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

Overall Legal and Ethical Framework v2

Deliverable number	D9.6	
Version:	2.0	
Delivery date:	July 2023	
Dissemination level:	PU	
Classification level:	Non classified	
Status	Final	
Nature:	Report	
Main author(s):	Prof. Dr. Marco Gercke Ulrich Gasper	CRI CRI
Contributor(s):	Thalia Prastitou Pedro Vicente Patrick De Smet Panagiotis Daousis Ntres Sigute Stankeviciute	EUC PJ NICC Europol L3CE

DOCUMENT CONTROL

Version	Date	Author(s)	Change(s)
1.0	31/12/2021	Ulrich Gasper, Thalia, Prastitou, Pedro Vicente, Patrick De Smet, Panagiotis Daousis Ntres, Sigute Stankeviciute	Submission of v1 of this document: D9.5 – Overall Legal and Ethical Framework v1
1.1	08/07/2023	Marco Gercke	Considerations for not updating D9.5 in chapter 1.
2.0	28/07/2023	Peter Leskovsky	Deliverable 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.

© Copyright in this document remains vested in the Project Partners

Table of Contents

Annexes

1. Introduction

Considerations related to the decision not to update D9.5:

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 Overall Legal and Ethical Framework (D9.5) the situation is different. The Overall Legal and Ethical Framework contains importance actionable guidance for all partners of the GRACE Consortium. The partners of the GRACE Consortium participating in WP9 continuously monitored potentially relevant developments related to the range of content addressed in D9.5. If such relevant developments had been identified, a process to update D9.5 would have been implemented immediately and prior to the due date of D9.6. This monitoring and update process was established to ensure that the guidance provided in D9.5 for all partners of the Consortium was accurate at any stage of the lifespan of the GRACE project. However, no relevant development requiring the need to initiate an update process has been identified. Therefore, the decision was taken to leave the content of D9.5 unchanged and only add this explanation in D9.6.

1.1. Overview

The DoA describes this Deliverable as:

> *D9.5 – This deliverable will define the overall legal and ethical framework of the GRACE project. [month M19]*

The description of the related Task T9.3 provides the following details:

> *T9.3 – Operating in line with ethical and legal standards is a top priority. Having identified that the research/development process on the one side, and the tool/operation on the other side, have unique needs with regard to legal and ethical aspects, the project will address them separately. T1.3 will deal with a focus on the research/innovation work of the consortium while T9.1-T9.4 will address legal and ethical concerns related to the tool and its operation. Based upon the work carried out under T9.1 and T9.2, the overall legal and ethical framework will be developed. While T9.1 andT9.2 identify the issues the focus of T9.3 will be on concrete recommendations related to the tool. The framework will include concrete recommendations for the project with regard to the issues identified in T9.1 and T9.2 and thereby provide guidance with regard to the overall project, the development of the GRACE capability and input related to guidelines for end users (T9.3).*

The main objective of this Deliverable D9.5 is to provide concrete recommendations related to the GRACE tools and system as well as their operation. While Deliverable D1.4 has provided Societal, Ethical, Legal and Privacy (SELP) guidelines for the research conducted in the course of the GRACE project, this Deliverable D9.5 addresses the concerns related to the GRACE tools as well as to the use and operation of the GRACE system. This Deliverable

D9.5 predominantly aims to remind and enhance Consortium Member's awareness of their responsibility as developers for the legal and ethical compliance of the GRACE tools and system by providing an overview of the full range of legal and ethical concerns emanating from the scientific basis elaborated in Deliverables D9.1 and Deliverable D9.3.

1.2. Approach

As described in section 2.5 of Deliverable D9.1, the "Ethics Guidelines for Trustworthy AI"[1] presented by the High-Level Expert Group on Artificial Intelligence (AI H-LEG)[2] in April 2019 provide the most comprehensive guidance for the development, deployment and use of AI systems. The tools and platform to be developed in the course of the GRACE project will not only contain AI components. Nevertheless, this guidance provides a most valuable framework for addressing the ethical and legal impact of autonomous systems. As their foundation, the "Ethics Guidelines for Trustworthy AI" identify the following four ethical principles and their correlated values that must be respected by any technology based on autonomous and automating algorithms (AI system):

- the principle of **respect for human autonomy**: AI systems should be designed to augment, complement and empower human cognitive, social and cultural skills (human centric design principles) and to secure human oversight over the work processes;

- the principle of **prevention of harm**: the operation of AI systems must not only be safe and secure, but also technically robust and not open to malicious use;
- the principle of **fairness**: in a substantial dimension, AI systems have to be free from unfair bias and discrimination and, in a procedural dimension, an entity accountable for a decision must be identifiable; and
- the principle of **explicability**: the capabilities and purpose of an AI system as well as its processes must be transparent and its decisions – to the extent possible – must be explainable.[3]

These fundamental ethical principles have to be balanced against each other and it is for the developers and end-users within the GRACE Consortium in concert with its Ethics Board to achieve the most appropriate equilibrium for the GRACE tools and platform (= the GRACE system).[4] In order to help developers and end-users with this task, the AI H-LEG elaborated seven key requirements that an AI system should meet in order to be trustworthy:

- human agency and oversight,[5]
- technical robustness and safety,[6]

[1] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[2] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.
[3] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, pp. 11-13, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[4] When tensions arise between the four fundamental ethical principles, it is for the developers and end-users to approach such ethical dilemmas and trade-offs via reasoned and evidence-based reflection. See: AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, pp. 13, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[5] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 15, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[6] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 16, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

- privacy and data governance,[7]
- transparency,[8]
- diversity, non-discrimination and fairness,[9]
- societal and environmental wellbeing[10] and
- accountability.[11]

Developers of AI systems are expected to implement and apply these seven requirements to their design and development processes, while *deployers* should ensure that their product and services meet these requirements and end-users should be informed accordingly and able to request that they are upheld.[12]

For each of the seven key requirements to become operational, the "Ethics Guidelines for Trustworthy AI" provide a concrete and non-exhaustive assessment list that has to be tailored to the specific use case.[13] In July 2020, the AI H-LEG made available a refined and final version of this list entitled Assessment List for Trustworthy AI (ALTAI)[14]. The Assessment List for Trustworthy AI (ALTAI) is intended for self-evaluation purposes and aimed at provoking appropriate action and nurturing an organisational culture committed to the protection of fundamental rights as enshrined in the EU Treaties and the EU Charter.[15] This Deliverable D9.5 aims to apply and tailor each of the seven key requirements to the envisioned GRACE system and provide practical recommendations for the GRACE Consortium.

In addition, this Deliverable D9.5 provides practical recommendations concerning the legal challenges regarding privacy and data protection, electronic evidence, searches in databases and the use of a search crawler.

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
D1.3	Ethical and legal guidelines for the project and data management and	Both cover data protection issues

[7] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 17, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[8] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 18, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[9] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 18, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[10] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 19, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[11] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 19, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[12] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 14, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[13] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, pp. 24-31, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[14] High-Level Expert Group on Artificial Intelligence (AI H-LEG), Assessment List for Trustworthy Artificial Intelligence (ALTAI), 17 July 2020.

[15] High-Level Expert Group on Artificial Intelligence (AI H-LEG), Assessment List for Trustworthy Artificial Intelligence (ALTAI), 17 July 2020, p. 3-4.

	protection plan	
D1.4	SELP guidelines for GRACE	Both address similar topics – however with D1.4 focusing on practical aspects of the research
WP2 deliverables	DESIGN - Use Cases, Requirements, Standardisation, Technical and Architecture Specification, Security and Auditing	Deliverables submitted in WP2 so far have tried to design the first version of GRACE platform in compliance with legislation.
D2.10	Technical and Architecture Specification v1	While D2.10 approaches technical safeguards from a technical perspective, D9.5 addresses technical safeguards from an ethical and legal perspective.
D2.14	Security and Auditing Mechanisms Report v1	While D2.14 approaches guidelines for technical safeguards from a technical perspective, D9.7 addresses technical safeguards from an ethical and legal perspective.
D9.1	Ethical Report	Some of the Ethical Aspects also have a legal implication
D9.7	Architecture for technical safeguards – “security and privacy by design” v1	Both address privacy by design
D10.6	Stakeholder and policy recommendations for addressing online CSEM	Links legislation with policy and phenomenon which will help as a guidance for the reader and the designers of the platform

Table 1 – Relation to other deliverables – receives inputs from

Provides outputs to:

Deliv. #	Deliverable title	How the two deliverables are related
D2.2	Use Cases, Process and Data Flows Refinement v2	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D2.3	Use Cases, Process and Data Flows Refinement v3	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D2.5	User requirements v2	General implication of the D9.5 deliverable
D2.6	User requirements v3	General implication of the D9.5 deliverable
D2.11	Technical and Architecture Specifications v2	General implication of the D9.5 deliverable
D2.12	Technical and Architecture Specifications v3	General implication of the D9.5 deliverable
D2.13	Technical and Architecture Specifications v4	General implication of the D9.5 deliverable

D2.15	Security and auditing mechanisms report v2	General implication of the D9.5 deliverable – especially with regard to the security related issues addressed in D9.5
D2.16	Security and auditing mechanisms report v3	General implication of the D9.5 deliverable – especially with regard to the security related issues addressed in D9.5
D2.17	Security and auditing mechanisms report v4	General implication of the D9.5 deliverable – especially with regard to the security related issues addressed in D9.5
D3.2	Data acquisition module v2	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D3.3	Data acquisition module v3	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D3.5	Data pre-processing module v2	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D3.6	Data pre-processing module v3	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D3.8	Data loading and mapping module v2	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D3.9	Data loading and mapping module v3	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D3.11	Content management and digital evidence tamper detection module v2	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D3.12	Content management and digital evidence tamper detection module v3	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D4.11	Digital evidence tamper detection module v2	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D4.12	Digital evidence tamper detection module v3	General implication of the D9.5 deliverable – especially with regard to data protection issues and crawlers
D5.2	Federated data annotation tools	General implication of the D9.5 deliverable – especially with regard to AI

D5.3	Report on Federated Learning strategies	General implication of the D9.5 deliverable – especially with regard to AI
D5.4	Secure data exchange mechanism	General implication of the D9.5 deliverable
D6.1	Module(s) to perform cross-matching and entity mapping between referrals	General implication of the D9.5 deliverable
D6.2	Module(s) to perform content analysis and classification	General implication of the D9.5 deliverable
D6.3	Module(s) to perform content-based geo-location	General implication of the D9. deliverable
D6.4	Module(s) to perform analysis of knowledge graphs for evidence data fusion	General implication of the D9.5 deliverable
D6.5	Module(s) to perform prioritisation on OSP referral data	General implication of the D9.5 deliverable
D6.6	Module(s) for predictive analysis of short and long-term trends in CSEM	General implication of the D9.5 deliverable
D7.4	GRACE System v2	General implication of the D9.5 deliverable
D7.5	GRACE System v3	General implication of the D9.5 deliverable
D7.6	GRACE Collaborative Application v1	General implication of the D9.5 deliverable
D7.7	GRACE Collaborative Application v2	General implication of the D9.5 deliverable
D7.8	GRACE Collaborative Application v3	General implication of the D9.5 deliverable
D7.9-D7.14	Technical Validation Report v1 – v6	General implication of the D9.5 deliverable
D8.8	Report on pilot’s execution v2	General implication of the D9.5 deliverable
D8.9	Report on pilot’s execution v3	General implication of the D9.5 deliverable
D8.10	Report on pilot’s evaluation & assessment v1	General implication of the D9.5 deliverable
D8.11	Report on pilot’s evaluation & assessment v2	General implication of the D9.5 deliverable
D8.12	Report on pilot’s evaluation & assessment v3	General implication of the D9.5 deliverable
D9.2	Ethical Report v2	General implication of the D9.5 deliverable

D9.4	Legal Report v2	General implication of the D9.5 deliverable
D9.5	Overall legal and ethical framework v1	General implication of the D9.5 deliverable
D9.8	Architecture for technical safeguards – “security and privacy by design” v2	General implication of the D9.5 deliverable
D10.7	Stakeholder and Policy Recommendations for Addressing Online CSEM v2	General implication of the D9.5 deliverable

Table 2 – Relation to other deliverables – provides outputs to

1.4. Structure of the Deliverable

This document includes the following sections:

- Section 2 focuses on practical recommendations for the GRACE Consortium how to implement and ensure human agency and oversight sufficiently in the design, development and application of the GRACE tools and platform.
- Section 3 focuses on practical recommendations for the GRACE Consortium how to implement and ensure technical robustness and safety sufficiently in the design, development and application of the GRACE tools and platform.
- Section 4 focuses on practical recommendations for the GRACE Consortium how to implement and ensure transparency sufficiently in the design, development and application of the GRACE tools and platform.
- Section 5 focuses on practical recommendations for the GRACE Consortium how to implement and ensure diversity, non-discrimination and fairness sufficiently in the design, development and application of the GRACE tools and platform.
- Section 6 focuses on practical recommendations for the GRACE Consortium how to implement and ensure the principle of societal and environmental wellbeing sufficiently in the design, development and application of the GRACE tools and platform.
- Section 7 focuses on practical recommendations for the GRACE Consortium how to implement and ensure the principle of accountability sufficiently in the design, development and application of the GRACE tools and platform.
- Section 8 focuses on practical recommendations for the GRACE Consortium how to implement adequate data governance ensuring privacy and data protection sufficiently in the design, development and application of the GRACE tools and platform.
- Section 9 focuses on practical recommendations for the GRACE Consortium on how to implement and ensure in the design, development and application of the GRACE tools and platform that the data

processed by the GRACE system will be able to meet national legal and technical requirements for admissibility and acceptability as evidence in court.

- Section 10 focuses on practical recommendations for the GRACE Consortium on how to implement and ensure database solutions in the design, development and application of the GRACE tools and platform which respect not only privacy guidelines but also the best practices for database design and operations.
- Section 11 focuses on practical recommendations for the GRACE Consortium concerning the potential integration of a tool for performing automated searches in the GRACE system. Because the he discussion whether and how to integrate an automated search tool (= crawler) in the GRACE system is still ongoing, the practical recommendations distinguish between the use of a crawler in individual investigations only (section 11.1. below) and for content related to CSE and CSEM in general (section 11.2. below). In both scenarios a crawler would be searching the surface web as well as the dark web.

2. Human Agency and Oversight

The use of AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy.[16]

2.1. Human Agency

In the interest of *human agency*, the overall principle of user autonomy demands that an AI system should only support individuals in making better, more informed choices in accordance with their goals.[17] For that purpose, the user needs to be provided sufficient understanding of the AI system enabling the user not only to interact but also to reasonably self-assess or challenge the AI system.[18]

- The GRACE Consortium has to be mindful that the GRACE system is designed to interact, guide and take decisions by human end-users that ultimately will affect humans and society. Therefore, the GRACE Consortium should consider to develop mechanisms to ensure that any end-user of the GRACE tools and platform will become adequately aware that a decision, content, advice or outcome presented by the GRACE system is the result of an algorithmic decision so that there can be no confusion about which results can be verified in a CSEM report referred to law enforcement in the EU and which results were generated and added by the GRACE system.
- The GRACE Consortium should consider developing and putting reliable procedures in place to prevent end-users from over-relying on the GRACE system and their decision-making process from being interfered with in any other unintended and undesirable way.
- The GRACE Consortium should consider incorporating measures in the GRACE tools and platform to mitigate the risk of manipulating the behaviour of the end-user instead of providing a solid basis for a human decision.

2.2. Human Oversight

Human agency can be achieved by *human oversight* which helps ensuring that an AI system does not undermine human autonomy or causes any other adverse effects. *Human oversight* requires a governance mechanism allowing meaningful human control such as a Human In The Loop (HITL), a Human On The Loop (HOTL) or a Human In Command (HIC) approach.[19]

The set of GRACE tools and the GRACE platform will allow LEAs to delegate the analysis and prioritisation of the content of all CSEM reports to an automated system which can be seen as a significant gain in efficiency. However, humans not only outperform AI systems in areas like common-sense reasoning, but also in recognising the bigger picture and adapting to unusual situations.[20] The GRACE tools' and platform's functionality of automatically

[16] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 15, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

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

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

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

[20] Zardiashvili/Bieger/Dechesne/Didgnum, "AI Ethics for Law Enforcement", *Delphi 4/2019*, p. 19.

analysing, categorising and prioritising the content of all CSEM reports involves a certain loss of *human oversight* and plays into the increasing human nature of conveniently delegating decisions to machines. This appears to suggests the following mitigation measures:

- The GRACE Consortium needs to establish *meaningful human oversight* for each functionality of a GRACE tool and the GRACE platform as well as for the GRACE system as a whole. For the oversight to be *meaningful*, it must provide a human with the time, ability and knowledge to intervene. The less oversight a human can exercise over an AI system, the more extensive testing and stricter governance mechanisms are required.[21] These governance mechanisms have to include periodical review of the functioning of the system, risk management and assessment of ethical and legal compliance.
- The GRACE Consortium should determine and document for each functionality the governance mechanism most appropriately ensuring such *meaningful human oversight* without defeating its benefits. Whereas a Human In The Loop (HITL) could create a speed bottleneck, the governance mechanisms of having a Human On The Loop (HOTL) or having a Human In Command (HIC) may seem preferable. While HOTL processes enable human intervention during an AI system's design cycle and monitoring the AI system's operation, HIC processes enable a human to oversee the overall activity of an AI system and to decide whether, when and how to use the AI system.
- The GRACE Consortium should consider establishing a comprehensive documentation of the education and specific training required for any human involved in a governance mechanism (HITL, HOTL, HIC) concerning a GRACE tool, the GRACE platform or the GRACE system as whole so that these humans can exercise *meaningful* oversight.
- The GRACE Consortium should consider to develop and establish suitable detection and response mechanisms for undesirable adverse effects of the GRACE system for the end-user. Such mechanisms should include something like a 'stop button' or a procedure to safely abort an operation when needed.
- The GRACE Consortium should consider to develop specific oversight and control measures to reflect the self-learning and/or autonomous nature of a GRACE tool or the GRACE system.

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

3. 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.[22] In this section a more detailed discussion on the issue of implementing technical robustness and safety is given. This section builds upon earlier discussion as, e.g., reported on in Deliverable D9.1. "Ethical Report". With respect to Deliverable D9.1 it should be noted that in that deliverable considerable attention has been paid to references and issues discussing, e.g., AI/ML related ethical considerations. In particular, artificial intelligence (AI) and machine learning (ML) deficiencies or attacks indeed exist that relate to the received input data (*data poisoning*), (implicit/unknown weaknesses in) either or both the data and the AI model or the underlying infrastructure, both software and hardware[23].

3.1. Overall Goal

The GRACE tools and platform will not only contain AI components. As such, any and all components of the GRACE system should thus undergo similar and equally thorough scrutiny; i.e., technical robustness and safety should be implemented for the full processing chain, starting in particular with data ingestion and any related data pre-processing steps (WP3). Given the nature of the GRACE project, this does not mean, however, that a full in-depth analysis and implementation of all possible considerations should be realized.

The overall goal, therefore, will be to make a careful risk analysis and prioritization of these tasks that should also be discussed and decided on together with the GRACE LEA partners and Europol. Additionally, the desired or technical feasibility of realistic and achievable Technology Readiness Levels (TRLs) for GRACE tools and components should be taken into account. Clearly, these activities will require collaboration between, e.g., WP2 (User Requirements), WP7 (Technical Validation) and WP9.

3.2. Appropriate Stress Testing

Robustness and security aspects should not only focus on the security of the code and tools during software development; e.g., by only making use of secure input and parsing functions, scrutinized use of already (existing) externally developed tools and libraries, etc. Rather, additional stress testing would be required. For example, tools should be tested for their robustness against unusual data inputs and variations or combinations thereof. Ultimately, this approach could then also escalate to the direct application of *anti-forensics methods* and related purposely created *poisonous data*. Well-known examples of such methods include, e.g., detection and mitigation of the "42.zip" or similar archive bombs[24], or, anti-forensic methods to fool certain anti-virus tools[25], etc.

Particular attention could or should be given to any possible sensitivities with regard to *digital image* and *video inputs*; e.g., not fully standard conforming codec usage or data formats, irregular image sizes, etc.

Another data ingestion related example that may need further study in GRACE, is the use of *traditional binary hashing algorithms*. As explained in Deliverable D3.10 "Content management and digital evidence tamper detection module", it is nowadays relatively straight-forward to generate MD5 hashing collisions. This means that it would thus be possible to generate many different files (containing CSEM content) that might yield an identical

[22] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 16, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[23] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 16, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[24] See https://en.wikipedia.org/wiki/Zip_bomb.
[25] See https://en.wikipedia.org/wiki/Anti-computer_forensics.

hash signature value. In turn, this could then result in an increased risk that these files may remain undetected or not fully processed if the data deduplication module would simply consider these files as redundant.

The *chain of custody* and other core system components should also be scrutinized for its proper functioning and robustness as already indicated in D7.3 GRACE system.

The examples given above are included here merely to illustrate how any and all (other) tools within GRACE should undergo a proper *"passive" risk and impact analysis* during the remainder of the project.

In order to implement a *"more" active approach*, core system components and resources could also be adapted and instrumented further to allow for their online monitoring and management (including e.g. CPU/GPU, I/O, storage usage and bandwidth monitoring tools, etc.).

3.3. Range of Measures

In order to ensure various other aspects related to tool robustness and safety the following measures should be evaluated:

- manual safety and robustness review of full processing flows for different types of cases (NCMEC referrals that may contain known problematic, partial or incorrect data); this approach may be generalized into:
- composition of unit tests with accompanying data or parameter settings that will test for proper handling of incorrect data, incomplete data or other possible processing issues, etc.
- use of unit or larger scale test scenarios based on unit testing described in tool documentation/design plans, and variations thereof.
- having a user accessible tool for signalling issues encountered with the tool or system during processing, which could thus trigger/drive a more detailed manual review process both during development and during real-life use.
- testing strategies for graceful degradation under variations of inputs (incl. parameters), outputs, internal parameters or limitations: e.g., simulating high load or starvation of hardware resources.
- stress testing, using a *white hat hacker* or *red team attack* approach; e.g., a user entering faulty data and/or (API) commands on purpose (also simulating real-life accidental mistakes).
- documenting or automatically extracting and listing of code or module dependencies. For individual tools such a list could possibly then be correlated with security related reports (e.g. CVE reports) etc.
- optimization, refactoring and trimming of APIs and code size as well as its embedded dependencies. This approach would be specifically aimed at reducing the size of the possible "attack surface" and any possibilities that might impact erroneous or non-robust behaviour.

3.4. Generation of Confidence Levels

Additionally, as noted by *Dechesne* et al. in their White Paper, "AI Ethics at the Police"[26]:

> "*Robustness can be enhanced by improving a system's ability to generalize to new situations, or by letting it calculate a confidence score about its decisions so it can alert a human (although these confidence scores may themselves be less reliable in novel situations), or by otherwise mitigating the impact of errors.*

[26] *Francien Dechesne/Virginia Dignum/Lexo Zardiashvili/ Jordi Bieger*, white paper: "AI & Ethics at the Police: Towards Responsible use of Artificial Intelligence in the Dutch Police", March 2019.

> *Regular maintenance and updates can avoid AI systems becoming outdated, but it's important to always test new versions thoroughly*."

Thus, for any tools that would produce a confidence level, a proper procedure for fuzzy evaluation of the obtained results *and* their related confidence measures may thus be valuable or become a requirement. In particular, if any (AI/ML) or other tools would adapt or be adapted (by developers) to compensate for changes in the amount, type, or any other data or execution environment related aspects, it seems logical to assume that this may generate some "natural" drift or changes in obtained results and/or confidence measures. Nevertheless, a careful, possibly periodic re-evaluation of these changes should be considered as *low-rate data poisoning attacks* or *accidental error accumulations*, etc., might also induce at first sight seemingly similar harmless effects (as anticipated for normal system evolution).

3.5. Measures Against Re-Victimization and Over-Exposure

Another aspect that may need further consideration is not only the technical safety of the system, but also the safety of the victims and the LEA staff members involved. More specifically, the system should be maximally secured in order to avoid possible re-victimization. In particular as the Europol system may evolve into one of the largest databases containing CSEM related content, any accidental or purposely generated access or data leaks should obviously be avoided at all cost. As such, the use of independent overall system and administration monitoring or oversight should be considered, thus also providing an answer to the well-known security paradigm *"Quis custodiet ipsos custodes?"*[27].

In terms of staff members working with CSEM, it would be advisable to provide sufficient psychological relief and support (organized at a collegial and professional level). As such the GRACE system could, e.g., keep track of the number of images or videos (and their duration) that a certain user may be exposed too.

Finally, it should be noted that ethical and legal aspects related to suspects or perpetrators should be taken into consideration. Hence, robustness and safe processing for any information that relates to them should also be given proper consideration.

3.6. Work in Progress

During the remainder of the GRACE project, the aim of Task T9.3 in collaboration with other Work Packages will be to:

- Update, prioritize and evaluate (e.g. technical feasibility of) all technical robustness and safety considerations discussed above.
- Provide, as also indicated in D3.1, a more detailed robustness and safety analysis of the AI/ML and federated learning methods used in GRACE, including their design and limitations, and their training and testing strategies, in order to mitigate any possible issues. A collaboration with other recently launched projects such as ALIGNER[28], POP AI[29], STARLIGHT[30], may also be of interest to the GRACE project and its partners.

The further updating and reporting on the points listed above will be provided in Deliverable D9.6 "Overall legal and ethical framework v2".

[27] See: https://en.wikipedia.org/wiki/Quis_custodiet_ipsos_custodes.

[28] ALIGNER project: https://cordis.europa.eu/project/id/101020574.

[29] Pop AI project: https://cordis.europa.eu/project/id/101022001.

[30] STARLIGHT project: https://cordis.europa.eu/project/id/101021797.

4. Transparency

A crucial component of achieving trustworthiness for the GRACE system is transparency. The requirement of transparency demands clear information about all human decisions taken at the time of an AI system's development regarding the data, the system and the business model.[31] The data sets and the processes yielding an AI 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.[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. 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) learning model were responsible for each specific outcome.[33] Besides (i) traceability and (ii) explicability, the ethical dimension of transparency also encompasses (iii) open communication about the limitations of the GRACE tools and platform in order to build trust in the GRACE system.

4.1. Traceability

Focussing 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.[34] 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. The GRACE Consortium should, therefore, consider developing and documenting measures that address the traceability of the GRACE system during its entire lifecycle:

- 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.[35]
- 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.
- The GRACE Consortium should consider to develop and establish measures to continuously assess the quality of the output(s) of the AI system which could take the form of a standard automated quality

[31] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 18, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419; INTERPOL-UNICRI, "Towards Responsible AI Innovation", Second Report on Artificial Intelligence for Law Enforcement, 2020, p. 34.

[32] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 18, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[33] INTERPOL-UNICRI, "Towards Responsible AI Innovation", Second Report on Artificial Intelligence for Law Enforcement, 2020, p. 34.

[34] Whitaker/Crawford/Dobbe/Fried/Kaziunas/Mathur/Myers West/Richardson/Schulz/Schwartz, "AI Now Report 2018", p. 5 et seq.; Dechesne/Didgnum/Zardiashvili/Bieger, "AI Ethics at the Police", White Paper, March 2019, p. 11; Samek/Wiegand/Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", 2017, arXiv/1709.08296.

[35] AI H-LEG, "Assessment List for Trustworthy Artificial Intelligence (ALTAI)", 17 July 2020, p. 14.

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).[36] In this respect, adequate logging practices should be put in place to record the decision(s) or recommendation(s) of the GRACE tool, platform and system.

4.2. Explicability

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. As indicated in section 4.1 above, 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.[37] This creates a need for the GRACE Consortium 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:

- *End-Users:* 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 human LEA officers. 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.

    ➢ For the *development phase*, the GRACE Consortium follows a Co-Creation[38] and Co-Design approach based on a constant and intensive dialogue between the technical partners and the LEAs which includes seven workshops[39] aimed at combining the views of those who create and those who consume the results of the GRACE tools and platform. In these workshops, the developers of the GRACE tools and platform address together with LEAs the end-user interactions with the GRACE tools and platform. The GRACE should be mindful that this direct involvement of LEAs reveals the level of explanation needed by a LEA officer to understand the decision(s) of the GRACE system.

    ➢ For the *after roll-out phase*, the GRACE Consortium should consider to incorporate mechanisms in the GRACE system which continuously survey the end-users whether and how they understand the decision(s) of the GRACE tools and platform.

- *Along Chain of Authorisation:* Because the individual end-users at a LEA are part of a particular unit and organisation within the law enforcement ecosystem, the GRACE Consortium should consider to prepare explanations providing sufficient insight into the use of the GRACE system according to the level of confidentiality along the chain of authorisation.

[36] AI H-LEG, "Assessment List for Trustworthy Artificial Intelligence (ALTAI)", 17 July 2020, p. 14.

[37] Dechesne/Didgnum/Zardiashvili/Bieger, "AI Ethics at the Police", White Paper, March 2019, p. 12 and 14.

[38] Pieters, Maarten; Jansen, Stefanie (2017). The 7 Principles of Complete Co-Creation. Amsterdam: BIS Publishers. p. 15.

[39] See Work Packages WP2 – WP9.

4.3. Open Communication

The third component of the ethical dimension of transparency requires to communicate appropriately an AI system's capabilities and limitations to the end-users. In this respect, it is important to explain 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 for what purpose, on the other. However, it is vital to indicate the accuracy of each tool and functionality and how a LEA can use it. Furthermore, an overview of all decisions automated by the GRACE tools and platform needs to be provided as well as an understanding of how they come about and on which criteria they are based on. This is all the more important, since the GRACE system is intended to help and suggest a prioritisation of the enriched CSEM reports which immediately affects the use of law enforcement resources.

Against this background, the GRACE Consortium should consider to develop and incorporate sound mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the GRACE tools and platform. Such mechanisms need to thoroughly communicate 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. In this respect, it seems more than recommendable to provide appropriate training material and disclaimers to end-users on how to adequately use the GRACE system.

5. Diversity, Non-Discrimination and Fairness

Closely linked to the principle of fairness, the requirement for fair and equal treatment demands compliance with the right to non-discrimination[40] and calls for inclusion and diversity throughout an AI system's entire life cycle.[41] Automated decisions may not be taken based on discriminatory or unjust attributes.[42] Regarding diversity, non-discrimination and fairness, in order to achieve trustworthy AI according to the AI H-LEG's "Ethics Guidelines for Trustworthy AI", inclusion and diversity throughout the entire AI's, and in this case the GRACE's, system life cycle shall be enabled. This requirement has three main components: the avoidance of unfair bias (see section 5.1 below), accessibility and universal design (see section 5.2 below), and stakeholder participation (see section 5.3 below).

The aim of this chapter, is to elaborate further on the meaning and content of Diversity, Non-discrimination and Fairness and more importantly to provide practical recommendations for the GRACE Consortium (the Consortium). This exercise would be completed based on the Assessment List for Trustworthy AI (ALTAI),[43] prepared by the AI HLEG. These practical recommendations aim to encourage thoughtful reflection among the Consortium members and, if needed, to provoke appropriate action for the development of a Trustworthy AI system.

5.1. Avoidance of Unfair Bias

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.*[44]

- The Consortium should consider the development of a strategy or a set of procedures in order to avoid creating or reinforcing unfair bias in the GRACE system, both regarding the *use of input data* as well as for the *algorithm design*. In detail, the Consortium should put in place processes to test and monitor for potential negative discrimination during the development, deployment and use phases of the GRACE system. More importantly it should put in place processes to address and rectify for potential negative discrimination.
- Taking into consideration, the special attention that must be ascribed to children and children's rights, the Consortium should put in place processes (a) to address and rectify for potential harm to children by the GRACE system and (b) to test and monitor for potential harm to children during the development, deployment and use phases of the GRACE system.

[40] Art. 20 EU-Charter.

[41] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 18, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419; INTERPOL-UNICRI, "Towards Responsible AI Innovation", Second Report on Artificial Intelligence for Law Enforcement, 2020, p. 33.

[42] INTERPOL-UNICRI, "Towards Responsible AI Innovation", Second Report on Artificial Intelligence for Law Enforcement, 2020, p. 33.

[43] 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.

[44] AI HLEG, ALTAI, p. 23.

- The Consortium should consider *diversity and representativeness* of different LEAs and Europol (being the end-users) in the data. At this point one can (a) test for specific target groups or problematic use cases, (b) research and use publicly available technical tools, that are state-of-the-art, to improve understanding of the data, model and performance and (c) assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the GRACE system.
- The Consortium needs to introduce *educational and awareness initiatives* to help GRACE designers and developers be more aware of the possible bias they can inject in designing and developing the GRACE system.
- The Consortium needs to introduce a mechanism that allows for the *flagging of issues* related to bias, discrimination or poor performance of the GRACE system. Regarding this issue one needs (a) to provide clear steps and ways of communicating on how and to whom such issues shall be addressed, (b) to identify the subjects that could potentially be (in)directly affected by the GRACE system, in addition to LEAs and Europol.
- According to the ALTAI Glossary "fairness refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally referred to as 'substantive' fairness. But fairness also encompasses a procedural perspective, that is the ability to seek and obtain relief when individual rights and freedoms are violated".[45] The Consortium should develop a *definition of fairness* provided that (a) various definitions of fairness are examined before choosing a specific one, as there is no single definition that applies in all cases, (b) a quantitative analysis or metrics to measure and test the chosen definition of fairness is completed and (c) mechanisms to ensure fairness in the GRACE system have been established. The chosen definition of fairness should be used and implemented in different phases of the process of setting up the GRACE system.

5.2. Accessibility and Universal Design

Accessibility: *Extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use (which includes direct use or use supported by assistive technologies).*[46]

- The Consortium should ensure that the AI system corresponds to the variety of preferences and abilities in society.
- The Consortium should involve or consult with LEAs and Europol in need for assistive technology during the planning and development phase of the GRACE system.
- The Consortium should take the impact of the GRACE system on LEAs and Europol into account (a) by confirming that its members that are involved in building the GRACE system engage, throughout its development process, with LEAs and Europol and (b) by assessing the risk of a potential unfairness of the GRACE system into the LEAs and Europol communities.
- The Consortium should assess whether there could be groups who might be disproportionately affected by the workings and outcomes of the GRACE system. Specifically, the Consortium should assess whether including a crawler in the GRACE system, that will search the internet (both surface and dark web) for additional data on persons/items of interest in order to supplement the data contained in a referral, would lead to discrimination.[47]

[45] AI HLEG, ALTAI, p. 27.

[46] AI HLEG, ALTAI, p. 23.

[47] For recommendations regarding the GRACE crawler see also section 10. below.

5.3. Stakeholder Participation

Stakeholders: *"Developers, deployers and end-users, as well as the broader society. By developers, we refer to those who research, design and/or develop AI systems. By deployers, we refer to public or private organisations that use AI systems within their business processes and to offer products and services to others. End-users are those engaging with the AI system, directly or indirectly. Finally, the broader society encompasses all others that are directly or indirectly affected by AI systems".*[48]

- The Consortium should ask for regular feedback even after deployment of the GRACE system and set up longer term mechanisms for stakeholder participation, throughout the whole process of implementing GRACE.
- The Consortium should include the participation of the widest range of possible stakeholders in the GRACE system's design and development (who may be directly or indirectly affected by the GRACE system throughout its life cycle).

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

6. Societal and Environmental Well-Being

In line with the principles of fairness and prevention of harm, society at large and the environment should also be considered as stakeholders throughout an AI system's life cycle.[49] While sustainability and ecological responsibility address the environmental friendliness of an AI system's development, deployment and use process as well as its entire supply chain, the social impact of an AI system needs to be considered at the level of an individual user and society at large.[50]

- Use of Already Existing Infrastructure: The GRACE Consortium should consider to ensure that the GRACE tools and platform should exploit as much already available infrastructure as possible for the operational coordination of LEAs across Member States. Such environmentally friendly approach should be maintained when considering resource usage and energy consumption in their development phase and after deployment.

- Usability for Individual End-User: The GRACE tools and platform aim to automate the analysis, categorisation and prioritisation of CSEM reports at a speed, scale and level of complexity that defy human capacity. While the GRACE system's suggestions on how to approach the overwhelming influx of CSEM reports reduce a stressful and time-consuming task and are likely to improve job satisfaction, the individual end-user has to understand how the GRACE system operates, which capabilities it has and which it does not have. For that purpose, the GRACE Consortium should consider to elaborate how best to train end-user's capability and skill in using the GRACE system and to establish their awareness of its limits.
- Protection of Individual End-User: The GRACE tools and platform support and harmonise the necessary evaluation of CSEM reports for police investigations. The final evaluation of a CSEM report's content data has to be carried out by human LEA officers as individual end-users. Therefore, the GRACE system has the potential to equally affect individual end-user's physical and mental well-being by exposing the human end-user to the content data of CSEM reports. The GRACE Consortium should consider to develop and integrate 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.

- Impact on Society: The vision for the GRACE system is that the overwhelming influx of CSEM reports will become much more manageable for LEAs and duplicated investigative efforts will be significantly minimised. For society at large to gain trust and potentially benefit from LEAs use of the GRACE system, the use of the GRACE tools and platform needs to lead to more CSE offenders being apprehended and more victims being rescued from their ordeal while, at the same time, reducing their risk of re-victimisation. The GRACE Consortium should consider to develop and integrate into the GRACE system mechanisms which ensure that the GRACE system stays within the margins of what seems absolutely necessary for acquiring evidence for prosecuting offenders and for first time identification of victims.

[49] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 19, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[50] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 19, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

7. Accountability

Accountability addresses the fundamental questions of who bears the responsibility for an action, choice or decision and whether there is a satisfactory justification for it.[51] This requirement 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.[52] From an ethical perspective, 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.[53] 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.[54]

In short: Accountability *"…refers to the idea that one is responsible for their action – and as a corollary their consequences – and must be able to explain their aims, motivations, and reasons"*[55].

To ensure responsibility Risk Management and Auditability mechanisms should be previewed and operative, and set on a **S.I.M.P.L.E.** environment[56]:

(1) **S**et expectations

(2) **I**nvite commitment

(3) **M**easure progress

(4) **P**rovide feedback

(5) **L**ink to consequences

(6) **E**valuate effectiveness

These six mechanisms should be put in place to ensure responsibility in all phases of the development, the deployment and/or the use of Grace solution, both in terms of teamwork underlying the project as well as regarding any outcome in the context of the Project.

7.1. Set Expectations

"It is important to set firm, clear, and concise expectations for any group. Accountability will not grow where team members are unsure of the group's purpose and vision. Teams need to know what is expected of them before they in turn can be expected to be held accountable"[57].

Given the nature and size of the GRACE project, the definition of expectations should be decentralized, starting from the most general aspects and ranging to the most particular ones, without losing sight of the overall mission and vision of the project.

On the other hand, given the variety of backgrounds among the Consortium partners, the definition of

[51] INTERPOL-UNICRI, "Towards Responsible AI Innovation", Second Report on Artificial Intelligence for Law Enforcement, 2020, p. 34.

[52] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 19, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[53] Dechesne/Didgnum/Zardiashvili/Bieger, "AI Ethics at the Police", White Paper, March 2019, p. 9.

[54] Dechesne/Didgnum/Zardiashvili/Bieger, "AI Ethics at the Police", White Paper, March 2019, p. 10.

[55] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[56] CJ Goulding - https://www.lifehack.org/articles/productivity/6-practical-ways-create-culture-accountability.html.

[57] CJ Goulding - https://www.lifehack.org/articles/productivity/6-practical-ways-create-culture-accountability.html.

expectations should adopt a common language which is understandable for all and does not rely on a supposed extra background knowledge by an element of each institution.

7.2. Invite Commitment

Divided into Work Packages, the distribution of tasks may be defined by the respective leader in terms of main assumption complemented by requests for revision; initial distribution of contributions to be integrated by the leader.

In any case, the designated leader should provide the direction he intends to give to the task in order to guide the contributing partners, promoting meetings (preferably face-to-face) in order to better convey the meaning of the contribution and better perceive the adhesion of the assigned partners.

7.3. Measure Progress

All decisions, activities, and results of teamwork should be permanently recorded and evaluated, either in the form of written documentation, or in the form of multimedia records of meetings and activities.

7.4. Provide Feedback

"After setting clear expectations, committing to set goals, and measuring progress, it is important to provide feedback to team members so that there can be improvement towards the goal"[58].

Feedback should be a permanent concern for both, the leader and the task partners. It should take the form of permanent contact and a debriefing meeting at the end of the task to discuss all procedural aspects and the final result in comparison with the initially intended one.

Feedback should also be provided at the stage of integrating the contributions of a task with a future task, of the same or another WP, in a sense this time ascending from the particular to the general.

7.5. Link to Consequences

Making public the degree of commitment and involvement of partners in a given task of the team work assigned for the GRACE Project, while providing feedback on their collaboration, generates institutional responsibility of the contributor, affecting positively or negatively their image.

Throughout the project, mechanisms must be created to allow the performance of assessments as well as to access records on said assessments. This does not necessarily imply that information about business models and intellectual property must always be openly available.

Those mechanisms implying traceability of the GRACE project's activities, the sourcing of training data and the logging of all development activities processes should be established for the benefit of both, Consortium partners as well as external monitoring entities or those directly or indirectly affected by the GRACE solution's results.

[58] CJ Goulding - https://www.lifehack.org/articles/productivity/6-practical-ways-create-culture-accountability.html.

7.6. Evaluate Effectiveness

In general terms, the architecture defined for team work in the GRACE project already allows the audit of activities and their results, with mechanisms for scheduling meetings, providing agenda information, and documenting work, as well as the preparation of contributions, reviewed by peers and by an external entity, aimed at documenting the progress of activities.

The commitment of the partners involved should be permanently evaluated, adjusting procedures and strategies in order to promote more effective and participatory activity and the possibility of reaching useful solution.

8. Privacy, Data Governance and Data Protection

As an integral part of human dignity, privacy is intertwined with the principle of prevention of harm and includes the dimension of data protection.[59] Both, the right to privacy and the right to the protection of personal data are at the core of the Police Directive[60], the General Data Protection Regulation[61] (GDPR) and the Europol Regulation (ER)[62]; rights which must be guaranteed by AI systems throughout their entire lifecycle.[63]

The three pieces of legislation, together with Regulation (EU) 2018/1725[64], represent the main body of legislation of the European Union which aims to safeguard these fundamental rights, and this chapter will present a list of recommendations that can be taken into consideration when partners of the GRACE Consortium design their compliance to the spirit and letter of these pieces of legislation. These pieces of legislation are aligned with each other, and they have distinct and complementary scopes. Since the GRACE Consortium comprises commercial entities, research institutes, LEAs and Europol, it is important to first distinguish which pieces of legislation are applicable to which type of partner and for what purpose.

GDPR is intended to apply to any processing action of personal data, both in the public and private sector. However, there are two exceptions relevant to the GRACE project:

- Firstly, processing of personal data by LEAs is excluded, when the purpose is explicitly for law enforcement purposes; in this case, the Police Directive is applicable instead of GDPR. As an example, a LEA processing personal data for the purposes of payroll falls under GDPR, but investigating or prosecuting a criminal offence falls under the Police Directive.
- Secondly, Europol is also excluded from GDPR when processing personal data for law enforcement purposes, in which case the ER applies instead. In order to provide additional context, the existence of Regulation (EU) 2018/1725 is also worth mentioning, which applies to "*EU institutions, bodies, offices and agencies*"; however, Article 2 of this Regulation excludes Europol from the Regulation's scope, until the ER is amended and adapts Regulation (EU) 2018/1725.

[59] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 17, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[60] Directive (EU) 2016/680 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, Official Journal of the EU 2016 L 119/89.

[61] Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, Official Journal of the EU 2016 L 119/1.

[62] 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
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0794

[63] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 17, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[64] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC (Text with EEA relevance.)
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R1725

This chapter presents key requirements of the GDPR in section 8.1, of the Police Directive in section 8.2, and of the Europol Regulation in section 8.3. Finally, section 8.4 takes these pieces of legislation into consideration and presents a list of practical recommendations, which aim to guide the GRACE Consortium partners on how to address compliance with this body of legislation on a technical level. It is worth noting, however, that these recommendations are not intended to substitute thorough study and understanding of these pieces of legislation, and do not function as an exhaustive list of compliance requirements. They can only serve as a general guidance providing context and direction to the Consortium partners. In order to assure compliance, partners are encouraged to engage in dialogue with their respective DPOs.

8.1 General Data Protection Regulation (GDPR)

GDPR was introduced on 27 April 2016, and as a Regulation, it is a binding piece of legislation without the need for the Member States to translate it to national legislation. The stated goal of GDPR is "*the protection of natural persons with regard to the processing of personal data and on the free movement of such data*", and is applicable to any organization processing personal data of EEA residents, and to any organization processing personal data within the EEA. GDPR introduces the rights of data subjects, the definition and responsibilities of data controllers and data processors, penalties for breach of rights, distinctions of competent authorities, provisions for transferring data between organizations or member states, etc.

While the following recommendations cannot serve as a substitute for a thorough study and understanding of GDPR, the partners can use them as a consultation list when trying to design their compliance with GDPR:

- *Data protection by design and by default:* This is one of the key requirements of GDPR, and it refers to the principle of placing data protection at the centre of any technical design or implementation. This means that the partners of the GRACE Consortium need to employ robust data protection management programs, which should include not only technical measures such as secure design, secure development and secure deployment, but also establish organizational measures which ensure the protection of a person's data, such as data access policies, secure data management policies and auditing processes.
- *Data Protection Officer:* Depending on the size of the organization, its core activities, and the nature of the data processing activity it performs within the context of GRACE, partners may be required to appoint DPOs.
- *Grounds for processing personal data:* GDPR describes the grounds under which personal data can be processed, such as consent, contractual obligation, legal obligation, etc. Partners should make sure that the legal grounds under which they will process personal data are compliant with GDPR.
- *Grounds for processing sensitive data:* GDPR also describes the grounds under which sensitive data can be processed, such as explicit consent, sensitive data manifestly made public by the data subject, etc. Likewise, partners should make sure that the legal grounds under which they will process sensitive data are compliant with GDPR.
- *Grounds for processing data of children:* GDPR introduces additional safeguards when processing data of children such as requiring the explicit consent of a parent or guardian, which partners need to comply with, if they are to process personal data of children.
- *Data discovery and report:* Partners should implement a solution which makes it possible to extract and provide all relevant data to a data subject upon request.

- *Right to be forgotten:* Partners should implement a solution were the data of a data subject can be permanently deleted upon request.
- *Change consent:* Partners should implement a solution were the data processing permissions that a data subject has given can be changed or revoked.
- *Record processing activities:* Partners should record the processing activities that process personal data, keeping information like *who* processed *what*, *when* and *why*.
- *DPIA:* Partners should consider if the type of processing activities they are going to perform warrants the use of a Data Protection Impact Assessment, before this processing action takes place.
- *Notification of data breaches:* any data breach should be communicated to the competent authority in a timely manner.

8.2 Police Directive

The Police Directive was also introduced on 27 April 2016. An EU Directive is binding to all Member States, but its aim is to provide a set of concrete goals that must be achieved, with Member States having to transpose it into their national legal framework by the date specified in the Directive. This practically means that Member States have more leeway and there may be differences in the way a Directive is transposed into national legislation, but typically the different national legislations are in broad terms still aligned across the EU. The stated goal of the Police Directive is "*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*".

As with the GDPR recommendations, the following list of recommendations is not a substitute for thorough study and understanding of the Police Directive and the relevant national legislation. Therefore, it is advised that the partners of the GRACE Consortium use the recommendations only as a guide, and consult with the local DPO to ensure compliancy to local legislation.

Since the Police Directive has been designed to be consistent with GDPR, as both are part of EU's Data Protection Reform package, some of the recommendations for the Police Directive (on high level) are identical to the ones for GDPR:

- Data protection by design and by default
- *Compliance to Directive:* The partners need to implement appropriate technical and organizational measures, in order to demonstrate that processing is performed in accordance to the Directive.
- DPO
- Record processing activities
- DPIA
- Data protection officer
- Consult the conditions to transfer data to third countries
- Notification of data breaches

There are, however, additional obligations specific to the Police Directive:

- *Distinction between data subjects:* Where possible, the personal data of data subjects should be categorised by the type of the data subjects (offender, victim, etc.).
- *Data quality:* Particularly since the GRACE system is going to integrate AI tools for processing data, It should be possible to distinguish between data generated by inferences of AI tools, and data which have been triaged and confirmed by humans.
- *Lawful processing:* Processing personal data should be *necessary* for the performance of a task executed by an LEA.
- *Special categories of personal data:* Processing special categories of personal data (religions, political opinions, etc.) should take place only when *strictly necessary*.

8.3 Europol Regulation

The Europol Regulation (ER) was adopted on 11 May 2016, and it introduced changes to the structure and organization of Europol. The following recommendations are intended as a guide for transparency and understanding regarding the scope and purpose of ER, and not as an exhaustive compliance checklist:

- *Data protection by design:* As with GDPR and the Police Directive, article 33 of the ER introduces the responsibility of Europol to perform data processing in a way that protects the rights of the data subjects concerned.
- *Processing action needs to be necessary and proportional:* The principle of proportionality is prominent across the ER, so according to articles 17, 18, 23, 30, and 31 of the ER, processing of personal data should be necessary and proportionate for the purposes for which they are processed.
- *Audit logs and audit trails:* Pursuant to Article 40 of the ER and Europol's Policy on the Control of Retrievals (EDOC #893185), any processing action of personal data should be recorded in audit logs (chronological record of activities performed on a specific technical application) and audit trails (chronological record of technical components allowing the reconstruction of a specific operation)
- *Data retention:* Storing personal data should be in line with Article 31 of the ER.
- *Development and use of AI:* As defined in Article 39 of the ER, any new type of processing activity is subject to prior consultation when (a) special categories of data as referred to in Article 30(2) are to be processed, and (b) the type of processing, in particular using new technologies, mechanisms or procedures, presents specific risks for the fundamental rights and freedoms, and in particular the protection of personal data, of data subjects.

  A prior consultation is initiated by Europol's Data Protection Function notifying EDPS of a new type of envisaged processing operation, an assessment of the risks to the rights of freedoms of data subjects, as well as safeguards, security mechanisms and measures to address those risks. EDPS will in turn provide their opinion to the Management Board of Europol, after all required information have been delivered. In case EDPS finds that the notified processing activity may involve a breach to the ER, Europol shall modify the processing activity accordingly.

  Processing personal data using AI technologies falls within the scope of prior consultation, so all AI tools developed during the GRACE project which will need to process personal data will need to be included to a prior consultation process before processing said data type, either during development or operations. For more context, the following documents should be read in conjunction with this guideline:

- Provide machine learning toolbox process description (EDOC #1160124)
- Europol Security Rules (EDOC #865874)
- Policy on the control of retrievals (EDOC #893185)

8.4 Practical Recommendations

In the light of the Police Directive, GDPR and the ER, this section introduces technical recommendations for the GRACE tools and the entire platform, as a general guide for compliance with these pieces of legislation. As mentioned before, this cannot be considered as an exhaustive list, and partners are advised to consult the local DPOs in order to assure compliance:

- Each individual GRACE tool and the GRACE platform as a whole should keep security principles such as secure design, secure development and secure code deployment at the centre of their development, deployment and run processes. These can include mitigation of risks found in the OWASP[65] list or known in industry in general, code validation and penetration testing, integration of automated security testing in the code deployment pipeline, etc.
- Addressing security and privacy by design, it is recommended that data are encrypted both in transit via TLS and other secure connection practices, and at rest via encrypting databases and file exports.
- The consent from data subjects or data owners to process personal data should be documented; depending on the implementation:
    - It should be possible for data subjects to be directly informed as to how their personal data will be processed and provide consent for these processing actions.
    - Any transfer of datasets to and between the GRACE Consortium partners should be accompanied by information about the dataset source, the method of acquisition, and a record exhibiting the data subjects' consent to the intended processing actions.
    - In cases where personal data of children are involved, it should be possible for their guardians to identify themselves and provide consent in order for children's personal data to be processed.

  These cases are only applicable to non-law enforcement use; alternatively, when LEAs or Europol require the use of personal data for law enforcement purposes, what needs to be documented is:

    - 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 should be recorded.
    - Only authorised personnel for authorized purposes should be able to process data which are intended for law enforcement purposes only.
- 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. Partners should have the

[65] https://owasp.org/

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 overseeing authorities, or 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.
- Building upon the capability of recording consent and tracing personal data of data subjects, it should also be possible for data subjects to modify or completely withdraw their consent for the use of their personal data. This is only relevant in cases where personal data are not used for the purposes of law enforcement.

  For law enforcement use, a similar capability should be offered to the LEAs who own the data.
- Every data piece should be accompanied by a "retention evaluation" or "retention expiration" date, and the GRACE platform should 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 tools and platform should make it clear to users as to which pieces of information have been inferred by automated processing activities, and which pieces of information are confirmed by users.
- There is also the need to establish a number of administrative processes:
  - Every partner should be aware of the authority to which they have to notify potential data breaches, and a process should be created for such notification to be provided to the competent authority in a timely and concise manner.
  - Before partners process 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.

9. Electronic Evidence

Every action on cyber space leaves a trace – electronic data. It potentially could be useful for law enforcement in crime investigation. However, electronic data needs to meet certain criteria in order to be accepted as judicial evidence. The national laws on criminal procedures set up the requirements for data to be admissible as evidence in court. While the precise requirements of data admissibility might differ from country to country, they can be divided into two categories: *legal* requirements (section 8.1. below) and *technical* requirements (see section 8.2. below). Whereas meeting the *legal* requirements depends on both, law enforcement and forensic science tool developers, meeting the *technical* requirements depends mainly on the forensic science tools of the developers but it is related to legal requirements which law enforcement has to meet. Failing to comply with *legal* requirements means that the electronic data will not be accepted by the court as evidence. Failing to comply with *technical* requirements means that the electronic data might be questioned in court.

9.1. Legal Requirement

9.1.1. The Appropriate Legal Authorization

Responsibility: law enforcement.

The national laws on criminal procedures regulate legitimate data collections, authorization procedures and requirements for evidence. Usually, the national laws on criminal procedures do not establish differing requirements for electronic evidence. However, it might be different in some countries. The authorising body for electronic data collection also depends on the national laws. It might be a court, a public prosecutor or the head of a law enforcement institution. In some cases, such authorisation might not be required at all. This might be the case in particular regarding the collection of *meta data*. Although legal scientists agree that meta data can also be personal data (ex. not only content of communication but also meta data of electronic communication is personal data), the authorisation of their collection might be regulated in the laws only regarding collecting the content of communication data, but not regarding the collection of meta data or their collection might underlie different procedures of authorisation. In some cases, a LEA has a right to collect information without prior authorisation. This right applies to electronic data as well. However, electronic data containing personal information should not be collected without authorisation, see Art. 8 of European Convention on Human Rights.

9.1.2. Compliance with Data Protection Requirements

Data Protection by Design

Responsibility: forensic tool developers.

The *data protection by design principle*[66] means that a technical tool is created in a such way that it ensures data protection requirements by itself throughout its entire lifecycle. The data protection issues are being implemented since the earliest stage of technology development.

[66] See Art. 20(1) Directive 2016 (EU) 2016/680 on data protection by design.

Data Protection by Default

Responsibility: law enforcement.

The *data protection by default principle*[67] means that law enforcement institution or other data controller ensures that only data strictly necessary for each specific purpose of the processing are processed by default (without the intervention of the user)[68]. It links to the principles of purpose limitation[69] and data collection minimisation[70].

Hence, the right to data protection is a fundamental human right, failing to comply with data protection requirements means the breach of European Convention of Human Rights, European Charter on Human Rights which renders the electronic evidence less acceptable. Additionally, the person who's right to privacy and data protection was violated could possibly seek for compensation of damage[71].

9.1.3. Documentation

Responsibility: law enforcement.

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[72] in order to establish the authenticity of the data and initiate the chain of custody[73]. 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[74].

[67] See Art. 20(2) Directive 2016 (EU) 2016/680 on data protection by default.
[68] https://edps.europa.eu/data-protection/our-work/subjects/privacy-default_en
[69] See Art. 4(1)(b) Directive 2016 (EU) 2016/680.
[70] See Art. 4(1)(c) Directive 2016 (EU) 2016/680.
[71] Art. 13 of European Convention on Human Rights.
[72] Council of Europe, Electronic Evidence Guide, p. 15.
[73] Casey, E. Digital Evidence and Computer Crime, 2004, 2nd Ed. Elsevier, p. 106.
[74] ISO/IEC 27037.

9.2. Technical Requirements

9.2.1. Authenticity and Integrity

Responsibility: forensic tool developer/law enforcement.

The authenticity requirement for electronic evidence means that the evidence must establish authentic facts in a way that authenticity thereof could not be disputed and is representative of its original state[75]. 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* (ex. MD2, MD4, MD5, SHA-1, SHA-2) with time could be used at each stage to prove integrity[76] of the electronic evidence and check for any errors in the evidence file. This way it would be possible, first of all, for law enforcement to identify changes if they occurred to digital evidence at any point of an investigation[77]. Secondly, it helps to (im)prove its value as evidence in court if a dispute occurs.

9.2.2. Audit Trail

The documentation of any action taken when handling electronic evidence, should be created and preserved in such a way that these actions can be audited by a third party. The results of a third party audit should be the same as what is presented in court[78].

9.2.3. Proportionality

Responsibility: forensic tool developer/ law enforcement.

The methods used to collect electronic evidence must be fair and proportionate to the interests of justice. It means that the level of intrusion or coercion interfering with the fundamental rights of an individual should not outweigh the *"probative value"* of electronic evidence as proof[79]. Therefore, when considering providing law enforcement with more powerful tools in order to meet possibilities to electronic space for crime investigation purposes, these tools and powers will have to include proper data protection safeguards ensuring compliance with fundamental rights and the principles of data protection by design and default.

9.3. List of Applicable Standards

1. ISO/IEC 27037 – guidelines for identification, collection, acquisition and preservation of digital evidence[80].

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

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

[77] 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.

[78] Council of Europe, Electronic Evidence Guide, p. 14.

[79] *Ibid.*

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

2. ISO/IEC 27041 – guidelines on assuring suitability and adequacy of incident investigation method[81].
3. ISO/IEC 27042 – guidelines for analysis and interpretation of digital evidence[82].
4. ISO/IEC 27043 – guidelines for incident investigation principles and processes[83].

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

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

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

10. Data Base Search

Pursuing the goal of assisting the investigations of LEAs in CSEM cases, the GRACE system will make extensive use of database searches for investigation activities, cross matching of extracted entities for the purposes of entity recognition or intelligence, de-duplication of CSEM reports received multiple times from NCMEC, data analytics, intelligence generation, etc. Considering the breadth and sensitivity of the data that will be searchable, the importance of establishing robust recommendations for designing and operating database solutions becomes apparent.

This chapter presents three different sets of recommendations, which collectively should guide the GRACE Consortium towards implementing database solutions that respect the various legal/privacy guidelines, as well as best practices in the areas of database design and database operations.

- The first set of recommendations originates from the domain of Information Security; the scope of these recommendations is wider than just database solutions, but it is applicable to them nevertheless. These recommendations have been introduced in D2.14 – Security and Auditing Mechanisms, so this Deliverable D9.5 will only include a brief summary in section 10.1 below; the full extent of these recommendations can be found in Deliverable D2.14, section 3.
- The second set is a list of general, high-level recommendations, which are more specific to database solutions, and encompass various best practices on designing and operating database solutions. These recommendations are described in section 10.2 below and can be considered as a list of baseline, minimum set of requirements that should be universally applied to any database solution.
- Since every database solution will store different types of data and will be used in different use cases, a further extension of recommendations might be applicable in certain cases. Therefore, section 10.3 below presents a set of questions that need to be answered for every database implementation. The answers to these questions can then be used as a guide to refine and potentially add additional recommendations.

10.1. Information Security Recommendations

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. The topic of Information Security has been elaborated in depth in Deliverable D2.14, sections 3.2.1 to 3.2.5, so this section only presents a short summary curated for the purposes of a high-level overview:

- *Security by design:* Creating a secure database design should be a part of the initial design requirements, and not an afterthought.
- *Least privilege:* The principle of least privilege is a universal best practice, also applicable for database solutions; a user should be granted only the minimum set of privileges required when accessing a database.
- *Need to know:* Users should only have access to data which they need to know, regardless of their role or their security clearance.
- *Defence in depth:* A database solution should employ multiple security layers, as this increases the effort required by a malicious party to access unauthorized information.

- *Access control:* There are various methods of controlling and restricting the access of users to a database which are considered best practices, such as implementing proper authentication and authorization mechanisms, account management, session management, etc.
- *Secure hosting:* The environment on which a database will be deployed should follow proper security principles (secure firewall configuration, network segmentation guidelines, etc.).
- *Information security incident management:* A deployment of a database solution should be accompanied by a process of responding to security incidents (data breaches, data leaks, etc.).
- *Disaster recovery management:* A deployment of a database solution should be accompanied by a process of recovering from disasters (hardware failure, software failure, loss of data, etc.).

10.2. General Recommendations

This section presents a list of recommendations which can be universally applied to any database solution, but are minimum requirements for the GRACE system:

- *Database encryption:* Encrypting a database (also known as "*encrypting data at rest*") 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.
- *Backup encryption:* Creating database backups is a common task when working with databases, for reasons ranging from testing to disaster recovery. Any database backup should be encrypted, same as the operational database itself.
- *Secure connections (TLS):* Every connection to a database should be encrypted.
- *Usage of stored procedures or database views:* When developing a database solution, the practices of using stored procedures and database views should be considered per use case, as they can provide the developers a wider range of options on managing and optimizing database operations and performance, than only direct database querying. These considerations should be under constant revision by operations personnel, when a database solution becomes operational.
- *Database Activity Monitoring solutions (DAM):* Monitoring and analysing the activity of a database is a good practice which provides operations personnel with valuable performance and security information of a database.
- *Database access auditing:* This is a security best practice, which refers to the monitoring which data have been accessed by which user/service.
- *Secure account management and secure authentication:* Account management and user authentication are core aspects of a database solution, and they need to follow security best practices.
- *Hardening:* Database hardening is the process of mitigating security risks by analysing the security vulnerabilities of a database and implementing security best practices and processes.
- *Change management and regular patching:* A robust change management processes should be implemented to handle database changes, and database software should be patched regularly.

10.3. Additional Questions

While the two previous sets of recommendations can be applicable to virtually every database implementation, additional recommendations may be applicable on a case-by-case basis in the scope of the GRACE system. This section presents a check list of questions which need to be answered per database solution, and the answers may lead to additional recommendations:

- Which types of data are searchable? Will the data contain personal information, will they contain links between persons/investigations, are the stored data created by inferences of AI tools or triaged by human users, etc.?
- Are there sensitive categories of personal data stored and accessible (biometric data, sex life, etc.)?
- Is Role Based Access Control applicable?
    - If so, what are the basic roles and what data is accessible to each?
- How is a proper distinction between results of victims and suspects ensured?
- Is fuzzy search possible?
- Are some results masked?
- Is it possible to do free search or are only predefined queries offered?

The answers to these questions can lead to a multitude of different recommendations, so this section will not document the potential recommendations which might be applicable per question; rather, these questions should be used to facilitate a discussion between the GRACE Partners involving the Project Manager so that the most fitting solution for each individual case can be achieved.

11. Use of a Search Crawler

From a purely investigative point of view, among the first steps of an investigation based on the GRACE system 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 system could, at some stage, be combined with some tools for searching externally the surface web as well as the dark web.

Originally, the GRACE project neither envisioned nor embarked on developing a tool for external searches. Rather, the focus of the GRACE project is to develop a platform and tools solely for analysing and categorising and managing the content data of referred CSEM reports. However, because of 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[84] 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 has been suggested to the GRACE Consortium in May 2021.[85] At the time of submission for this Deliverable D9.5, the discussion whether and how to integrate a search tool in the GRACE system is still ongoing and the following three different options for data acquisition[86] are considered:

(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.

Each search option could be triggered either automatically or manually and could be implemented either in the central GRACE data acquisition tool or in the set of GRACE tools for a LEA at Member State level.[87] While option (i) seems least invasive concerning the fundamental rights of all victims and any potential suspects mentioned in a CSEM 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.

Even though the GRACE Consortium has not yet decided whether to integrate a tool for external searches or not, the mere possibility of integrating such a search tool renders it useful to have practical recommendations already prepared, if only to advise the Consortium's choice whether to integrate such a search tool. For the practical recommendations, it seems appropriate to distinguish between the use of a crawler in individual investigations (section 11.1. below) and for CSE content (section 11.2. below):

11.1. Use of Crawler for Individual Investigations

This section focuses on using a crawler restricted to the preparation of an individual investigation based on a

[84] ISF-P grant no. 821841.

[85] See "T3.1 Memo - targeted crawling of open source information", 19 May 2021.

[86] Presented in the order of increasing concerns from a legal and ethical perspective.

[87] See section 3. of "T3.1 Memo - targeted crawling of open source information", 19 May 2021.

CSEM report. Taking the evidence contained in the CSEM report which has been analysed and categorised by the GRACE system as a starting point, such a crawler could verify the CSEM report's content data as in option (i) as well as update the CSEM report's content data and supplement it with additional information about fresh sources, accounts etc. for the victim(s) and suspect(s) as in option (ii). Either option would add new data to the referred CSEM report because even mere verification of already existing CSEM content data as in option (i) would add new information about the availability of these data.

From a technological perspective, such restricted use of crawler could be implemented in the GRACE system either before suggesting a prioritisation for the referred CSEM report or be made available as an optional tool for the national LEA investigating the case.

- From a legal perspective, LEAs need to be able to base their investigations on procedural instruments that enable them to take the measures that are necessary to identify an offender and collect the evidence required for the criminal proceedings. As emphasised in section 8.1 of Deliverable D9.3, there is no specific international or European legal framework addressing the use of crawlers by law enforcement which seems to advocate against basing the GRACE system's automated prioritisation of CSEM reports on the results of an automated crawler. Because it is the general national legal framework that applies to the legal evaluation of a LEA's authorisation to use a crawler and because an investigating LEA might need to follow national preferences, the GRACE Consortium should consider integrating the use of the crawler as optional tool for the investigating LEA, especially since any verification of sources might need to be refreshed at a later stage for investigative measures.
- While a CSEM report referred to law enforcement presents sufficient evidence for human officers at LEAs to start a proper investigation, the crawler's element of automation reduces the amount of human agency. A human officer has better common-sense reasoning, a chance to recognise the bigger picture and unusual context. Therefore, a human officer might select only particular parts of the evidence for investigation which may also have to take place in a strategic sequence. The GRACE Consortium should therefore consider to develop for such a crawler a governance mechanism which ensures meaningful human oversight (see section 2.2. above) at the investigating LEA.

11.2. General Use of Crawler for CSE Content

If the use of a crawler was not dependant on individual investigations based on referred CSEM reports, then the general use of such a crawler would tend to monitor the surface web as well as the dark web for any potential content related to CSE and CSEM. The use of such an automated search tool may be intended to automate investigations for leads which currently need to be carried out by humans. For that purpose, such automated search tool should provide reliable data concerning CSEM or CSE activities which may serve as evidence (justifying at least further investigations if not already reliable in court) and trigger further action by a LEA. 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.

Whether an automated online search tool (= crawler) will be embedded in the central GRACE data acquisition tool or in the set of GRACE tools at Member State level,[88] the recommendations in the following subsections should be taken into account:

11.2.1. Evaluation of the Crawler's Evidence

The quality of the evidence produced by such a crawler[89] suffers from a risk of inconclusiveness, a risk of the algorithm's inscrutability and potential bias.[90] As minimum safeguard, the search results of such a crawler would have to be evaluated by a human LEA officer before they may trigger any decision or action based on them. Such a 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;[91] awareness of the ethical risks and dangers involved with the inscrutability of algorithms;[92] awareness that meaning is not self-evident in statistical models and that the explanation of any correlation requires additional justification;[93] and awareness of the risk of unfair discrimination by or based on the profiling by the crawler.[94] In the light of this, the GRACE Consortium should consider to develop appropriate training material for LEA officer's use of such a crawler.

11.2.2. Monitoring the Crawler's Effectiveness

At the time law enforcement would use such a crawler, the probability of a CSE-related harm is usually neither established as high nor does it seem very likely that a source for it might be established. As a consequence, the effectiveness of crawler's use appears in doubt while the probability of intruding deeply and without sufficient justification into the lives of individuals who are not involved in CSE seems rather high.[95] Therefore, the GRACE Consortium should consider to develop mechanisms allowing LEAs to monitor the crawler's effectiveness so that they will have an evidence base from which to draw for future decisions about its use in operations.

11.2.3. Suitability of Selectors

The crawler would filter the gathered information automatically on the basis of selectors. While automatically excluding the vast majority of material from intrusive further inspection by a LEA, such selectors are difficult to establish ethically acceptable.[96]

A keyword used as selector for the crawler should be reasonable, evidence-based and non-discriminatory. In addition, it has to be born in mind that suitable selectors have to be flexible enough to respond to suspects

[88] For an overview of the options discussed for the GRACE system, see: section 7.2. of Deliverable D9.7.

[89] Should such a crawler employ external search engines, the points made here extend to their results as well.

[90] See section 4.2.2 of Deliverable D9.1.

[91] See section 4.2.2.1 of Deliverable D9.1.

[92] See section 4.2.2.2 of Deliverable D9.1.

[93] See section 4.2.2.3 of Deliverable D9.1.

[94] See section 4.2.3 of Deliverable D9.1.

[95] See section 4.2.5 of Deliverable D9.1.

[96] See section 4.2.6 of Deliverable D9.1.

who are forensically aware and aim to avoid the use of incriminating language.[97] Specific keywords or group names may also enjoy prominence in the press.[98] 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 the crawler.[99] Against this background, the GRACE Consortium should consider to ensure that the appropriateness of a selector for the crawler would be determined by at least the following considerations:

- How likely are CSE offenders and their associates to be using those terms?
- What is the ratio between such suspicious potentially CSE-related users of a selector like a specific group name and the innocent people using the same term?
- How easy is it to distinguish between suspicious potentially CSE-related users and innocent users of those terms once the selection has been made?

11.2.4. Avoidance of Chilling Effects

For the use of the crawler by law enforcement, it seems a valid evaluation that the stronger the evidence is, the more justified appears the use of highly stigmatising measures of suspicion.[100] Placing suspicion on innocent people behaving in such a way as to fit a profile for affiliation to CSE activities is undeserved, but not ethically unfair if inflicted only to the extent proportionate and necessary to fight against CSE. The right not to be stigmatised as suspicious has to be balanced against the need for LEAs to have sufficient powers at their disposal to be able to prevent and investigate CSE activities. These powers must be sufficiently broad to allow LEAs to cast a net wide enough to catch CSE offenders and to pursue tentative leads. The GRACE Consortium should consider to develop mechanisms for LEA officers using the crawler not to 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.

11.2.5. Illegal Content

One general area of concern regarding automated searches is that it could lead to the collection of illegal content. With regard to some types of illegal content – namely CSEM – the mere possession is illegal and could, therefore, lead to criminal investigations against the operator of the crawler. Within the GRACE project, this risk seems less relevant as the crawler is utilized in the context of CSEM investigations and by authorized LEAs.

However, there are potential concerns with regard to other categories of illegal content such as violent extremism and terrorism. It should be pointed out that the degree of criminalization and, therefore, the potential concerns related the accidental collection of the text and audio-visual material containing violent extremisms and terrorism is not equivalent to the level regarding CSEM (in which case the mere possession in many jurisdictions is a crime). There are two key considerations:

First of all, the crawler would aim to enable LEAs to search and crawl both the surface and dark web for CSEM and CSE-related content. This could take place for the purpose of collecting information about group dynamics

[97] See first difficulty regarding selectors discussed in section 4.2.6 of Deliverable D9.1.
[98] See second difficulty regarding selectors discussed in section 4.2.6 of Deliverable D9.1.
[99] See third difficulty regarding selectors discussed in section 4.2.6 of Deliverable D9.1.
[100] See section 4.2.7 of Deliverable D9.1.

or planning of CSE offences. However, the publication of some of the content may by itself be illegal – even if it does not lead to CSE. The prosecution of offences related to illegal content could in some cases be easier than to prove the involvement in a broader plot. If the latter fails, it could be interesting for LEAs to have the possibility of prosecuting on the basis of illegal information. To support this possibility, the GRACE Consortium should consider developing a filter capable of identifying illegal material and thereby allowing to separate this illegal material from other content.

The second consideration for the GRACE Consortium is a potential criminal liability for material collected by the crawler. It is possible that the crawler collects illegal material. Already in 2008 there were reports that terrorists used child pornography websites to exchange information.[101] Some criminal law systems do explicitly de-criminalize acts undertaken by law enforcement officials. However, such de-criminalisation might be limited to LEA officers acting for specific purposes only.

The collection and exchange of information would be fundamental components of the GRACE system. Within the design of a crawler, the GRACE Consortium should be mindful about the fact that the collection and exchange of information may include illegal material and, despite harmonisation approaches, the criminal law systems of the Member States show differences when it comes to the criminalisation of illegal content and to preventing law enforcement officials from being prosecuted for interacting or exchanging such material.

11.2.6. Copyright Issues

If the crawler is designed in a way that it collects large quantities of content, such collection process could go along with risks related to copyright violations. The crawler might copy and save in a database content that is protected by copyright laws. This issue is among the most frequently discussed legal issues related to web-crawlers used by search engines. However, the legal framework for the use of web-crawlers by search engines cannot simply be transferred to crawlers utilized by LEAs and it is, therefore, not possible to refer to search engines operating web-crawlers when discussing the legal basis.[102] In the light of this, the GRACE Consortium should consider to develop mechanisms to ensure that the crawler would avoid potential copyright violations.

[101] Tibbetts, Terrorists use child porn to exchange information, The Telegraph, 10.10.2008.

[102] See section 8.5 of Deliverable D9.3.

12. Conclusion

12.1. Summary

This Deliverable D9.5 has provided an overview of the full range of legal and ethical concerns emanating from the scientific basis elaborated in Deliverables D9.1 and Deliverable D9.3. With the help of this overview the members of the GRACE Consortium can increase their awareness of their responsibility as developers for the legal and ethical compliance of the GRACE tools and platform as well as the entire GRACE system as a whole.

In section 2 the practical recommendations have presented guidance for the GRACE Consortium on how to integrate adequate elements of human agency (section 2.1. above) and human oversight (2.2.) in the design, development and application of the GRACE tools and platform.

In section 3 the practical recommendations based on the overall goal (section 3.1. above) have presented guidance for the members of the GRACE Consortium on how to integrate not only appropriate stress testing (section 3.2. above) but also a range of measures (section 3.3. above) ensuring technical robustness and safety in the design, development and application of the GRACE tools and platform. Furthermore, the benefits of generating confidence levels to compensation for changes (section 3.4. above) have been pointed out and the need for measures against re-victimisation and over-exposure (section 3.5. above) has been highlighted, while the work in progress (section 3.6. above) has been outlined.

In section 4 the practical recommendations regarding transparency have presented guidance for the GRACE Consortium on how to implement and ensure the auditability of the GRACE system by integrating traceability mechanisms documenting the methods used for its development (section 4.1. above). Furthermore, the need for producing various degrees of explanations has been addressed 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 (section 4.2. above). Especially for all end-users, sound mechanisms have been called for to inform all users about the purpose, criteria and limitations of the decision(s) generated by the GRACE tools and platform.

In section 5 practical recommendations have presented guidance for the GRACE Consortium on how to implement and ensure diversity, non-discrimination and fairness sufficiently in the design, development and application of the GRACE tools and platform. These recommendations have especially focused on the avoidance of unfair bias (section 5.1. above), the accessibility and universal design (section 5.2. above) and stakeholder participation (section 5.3. above).

In section 6 the practical recommendations have reminded the GRACE Consortium of the environmental need to integrate in and build the GRACE tools and platform on already existing infrastructures and of the end-user's two most urgent needs (usability & protection) as well as the societal need for keeping the acquisition of evidence to what seems absolutely necessary.

In section 7 the practical recommendations have presented guidance for the GRACE Consortium on how to ensure that the design, development and application of the GRACE tools and platform create the responsibility and accountability necessary for its use in the law enforcement ecosystem by setting expectations (section 7.1. above), inviting commitment (section 7.2. above), measuring progress (section 7.3. above), providing feedback (section 7.4. above), linking to consequences (section 7.5. above) and evaluating effectiveness (section 7.6. above).

In section 8 the practical recommendations for the GRACE Consortium have addressed the need for data protection compliance of the GRACE Consortium and the GRACE tools and platform by drawing on GDPR for the technical partners (section 7.1 below), while drawing on the Police Directive for LEAs (section 7.2 below) and on the Europol Regulation for Europol (section 7.3 below).

In section 9 the practical recommendations for the GRACE Consortium on how to implement and ensure in the design, development and application of the GRACE tools and platform that the data processed by the GRACE system will be acceptable as evidence in court, have been in divided into two categories: Whereas fulfilling the *legal* requirements (section 8.1. above) depends on the developed capabilities of the GRACE tools and platform as well as the on how the GRACE system will be used by a LEA, the fulfilment of the *technical* requirements (section 8.2. above) depends predominantly on the forensic science tools of the developers.

In section 10 the practical recommendations have presented guidance for the GRACE Consortium on how to implement and ensure database solutions in the GRACE system respecting privacy guidelines as well as best practices for database design and operations.

In section 11 practical recommendations for the GRACE Consortium have been included concerning the potential integration of a tool for performing automated searches in the GRACE system. Because the discussion whether and how to integrate an automated search tool (= crawler) in the GRACE system is still ongoing, the practical recommendations have been focused on the use of a crawler in individual investigations only (section 11.1. below) as well as on searches for content related to CSE and CSEM in general (section 11.2. below).

12.2. Evaluation

The ethical and legal issues identified and analysed in Deliverables D9.1 and D9.3 have rendered the risks faced by the GRACE project very significant. The suitable mitigation of these risks requires careful evaluation and diligent management by every single member of the GRACE Consortium. The practical recommendations provided in this first iteration of the overall legal and ethical framework as Deliverable D9.5 have been oriented on the key principles elaborated in AI H-LEG’s “Ethics Guidelines for Trustworthy AI”. These key principles already include legal aspects which have been supplemented by practical recommendations on electronic evidence and database searches.

The practical recommendations regarding the use of a crawler either in individual investigations only or in general for CSE content have been included as part of the ongoing discussion within the GRACE Consortium whether and how to include a crawler searching the surface web and the dark web in the GRACE system. Only once the GRACE Consortium has decided to integrate a crawler in the GRACE system and opted for a particular option of data acquisition by the crawler, the practical recommendations for its development and use can be tailored more specifically.

12.3. Future Work

As the design of the GRACE tools and platform will evolve, the practical recommendations for the GRACE Consortium will be constantly adjusted culminating in the section iteration of this overall legal and ethical framework in Deliverable D9.6.

ANNEX I - GLOSSARY AND ACRONYMS

Term	Definition / Description
AI	Artificial Intelligence
AI H-LEG	High-Level Expert Group on Artificial Intelligence
ALTAI	Assessment List for Trustworthy AI
CRC	Cyclic Redundancy Check
CSE	Child Sexual abuse and Exploitation
CSEM	Child Sexual abuse and Exploitation Material
DAM	Database Activity Monitoring
DPO	Data Protection Officer
DPIA	Data Protection Impact Assessment
ER	Europol Regulation (EU) 2016/794
ETL	Extract, Transform, Load
GDPR	General Data Protection Regulation (EU) 2016/679
HIC	Human In Command
HITL	Human In The Loop
HOTL	Human On The Loop
ISP	Internet Service Provider
LEA	Law Enforcement Agency
ML	Machine Learning
NCMEC	National Center for Missing & Exploited Children
P2P	Peer-to-Peer
SELP	Societal, Ethical, Legal and Privacy
TLS	Transport Layer Security
TRL	Technology Readiness Level

Table 3 - Glossary and Acronyms