL and participating LEAs represent business stakeholders and as such can test the GRACE system against their original business requirements. This environment is a nearly exact replica of a production environment for software testing.

  The infrastructures that make up this environment will be deployed mainly in the facilities of EUROPOL and the LEAs.

It is very important to note, that, as a research and innovation project, the GRACE project does not and will not include the Production Environment as part of its scope. However, the Production Environment is already also considered in order to have the overall picture.

Scenarios

This Deliverable D9.7 builds upon the results of Tasks T9.1 and T9.2 that identified various ethical and legal issues of relevance during the development of the GRACE solution. This Deliverable D9.7 explains the methodology that was applied and has allowed to identify particularly five critical scenarios which require technical safeguards by design from a legal and ethical perspective:

(i) unauthorised access (section 3. below);

(ii) chain of custody (section 4. below);

(iii) audit trail (section 5. below);

(iv) bias detection (section 6. below); and

(v) restrictions for targeted OSINT crawler (section 7. below).

Aim

At the time for submission of this Deliverable D9.7, the technical development is still at a very early stage. The purpose of this Deliverable is, therefore, not to provide a conclusive technical safeguard architecture. Instead, the focus is on providing the underlying framework for the entire process and introduce the first five technical safeguards that have already been identified at this stage of the development process. Task T9.4 is an ongoing task and, as a consequence, the work on developing the security architecture will continue until month M33. It is possible that during the continuing technical development needs for additional technical safeguards will be identified and added to the Deliverable D9.8. However, the basic concept of technical measures preventing

abuse and ensuring legitimate and ethically acceptable use of the GRACE tools and platform, that is laid out in this Deliverable D9.7, will remain unchanged.

The architecture for technical safeguards ensuring compliance by design for the GRACE system during the development in the course of the project and beyond after a potential roll-out is addressed from all crucial perspectives. While this Deliverable D9.7 approaches the architecture for technical safeguards from an ethical and a legal perspective, the user perspective as well as the technical perspective are advanced in Deliverable D2.10 elaborating the *"Technical and Architecture Specifications"* of the GRACE tools and platform and in Deliverable D2.14 documenting the evolving *"Security and Auditing Mechanisms"* implemented in the entire GRACE system. The GRACE project will document the development of each, the *"Technical and Architecture Specifications"* and the *"Security and Auditing Mechanisms"* in four iterations creating the transparency necessary for an artificial intelligence (AI) system to be operated within the law enforcement ecosystem in an area as sensitive as CSEM. While Deliverable D2.10 and Deliverable D2.14 are each the first of four iterations, Deliverable D9.7 is the first of only two iterations on the *"Architecture for Technical Safeguards"*. Deliverable D9.7 provides an initial guidance for preventing violations of ethical and legal standards, while the final assessment will be reported in the second iteration of the *"Architecture for Technical Safeguards"* in Deliverable D9.8 which will be provided after the third iteration of both, the *"Technical and Architecture Specifications"* in Deliverable D2.12 and of the *"Security and Auditing Mechanisms"* in Deliverable D2.16, so that their fourth and final iteration in M40 could still be adjusted accordingly. Ultimately, all three perspectives (technical, ethical and legal) supplement each other focussing on vital aspects of a complex AI system's development.

1.2. Relation to Other Deliverables

This Deliverable D9.7 is related to the following other GRACE deliverables:

Receives inputs from:

Deliv. #	Deliverable title	How the two deliverables are related
D2.10	Technical and Architecture Specification v1	While D2.10 approaches technical safeguards from a technical perspective, D9.7 addresses technical safeguards from an ethical and legal perspective.
D2.14	Security and Auditing Mechanisms v1	While D2.14 approaches guidelines for technical safeguards from a technical perspective, D9.7 addresses technical safeguards from an ethical and legal perspective.

Table 1 – Relation to other deliverables – receives inputs from

Provides outputs to:

Deliv. #	Deliverable title	How the two deliverables are related
D2.10	Technical and Architecture Specification v2	While D9.7 expands the technical safeguards to be implemented from an ethical and legal perspective, D2.10 v2 will approach the technical safeguards for implementation from a technical perspective.
D2.14	Security and Auditing Mechanisms v2	While D9.7 expands the technical safeguards to be implemented from an ethical and legal perspective, D2.14 v2 will approach the technical safeguards for implementation from a technical perspective.

Table 2 – Relation to other deliverables – provides outputs to

1.3. Structure of the Deliverable

This document includes the following sections:

- Section 2 outlines how the ethical and legal perspectives interrelate with the technical perspective in order to properly address the ethical and legal challenges for the GRACE system through security-by-design technical safeguards. Highlighting the necessary interplay between ethics, law and technology (section 2.1.) the methodology applied for this Deliverable D9.7 (section 2.2.) and the results (section 2.3.) are presented which led to the critical scenarios identified (section 2.4.).
- Section 3 describes for the critical scenario of unauthorised access to the GRACE solution the legal and ethical values at stake and provides guidance for a technical solution.

- Section 4 describes for the critical scenario of chain of custody the legal and ethical values at stake and provides guidance for a technical solution.
- Section 5 describes for the critical scenario of audit trail the ethical values at stake and provides guidance for a technical solution.
- Section 6 describes for the critical scenario of bias detection the ethical values at stake and provides guidance for a technical solution.
- Section 7 describes for the critical scenario of finding suitable restrictions for a targeted OSINT crawler the ethical values at stake and provides guidance for a technical solution.
- Section 8 presents as conclusion first a summary (section 8.1.), then an evaluation (section 8.2.) and finally the necessary future work (section 8.3.).

2. Addressing Ethical and Legal Challenges Through Technology

From the time of proposal up until now, it has been apparent that while the GRACE tools and platform could be powerful instruments providing unique opportunities for LEAs and through this process to society in general in the fight against CSEM, there are, at the same time, inherent risks associated with the features and advanced methods utilised. In this regard, it is important to distinguish between two reasons for concerns:

(a) *Use:* The first category of issues is related to legitimate ethical and legal concerns. As pointed out in the Ethics Report[1] drafted during the initial phase of the project there are several ethical issues related to the tools and FL platform for the processing of CSEM reports envisioned as the GRACE system. Anybody familiar with AI/ML will identify the ethical challenges. Being in a position to respond to such criticism and undertaking measures to demonstrate that they have not only been identified and discussed but also addressed and mitigated by technology has been a driving factor for the technical solutions described in this Deliverable D9.7.

(b) *Misuse:* The second category is related to the potential of intentional or unintentional misuse of the tool's functions and capacities. People with only rare and limited access to the required resources might get a wrong impression of the tools' functions, abilities and restrictions. Having technical measures in place not only indicating transparently the intended use for legitimate purposes, but also mitigating potential conflicts seems necessary to ensure proper use of the GRACE tools and platform.

The development of the guidance for technical solutions contained in this Deliverable D9.7 strongly builds on the prior work carried out under WP9. The core aim of WP9 is to identify and to analyse not only any real and potential ethical and legal implications for online investigations which involve gathering, analysing and exchanging information, but also legal considerations that need to be taken into account when developing such tools. The research carried out in this context revealed various areas in which – due to the technical capacities - an unintended operation of the tools could interfere with legal and especially ethical standards. These standards range from the requirement of a legitimate legal authorisation for using such tools to the need for protecting victims and their rights. The findings are described in the Ethics Report[2] and in the Legal Report[3].

When designing WP9 the drafters did not consider the Ethics Report and the Legal Report to be the final output. The aim of WP9 is not only to deliver research going beyond state-of-the-art with regard to automated investigations. The purpose of the research is also to ensure that potential legal, ethical and societal issues can be addressed during the process of developing the tools. In this regard, the driving idea for Deliverable D9.7 is to raise awareness of these issues among developers, on the one hand, and to provide the scientific basis for a discussion about suitable technical solutions addressing these issues within the design process. As a result, the GRACE tools and platform will be deeply rooted in the developers' shared believe that "security/safeguards by design"[4] can strengthen the acceptance of the tools and platform.

[1] Ethics Report, Deliverable D9.1.
[2] Ethics Report, Deliverable D9.1.
[3] Legal Report, Deliverable D9.3.
[4] Regarding the principle "security by design" in software development see: *Othmane/Jaatun/Weippl*, Empirical Research for Software Security, 2018.

Against this background, the findings in the WP9's Ethics Report and Legal Report as well as the results of a red teaming exercise[5] have been combined to establish a solid foundation for developing realistic case scenarios. Because the Legal Report in Deliverable D9.3 was finalised only a month before this Deliverable D9.7 has to be submitted, the scenarios presented here seem to draw more from the ethical perspective. A more fundamental reason for the predominant focus on the ethical perspective is that the awareness of ethical values usually emerges well before any safeguards for their guarantee are established as legal obligations and/or as technical standards. A case in point is the evolution of ethical guidelines for trustworthy AI having started with ethical frameworks and guidelines issued by public, private or academic organisations,[6] and the "Ethics Guidelines for Trustworthy AI" elaborated by the AI H-LEG,[7] leading to the European Commission's Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)[8] from the legal perspective and the standards related to AI systems suggested by the IEEE Standards Organisation,[9] from the technical perspective. So far, five scenarios have been identified:

1) unauthorised access (section 3. below);
2) chain of custody (section 4. below);
3) audit trail (section 5. below);
4) bias detection (section 6. below); and
5) restrictions for targeted OSINT crawler (section 7. below).

The Consortium Partners involved in WP9 are tasked to develop technical solutions that address these issues by providing state-of-the-art preventive measures to hinder or mitigate legal and ethical violations and, as a minimum, create unalterable records that allow for retrospective auditing. The sections 3. – 7. below provide background information describing the scenarios as well as the proposed solutions.

2.1. General Observation

This Deliverable D9.7 describes the technical solutions to prevent and mitigate misuse of the GRACE tools and platform. Although technology has a great potential to prevent misuse of a tool and is rightly considered a major factor in any prevention regime, it is important to highlight that technical solutions are limited. By far not every misuse can be identified and prevented by technology. Therefore, the third issue ("audit trail" emanating from the fear of an inability to reconstruct abusive behaviour) had to be added to the list.

[5] See section 2.2(a) and 2.3. below.
[6] Ranging from mere recommendations over voluntary commitments up to binding policies, see section 2.4 of Deliverable D9.1.
[7] Also not (yet) binding, see section 2.5 of Deliverable D9.1; AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[8] Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM(2021) 206 final, 21 April 2021.
[9] IEEE Standards Association, "IEEE portfolio of AI systems technology and impact standards and standards projects", available at: https://standards.ieee.org/initiatives/artificial-intelligence-systems/standards.html.

Having technical solutions in place enabling retrospective audits will help mitigating those scenarios in which technical prevention measures are insufficient to prevent a misuse of the technology and can only mitigate. It is equally important to emphasise that technical prevention/mitigation measures are neither discussed nor implemented out of a lack of trust in the work of LEAs. Rather, the concept of technical prevention/mitigation measures is aimed to increase the trust of the broader society in LEA's and their use of such tools.

2.2. Methodology

When designing the work plan establishing the architecture for technical safeguards, the working group in charge decided to undertake a three-step approach:

(a) The initial step was a red teaming exercise. Red teaming describes the methodology to challenge concepts and set mines related to certain issues.[10] Military forces use this methodology for decades.[11] This approach was undertaken to ensure that potential issues can be identified and are not neglected just because team members are solely focused on the constructive development of an underlying solutions.

(b) The second and crucial step was a careful counter-evaluation of existing technological deliverables focusing on legal and ethical aspects (as outlined in the Legal Report[12] and the Ethical Report[13]) as well as on the documentation of technology, especially in Deliverables D2.10 and D2.14, in an effort to identify possible issues that had not been identified in the course of the red teaming exercise.

(c) The final step will be the development of technical solutions based on the issues identified in both previous steps.

2.3. Results of the Red Teaming Exercise

The red teaming exercise was designed and carried out by a team frequently tasked with carrying out such exercises in the field of ethics and legislation.[14] While usually such exercise is carried out by involving other contributing Consortium Members, COVID-19 restrictions did not allow such broader integration and the exercise was carried out by CRI members only. For the purpose of carrying out step one (see 2.2.(a) above), the team created two teams – a red team and a blue team. The task of the red team was to identify any

[10] With regard to the concept of red teaming see: *Herman/Frost/ Kurz*, Wargaming for Leaders. 2009; *Sabin*, Simulating War, 2012; *Lauder*, Red Dawn: The Emergence of a red teaming capability in the Canadian Forces, Canadian Army Journal, Vol. 12.2, 2009; *Longbine*, Red Teaming: Past and Present, 2008; *Wood/Duggan*, Red Teaming of Advanced Information Assurance Concepts, DARPA Information Survivability Conference and Exposition, 2002. DISCEX 00 Proceedings, Vol. 2, page 112 et.seq.

[11] See in this regard: *Longbine*, Red Teaming: Past and Present, 2008. *Lauder*, Red Dawn: The Emergence of a red teaming capability in the Canadian Forces, Canadian Army Journal, Vol. 12.2, 2009, page. 28.

[12] Legal Report, Deliverable D9.3.

[13] Ethics Report, Deliverable D9.1.

[14] Regarding the use of red teaming in the field of legislation see: *Gercke*, Red Teaming Approaches for more effective legislative drafting?, Computer und Recht 2014, page 344 et seq.

criticism, weaknesses and areas how the tool could be misused in a way that could lead to a violation of legal and ethical standards and consequently prompt legitimate criticism. The task of the blue team was to anticipate criticism and develop concepts for technical solutions which address the anticipated issues of the GRACE tools and platform. The exercise was carried out over two days and led to an initial list of three areas in which challenges had the potential to be addressed through technical measures. It is a sign of the effectiveness and usefulness of such an approach, that the final list of topics covered in this Deliverable D9.7 is largely identical with those topics identified in the initial process. One issue (restrictions for targeted OSINT crawler) was added at a later stage to address potential abuses that seem preventable by technical measures.

2.4. Identification of Critical Scenarios

A key focus of the GRACE project is to ensure and maintain ethical and legal compliance of the entire project. For that purpose, all participants of the project have to cooperate and exchange useful findings so that these can be born in mind for the course of the project and allow suitable pre-emptive measures tackling all issues raised from a legal and an ethics perspective.

In general terms, it is vital that appropriate procedural safeguards are designed and built into the GRACE solution. This is a fundamental element of the lawfulness requirement of Art. 8 ECHR[15] and is also a matter of common sense. Procedural safeguards include

(a) audit capability,

(b) recording of investigations, and

(c) the facility to set time-limits for the retention and destruction of information.

The European Court of Human Rights (ECtHR) has stated that "it would be entirely contrary to the need to protect private life under Article 8, if the Government could create a database in such a manner that the data in it could not be easily reviewed or edited, and then use this development as a justification to refuse to remove information from that database."[16] Guided by this approach, its seems helpful to distinguish between "*ordinary issues*" and "*capital issues*":

(i) Whereas "*ordinary issues*" merely create an awareness of the ethical value underlying important decisions made for and during the development of the GRACE tools and platform,

(ii) "*capital issues*" pose serious threats to the legality and societal acceptability of the GRACE project as well as later use of the tools and platform developed by the GRACE project.

Because of their inherent nature and far-reaching dangerous capacity, "*capital issues*" require technological countermeasures to be built into the GRACE tools and platform preventing their unethical and illegal use.

[15] European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR; as amended by Protocols Nos. 11 and 14 and supplemented by Protocols Nos. 1, 4, 6, 7, 12 and 13), 4 December 1950.

[16] ECtHR, *Catt v United Kingdom*, judgment of 24 January 2019 at paragraph 127.

Accordingly, when a "*capital issue*" has been identified, not only every partner of the GRACE Consortium involved with a specific deliverable entailing the "*capital issue*" but also the entire project have to react and support suitable technological countermeasures in an effort to keep the capital legal, ethical and/or societal risks at bay. If at all possible, technical safeguards and solutions are expected to be developed and integrated in the functionality of the respective GRACE tool as well as the entire GRACE platform in order to comply with the highest ethical standard required for a project like GRACE.

Against this background, five *"capital issues"* have been identified so far to which the participants of the GRACE project have to react to and adjust the respective functionalities. These *"capital issues"* are critical not only for developing helpful functionalities of the GRACE tools and platform, but also for avoiding unwanted complications or hindrances of the intended (commercial) exploitation of the GRACE tools and platform.

3. Unauthorised Access

The use of the GRACE tools and platform involves a large-scale analysis of CSEM data. Therefore, access to the functionalities of the GRACE tools and platform has to be guarded at every possible level. Only the highest level of security can ensure that the GRACE tools and platform properly assist LEAs by automating necessary and proportionate data analysis.

3.1. Legal Issues at Stake

CSEM is one of the most regulated areas of illegal content with broad international consensus that the harm of such material is so substantial, that it requires extensive criminalisation. The applicable international and regional frameworks do not include any specific exemption from criminal liability for individual LEA officers or researchers.[17] In order to prevent criminal liability for interacting with CSEM, the GRACE system has to ensure that only LEA officers and researchers gain access to it for whom national law provides an exemption from criminal liability for their respective activity. In this case it is important to underline that not even all LEA officers are authorised to access such material. As a consequence, the status "LEA officer" alone is not a suitable access condition.

3.2. Ethical Values at Stake

In addition to the legal issues, the ethical dimension of preventing harm necessitates adequate data governance which includes monitoring the processing and access protocols.[18] Considering the envisioned pivotal role in coordinating law enforcement's response to CSEM reports, the GRACE system has to be protected against vulnerabilities allowing them to be exploited by adversaries. Security measures need to prevent unauthorised entities as much as possible from gaining access to the GRACE tools or platform and/or tampering with the data of the CSEM reports.

3.3. Technical Solution

The mitigation of risks concerning unauthorised access and use of data has already been identified and addressed under the heading of information security in Deliverable D2.14.[19] In this regard, the technical perspective is aligned with the ethical and legal perspective, so that it is appropriate to reiterate here three of the guiding general security principles highlighted there which are key to a technical solution preventing unauthorised access to the GRACE tools and platform:[20]

- *"Least privilege: Only the minimum possible privileges should be granted to a user, or a process for accessing a resource.*

[17] See section 2. Of Deliverable D9.3.
[18] See section 3.3 of Deliverable 9.1.
[19] See section 3.2 of Deliverable 2.14.
[20] See section 3.2.1. of Deliverable D2.14.

- *Need to know: Access to information shall be restricted to those who have the need to know regardless of their security clearance.*
- *Defence in depth: Using layers of security increases the level of effort required by an attacker to gain unauthorised access to a system or application. In the event one security control fails or is compromised, another security control should prevent the exposure of information or an information system."*

Further, Deliverable D.2.14 has also already established the following fundamental requirements concerning access control:[21]

"Access control aims to ensure that data and functionality are accessed only by authorized users for the legitimate purposes of performing their business tasks, by providing a set of guidelines on how to manage users, access, etc.

- *Authentication: access to GRACE components must be done only after authentication.*
- *Authorisation: application level authorisation must be enforced to ensure GRACE functionality and data access are restricted in line with the authorisation rules which will be agreed.*
- *Principle of least privilege: users/programs/processes should make use of the minimum privileges possible in order to perform their tasks.*
- *User registration process: there will be a clear process where new user accounts will be created.*
- *Deactivation and deletion of user accounts: unnecessary user accounts are removed.*
- *User accounts: all users will make use of personal accounts in order to access the GRACE platform.*
- *User roles: users are only allowed to use the GRACE platform within the scope of the permissions and rights allocated to the role, which has been assigned to the user by the platform/sub-site manager.*
- *User privilege management: constrain the allocation of privileges to users and applications to that required to perform the business function. Review the allocated privileges periodically and revoke privileges when no longer needed.*
- *Password policy: enforce password complexity requirements established by policy or regulation. Authentication credentials should be sufficient to withstand attacks that are typical of the threats in the deployed environment (e.g., requiring the use of alphabetic as well as numeric and/or special characters).*
- *Session management: ensure proper configuration of session parameters across all component (time-out, session id generation, auditing, etc).*
- *Logging: all user actions must be logged".*

Implementation

At the current stage of the GRACE project, the Development Environment has already been implemented and made available to technical Consortium partners to carry out the tasks described in section 1.1. above. The Development Environment is located in the facilities of the technical Consortium partner Vicomtech. All the

[21] See section 3.2.2. of Deliverable D2.14.

devices are inside a *virtual private network* (VPN) completely independent of any other Vicomtech business network and isolated from any external connection. In order to ensure only GRACE project researchers can access the resources within this VPN, a firewall has been deployed that can only be surpassed by users previously accredited by Vicomtech's IT team using each individual's private and personal credentials. If any of the services deployed within the VPN requires exposing an external connection to the internet for non-accredited users to access a specific functionality (e.g. user interface of the federated learning platform), a perimeter subnet (DMZ) has been enabled, isolating the exposed service from the rest of the private network and thus adding an extra layer of security.

Strategy

Fully aware of the legal and ethical issues at stake,[22] the main objective of the separating the *Development Environment* from the *Staging Environment*[23] is the control of access to the data of CSEM reports.

To ensure comprehensive compliance by design with these security and privacy protocols, in no case access to CSEM data will be granted in the *Development Environment*. Therefore, the security and privacy by design strategy is based on the fact that the various platforms of the GRACE system will be fully integrated and tested in the *Development Environment* until a high level of confidence is achieved that the different components are functional based on the GRACE project requirements. Only once this level of compliance has been achieved and verified, the various platforms will be deployed in the *Staging Environment* under the supervision of the Consortium partners EUROPOL and participating LEAs in order to access their CSEM data.

In the *Staging Environment*, access to CSEM data will be restricted, controlled and partial to create and train machine learning models that are as accurate and reliable as possible, in order to achieve all objectives set out for the final stage of research and development in the course of the GRACE project. During the training process of new machine learning models, the requirements regarding access control (and data privacy) will remain by design. For this purpose, a Federated Learning platform is being developed, capable of training a machine learning model by bringing the model to the data for training, and therefore without requiring direct access to the CSEM data for GRACE researchers.

The Federated Learning Approach

Federated Learning (FL) provides a structure for decomposing the overall *Machine Learning* (ML) work-flow into the approachable modular units. The GRACE project aims to provide an EU-wide FL platform that will exploit the available infrastructure to train *Neural Network* (NN) models.

In the *Development Environment*, the key design principles and beneficial characteristics for the GRACE FL platform are:

(i) Design of a suitable FL programming environment that enables the development and application of ML tools;

(ii) Efficient handling of FL network security, privacy and legal aspects, where initially requirements related to data privacy, security, exchange, integrity and controlled/authorised access to the data by each stakeholder in a federated topology are investigated and subsequently a corresponding access scheme satisfying them will be implemented; and

[22] See sections 3.1. and 3.2. above.
[23] See section 2.1. above.

(iii) Design of knowledge aggregation strategies for federated systems, where efficient and effective methodologies are developed that guarantee convergence and robustness of the FL system as well as communication efficiency between the nodes and the aggregator, ensuring the minimal data transfer between the nodes and the central server.

Concerning the *Staging Environment*, one of the primary attractions of the envisaged GRACE FL platform is that it provides an effective implementation of the fundamental principle of *data minimisation*.[24] The benefits of this approach include not only an enhanced privacy for all participating users of the GRACE solution, but also a by-definition privacy-preserving approach:

> *Data Minimisation:* The CSEM raw data processed by a user never leaves the local device on a LEA's premises (*MS node*), and only updates to models (e.g., gradient updates) which are sent to the *master node*, enabling the processing of data in the premises of the data owners.

In particular, the GRACE FL platform will follow a distributed processing paradigm according to which CSEM raw data (e.g. images and audio samples) is processed locally by each participating LEA as data owner (*MS node*) and contributes to the development of a *global network node model* by computing updates, using the CSEM data which is available only locally. It is critical to highlight that according to the adopted FL methodology no actual CSEM data is exchanged among the *network nodes (NN)*. Rather, the *master node* (EUROPOL) collects the updates from all distributed sources (= *MS nodes*) and combines them in a *global NN model* for each targeted functionality (or data type). These model updates transferred through the network are more focused on the learning task at hand than on the CSEM raw data (i.e. they contain strictly no additional information about the user (= *MS node*), and, therefore, are typically significantly less in size compared to CSEM raw data), and the individual model updates only need to be held ephemerally by the server. However, the capability of maintaining a customised version of each tool/NN at every node will be provided. By applying the FL approach to the challenge of optimising analysis and information flow, the GRACE solution enables cooperation between LEAs by improving their own capabilities and harnessing experiential knowledge.

Preventing Model Inversion and Reconstruction

Deep Neural Network architectures like the one of the GRACE FL platform represent also a form of memory mechanism, incorporating high-level and compressed representations of the input/training data stored within their weights. This aspect creates a vulnerability to attacks aiming for model inversion or reconstruction by attempting to reverse-engineer parts of the training data from the algorithm weights on decentralised nodes. This form of attacks can create severe data leakage and enable possible access to the original data by third parties.

Against this attack vector, FL provides an innovative and efficient infrastructural approach to security, especially if additional measures were adopted. In this direction, *differential privacy* is a promising framework that randomises part of a mechanism's behaviour to protect its content. For the scenario of the GRACE FL

[24] Note that Art. 4(1)(c) Directive (EU) 680/2016 requires that processing shall be adequate, relevant and "not excessive" in relation to the processing purposes, while Art. 5(1)(c) GDPR and Art. 28(1)(c) Europol Regulation require adequate, relevant and "limited to what is necessary" in relation to the processing purposes.

platform, the mechanism is identified as the learning algorithm.[25] The motivation behind adding randomness to a learning algorithm is to make it impossible to reveal behaviour aspects that correspond either to the model and the learned parameters, or to the training data. Without adding any such randomness, adversaries could claim answers either regarding the parameters that are required for the learning and convergence procedures on standard datasets, or regarding the probabilities in which the learning algorithms will choose parameters within a set of possible learning parameters for a specific dataset. The implementation of differential private techniques is capable of eliminating this risk. Generally, *differential privacy* may be separated into *Local Differential Privacy* (LDP), and *Global Differential Privacy* (GDP):

- **Local Differential Privacy** (LDP) allows statistical computations by simultaneously protecting the individual users' privacy. Additionally, LDP ensures that no trusted party is required, since the individual users are responsible for adding noise to their own data before sharing it.
- **Global Differential Privacy** (GDP) techniques create a *central aggregator* (i.e. *trusted curator*) with access to the raw data. Specifically, each individual user sends their data to the *central aggregator* without adding extra noise. The *central aggregator* considers the input data and transforms it with a differentially private mechanism by adding noise. When an untrusted (third) party addresses the trusted *central aggregator* with a specific query, an answer will be provided which is mathematically impossible to be reverse-engineered so that it is impossible to know the precise answer concerning the raw data.

In general, GDP systems tend to be more accurate since all the analysis is implemented on noise-free data, and only a small amount of noise is added at the end of the process. However, the efficiency of global privacy models depends entirely on the amount users' trust to the *trusted curator*.

The solution for the GRACE FL platform is intended to achieve the maximum of security features. This typically requires not only composing many of the tools and technologies into an end-to-end system, but also using different strategies to protect different parts of the GRACE system. As the GRACE project evolves, adjustments on the methodologies of the privacy and encryption aspects within the FL will be implemented, in order to meet all envisioned objectives for the GRACE architecture.

[25] However, differential privacy techniques can be applied to any algorithmic formulation.

4. Chain of Custody

There is also the risk that the data and intelligence contained in the GRACE system are modified. Such modification could be accidental or intentional nurturing of any kind of ethically unacceptable bias. At the heart of this phenomenon is the authenticity of the data contained in the GRACE system. This fundamental requirement for the admissibility of any evidence constitutes the legitimacy of evidence. This fundamental principle requires digital evidence to be collected, analysed, preserved and ultimately presented in court in accordance with the appropriate procedures and without violating the fundamental rights of the suspected individual.[26]

4.1. Legal Issues at Stake

"Chain of Custody" is – especially in common law countries - an important principle when it comes to the admissibility of evidence in court.[27] It requires that the data which is copied or removed during an investigation be retained in the state in which they were found at the time of the seizure and remain unchanged during the time of criminal proceedings.[28] If data processed by the GRACE solution should be used as evidence in court, it will be essential not only to ensure the "chain of custody" but also be able to provide records that prove it.

4.2. Ethical Values at Stake

The ethical dimension of preventing harm from privacy necessitates adequate data governance aiming to ensure the quality and integrity of the data used as well as monitoring the data's relevance, processing and access protocols.[29]

Further, the ethical requirement of transparency demands clear information about all human decisions taken at the time of the GRACE system's development regarding the data, the system and the business model.[30] The datasets and the processes yielding the GRACE system's decisions including those of data labelling, data categorisation and selection of algorithms need to be documented to the best possible standard to allow for traceability.[31] In this regard, transparency is closely linked to the principle of explicability which requires that all algorithmic decisions of the GRACE system can be understood by end-users in non-technical terms outlining what elements used in the (machine) learning model were responsible for each specific outcome.[32]

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.

Ultimately, transparency concerning the reasons for AI-generated decisions amounts to explicability and primarily serves to maintain *meaningful human oversight* over the decisions an algorithm makes. Such

[26] See section 6.1 of Deliverable 9.3.
[27] See: *Nemeth*, Law and Evidence, 2nd Edition, 2011, page 68.
[28] *Becker*, Criminal Investigation, 2008, 473.
[29] See section 3.3 of Deliverable 9.1.
[30] See section 3.4 of Deliverable 9.1.
[31] See section 3.4 of Deliverable 9.1.
[32] See section 3.4 of Deliverable 9.1.

meaningful human control is necessary to trace moral accountability for the outcomes of machine learning algorithms back to human beings.

4.3. Technical Solution

In general, for the prevention of any tampering with data contained in the GRACE system a technical solution could be implemented to prevent accidental or intentional modifications of documents. The system could use **hash values** (*block-chain like*) and **auditing trails** to prevent unrecognised interactions. These **hashes** could be applied to all core GRACE data and their metadata at source allowing the verification of the state of the data from the source module that acquired it. These hashes could also be cryptographically signed by a module that is creating or enriching it indicating which data they have added. This signing would allow the authenticity of the data to be verified back to its source.

At the **audit trail** level, by hashing and signing each audit line against the previous audit line it would become near-impossible to manipulate the audit (introducing new or modified data) without rewriting the entire audit - even this can be mitigated by using a separate module to initiate the first audit. It could further more flag detected irregularities for quick auditing.

More specifically and because the technical perspective is on par with the ethical and legal perspective in this regard, the requirements necessary to be observed for properly incorporating *chain of custody* into all GRACE tools and the GRACE platform are appropriately outlined in Deliverable D2.10[33] and, therefore, can be reiterated here:

"The chain of custody (CoC) implements the technological solution to track the accesses and the operations that can be done over stored resources. The scope is to prove that the originality and integrity of those resources have been maintained since the acquisition time. The integration of the CoC solution into the GRACE platform is needed in order to avoid any unauthorised data access and manipulation which could compromise current investigations.

In particular, relating the GRACE platform we have to consider two different kinds of information:

- ***Resources Files** (Text, Videos, Audio, Images) which will be acquired as potential evidence in the GRACE platform*
- ***Metadata**, which will be generated manually or after validation of analysis results.*

The chain of custody must be maintained for both resources and metadata. The chain of custody tracks the logging activities that have been performed by an actor (user or system) over stored information. This includes Create-Read-Update-Delete (CRUD) operations includes the operation done by users as well as the access information done by background services (enrichment, analyses, reasoning tools, etc.).

Additionally, for the resources files is also necessary to maintain the chain of evidence. Chain of evidence consists of the calculation of a digital mark that uniquely identifies a resource (like the hash value) every time it is modified.

[33] See section 3.8 of Deliverable D2.10.

This mechanism recalculates the hash value every time a given resource is processed and compares it to the known ones, in order to verify its integrity.

Whenever a hash value of a resource does not correspond to that computed at acquisition time, this means that the resource has been corrupted in an unauthorized way. Chain of custody, together with a chain of evidence in case of resources, enables the forensic reuse of information stored into the GRACE system."

Implementation

The decentralised nature of the GRACE FL platform's architecture complicates data curation to ascertain the integrity and quality of the results in the *Staging Environment*. Consequently, extended research is performed in the *Development Environment* to determine the optimal method for updating the central model state (i.e. distributed optimisation, federated averaging, etc.).

5. Audit Trail

In contrast to the chain of custody, the audit trail allows retrospectively to trace all activities within and allocate responsibility.

5.1. Ethical Values at Stake

The ethical requirement of accountability is closely linked to the principle of fairness and demands mechanisms to be put in place to ensure responsibility and accountability for an AI system and its outcomes throughout the entire AI system's life cycle.[34] This responsibility must always be assigned to a moral agent or a legal person and is particularly important in the law enforcement domain where it means holding individual human officers as well as (their) units and LEAs responsible for effectively delivering the basic services of crime control and maintaining order.[35] Within the law enforcement ecosystem, LEAs are permanently monitored by superior government branches of the executive and law enforcement is constantly observed by the public for their ethical and legal behaviour which is essential for the public's trust in law enforcement at societal level.[36]

5.2. Technical Solution

Regarding the requirements of accountability, the technical perspective is on par with the ethical and legal perspective, so that it is appropriate to reiterate here the key guidance for audit trail established in Deliverable D2.14:[37]

"Auditing is the main mechanism with which compliance is monitored. For this purpose, Europol has established the "Policy on the control of retrievals", which sets the requirements on how every user action which accesses personal data is logged and audited. The "Policy on the control of retrievals" describes two objects which are needed in order to perform reliable and meaningful auditing; audit logs and audit trails.

Audit Logs

All user actions need to be audited. The 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[38] and the relevant implementing rules. Any application used to process personal data shall log activities related to the access of data that it controls.

It shall be possible to ascertain from audit logs the following information as a minimum:

[34] See section 3.6 of Deliverable D9.1.
[35] See section 3.6 of Deliverable D9.1.
[36] See section 3.6 of Deliverable D9.1.
[37] See section 3.1.3. of Deliverable D2.14.
[38] 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.

- *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 are accessed or consulted;*
- *The identification of the user, such as User Name 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;*
- *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;*
- *Logon and logoff attempts to the application and their outcome.*

Audit Trails

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 particular on the level of:

- *A web-server level, whose primary function is to serve content to the client; e.g. Apache, Microsoft IIS, etc;*
- *A database level, whose primary function is to store the data used by the application;*
- *A server level, whose functions vary depending on the services provided to the application by the server (such as file server, platform to install the web server, DHCP server, DNS server, NTP server, etc);*
- *Authentication services used to access systems processing personal data, such as IAM;*
- *Network equipment used to transmit data (such as proxies, routers, firewalls, etc). For any technical component participating in a transaction linked to retrieval of information, where possible, audit trails shall record the following information:*
- *User Name 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.*

Furthermore, where possible, for each technical component participating in a transaction participating in a transaction linked to retrieval of information, the following actions on the technical component shall be logged:

- *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."*

6. Bias Detection

An area seemingly beyond the technical perspective is the need to ensure that appropriate considerations of fairness and non-discrimination are built into the GRACE system. The incorporation of such considerations poses a problem at technical level because scalable automated methods to detect and combat discriminatory decision-making require clear-cut rules or quantifiable thresholds. In contrast, the notion of fairness and non-discrimination has historically been specified contextually according to the details of the case and defined in European jurisprudence by judicial intuition, instead of by statistics.[39]

The (judicial) interpretative flexibility is not a 'bug' of the notion of fairness and non-discrimination, but rather intentional and essential. Therefore, the technical perspective is vital in providing statistical evidence as well as developing tools for the detection of bias and measuring fairness, while the legal concept of *"contextual equality"* needs to be guaranteed and exercised by the judiciary, legislators and regulators.[40] In order to combine these strengths of both, the technical and the ethical (and legal) perspective, the GRACE system would need to be designed with an *'early warning system'* for automated discrimination which produces consistently the types of statistical evidence necessary for a human (and ultimately judicial) detection of unfairness and discrimination.[41]

6.1. Ethical Values at Stake

Emanating from the ethical principle of fairness, the requirement for fair and equal treatment demands compliance with the right to non-discrimination[42] and calls for inclusion and diversity throughout an AI system's entire life cycle.[43] Automated decisions may not be taken based on discriminatory or unjust attributes.[44]

6.2. Technical Solution

The development and the design of the GRACE tools and platform could incorporate measures which together amount to an effective *'early warning system'* for unfairness and discrimination.

The forensic analysis tools for the GRACE system will be developed and trained with non-CSE specific data and biases could easily slip in through selections in the training data or in the tool's design. Because each specific tool may either involve or lead to a trade-off concerning fairness and non-discrimination, the GRACE system could include a mechanism for monitoring its behaviour and results closely for potential changes due to the input of real CSEM report content data and their use. The notion of fairness and non-discrimination requires that the GRACE system will be rigorously audited continuously.[45]

[39] Wachter/Mittelstadt/Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI", 3 March 2020, arXiv/2005.05906, p. 1 (p. 44).
[40] Wachter/Mittelstadt/Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI", 3 March 2020, arXiv/2005.05906, p. 1 (p. 46).
[41] See section 3.5 of Deliverable D9.1.
[42] Art. 20 EU-Charter.
[43] See section 3.5 of Deliverable D9.1.
[44] See section 3.5 of Deliverable D9.1.
[45] See section 3.5 of Deliverable D9.1.

7. Restrictions for Targeted OSINT Crawler

The GRACE platform and tools are envisioned only for analysing and categorising and managing the data contained in the CSEM reports. From a purely investigative point of view however, among the first steps of an investigation is the verification of facts followed by an update of the evidence which typically includes a search for potential fresh evidence regarding the investigated suspect(s) and victim(s). Therefore, it would appear helpful for LEAs if the GRACE platform could, at some stage, be combined with some tools for searching the surface web as well as the dark web. Once a CSEM report is uploaded onto the GRACE system, such tools could automatically either

(i) verify the data contained in the CSEM report and update as well as supplement the CSEM reports data with fresh sources,[46] or

(ii) separately search even for new CSE related content creating new CSEM reports of its own[47].

Because of the investigative necessity to verify and the convenience to update and expand the data contained in a CSEM report at some stage, it appears technologically viable and beneficial for LEAs that the GRACE system may be combined with such a search tool at some point. The technological design of the GRACE system appears at least open for being combined with such automated search tools for investigative evidence.

If the GRACE system was combined with an automated search tool as currently suggested to the Consortium,[48] such automated search tool would have to filter the information collected by using search terms. 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.[49]

7.1. Ethical Values at Stake

If the GRACE system were to include tools for external searches in the surface web and the dark web, the use of the GRACE system by LEAs would affect where individuals can reasonably expect to be private. For maintaining citizen's trust in the policing authorities, also covert activities like the collection of information online requires transparency by informing and gaining permission before using any kind of search term. This is all the more necessary because placing suspicion on innocent people behaving in such a way as to fit a profile for affiliation to CSE activities is ethically fair only if inflicted to no more than the extent proportionate and necessary to fight against CSE.[50]

Closely intertwined with maintaining citizen's trust is the ethical requirement of *human agency* demanding that the GRACE system should only support human individuals in their decision making.[51] Humans not only outperform AI systems in areas like common-sense reasoning, but also in recognising the bigger picture and adapting to unusual situations.[52] The design of the functionalities has to allow for *meaningful human oversight*

[46] See section 4.1. of Deliverable D9.1.
[47] See section 4.2. of Deliverable D9.1.
[48] See "T3.1 Memo - targeted crawling of open source information", 19 May 2021.
[49] See section 4.2.6. of Deliverable D9.1.
[50] See sections 4.2.3. and 4.2.4. of Deliverable D9.1.
[51] See section 3.1 of Deliverable D9.1.
[52] See section 3.1 of Deliverable D9.1.

to ensure that the GRACE tools and platform support LEAs in making faster, better, and more considerate decisions in prioritising their investigations based on CSEM reports in their fight against CSE.

Suitable selectors in a search term have to be sufficiently flexible to respond to suspects who are forensically aware and aim to avoid the use of incriminating language.[53] Specific keywords or group names may also enjoy prominence in the press.[54] Concerning the use of names of particular individuals, the number of false positives has to be factored into any assessment of proportionality of the intrusion by an automated search tool within the GRACE system.[55]

Against this background, the appropriateness of the selectors becomes extremely sensitive for the ethical and societal acceptability of a LEA using an automated search tool in combination with the GRACE system. Especially because the GRACE tools and platform incorporate various elements of machine learning, the training data for these self-learning functionalities will significantly evolve once the GRACE tools and platform are rolled out and in use by LEAs. Therefore, the use of selectors in a search term constitutes a "*capital issue*" for development and the later commercialisation of the GRACE tool and platform.

7.2. Technical Solution

A technical solution could be implemented to automatically check searches before they are executed, so that they can be entered in audit log or even prevented. At the current stage of the GRACE project, the Consortium has neither decided whether nor where best to integrate an automated search tool in the GRACE solution and for what purpose.

Purpose

Concerning the purpose of an automated search tool, the following three options are currently discussed for data acquisition:[56]

(i) *Verification:* The automated search tool merely verifies the (continued) availability of pre-defined elements (i.e. email addresses or user accounts) of a CSEM report in publicly available sources.

(ii) *Verification & Extension:* The automated search not only verifies the availability of pre-defined elements like in option (i), but also uses any links provided in a CSEM report as potential source for extending the content data of the CSEM report.

(iii) *Verification, Extension & External Search:* The automated search verifies the CSEM report data like in option (i), extends it like in option (ii) and uses external search engines to discover any other relevant information available online.

Any of these three options would add new data to a CSEM report because even mere verification of already existing CSEM data adds the information that the data are still available. While option (i) seems to be the least invasive concerning the fundamental rights of all victims and any potential suspects mentioned in a CSEM

[53] See section 4.2.6. of Deliverable D9.1.
[54] See section 4.2.6. of Deliverable D9.1.
[55] See section 4.2.6. of Deliverable D9.1.
[56] Presented in the order of increasing concerns from a legal and ethical perspective.

report, option (ii) increases the risk of automated encroachment on the fundamental rights of these individuals and option (iii) involves the maximum data collection by using a CSEM report as trigger for automatically searching the entire surface web and dark web for any additional relevant information. However, even if an automated search was only to verify a first name or surname of combination thereof, such names of particular individuals would inevitably cause a number of false positive search results the desirability of which seems questionable and would, at the very least, have to be factored into any assessment of proportionality of such an intrusion.

Processing Stage of CSEM Reports

An equally strategic and practical consideration for a technical solution is the question at what stage of processing a CSEM report by the GRACE solution any additional search should be triggered, either automatically or manually. In this respect, it is currently discussed to incorporate an automated search tool:

- Either in the central GRACE data acquisition tool,
- Or in the set of GRACE tools at Member State level.[57]

At either stage, the search could be restricted by specific characteristics of a CSEM report, i.e. containing already known CSEM or relevance for selected countries. The strictest restriction would be to allow such additional searches only to be performed manually by human data analysts of a competent LEA.

Next Steps

The discussion about the ideal integration of an automated search tool in the GRACE solution has just started in May 2021.[58] One benefit of such integration would be a synergy effect with the results of the EU-funded AviaTor project[59] which has developed a Targeted Online Research as optional functionality for the AviaTor solution. The GRACE Consortium is determined to thoroughly explore the advantages and disadvantages of all options from all perspectives to find the ideal solution for the GRACE system. The result of this discussion will be presented in Deliverable D9.8. and may possibly already be included in the second iteration of the "Technical and Architecture Specifications" in Deliverable D2.11. and the "Security and Auditing Mechanisms" in Deliverable D2.15.

[57] See section 3. of "T3.1 Memo - targeted crawling of open source information", 19 May 2021.
[58] See "T3.1 Memo - targeted crawling of open source information", 19 May 2021.
[59] ISF-P grant no. 821841.

8. Conclusion

8.1. Summary

This document has described the architecture for technical safeguards to be implemented in the GRACE system from an ethical and legal perspective. Five critical scenarios have been identified for each of which the ethical values at stake as well as guidance for a technical solution have been provided. Three of these scenarios (unauthorised access, chain of custody and audit trail) are already in view of technical perspective. The critical scenario of bias detection still has to be added, while the implementation of a targeted OSINT crawler can only be added once the Consortium has decided whether and how to embed such an automated search tool in the GRACE system.

8.2. Evaluation

This is the first of ultimately two versions of the "Architecture for Technical Safeguards" realised in the GRACE system. Because the technical development is still at a very early stage, this Deliverable D9.7 can introduce only general guidelines for the further development. Once the requirements and design of the GRACE platform have matured, the second version of the "Architecture for Technical Safeguards" will refer to the concrete technical solutions implemented in the GRACE system.

8.3. Future Work

This version merely represents the starting point of the "Architecture for Technical Safeguards" ensuring security and privacy by design in all GRACE tools and the GRACE platform. The next step will be to add the critical scenario of bias detection to the list of requirements for the GRACE system's compliance by design and to find a decision within the GRACE Consortium whether and how to embed a targeted OSINT crawler in the GRACE system.

ANNEX I - ACRONYMS

Term	Definition / Description
AI	Artificial Intelligence
COC	Chain of Custody
CSE	Child Sexual Abuse and Exploitation
CSEM	Child Sexual Exploitation Material
DoA	Description of Action
FL	Federated Learning
H-LEG	High-Level Expert Group
ML	Machine Learning
NN	Neural Network
LEA	Law Enforcement Agency
OSINT	Open Source Intelligence

Table 3 - Glossary and Acronyms