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

Grace

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

D9.1 Ethical Report

Deliverable number	D9.1	
Version:	1.0	
Delivery date:	March 2021	
Dissemination level:	PU	
Classification level:	Non classified	
Status	Final DRAFT	
Nature:	Report	
Main author(s):	Ulrich Gasper (CRI)	
Contributor(s):	Rimantas Zylius (L3CE), Patrick De Smet (NICC)	

DOCUMENT CONTROL

Version	Date	Author(s)	Change(s)
0.1	06/02/2021	Ulrich Gasper (CRI)	TOC and first draft
0.2	22/02/2021	Rimantas Zylius (L3CE)	Refinement + peer review comments
0.3	02/03/2021	Patrick De Smet (NICC)	Refinement + peer review comments
0.4	06/03/2021	Ulrich Gasper (CRI)	Updating + final draft
0.5	14/03/2021	Ulrich Gasper (CRI)	Final adjustments
1.0	16/03/2021	Ulrich Gasper (CRI)	Final version after PC + QA reviews – ready for submission.

DISCLAIMER

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

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

Table of Contents

Annexes

No se encuentran elementos de tabla de ilustraciones.

Figures

1. Introduction

1.1. Overview

The DoA describes this Deliverable as:

> *D9.1 - This deliverable will define the potential ethical concerns related to use of Big Data, Machine Learning and AI with regard to investigations concerning child sexual exploitation and abuse material (CSEM). [M33, draft M7]*

The main objective of this Deliverable is to cover the full range of possible ethical concerns related to the interaction of Law Enforcement Agencies (LEAs) with Big Data and techniques in the field of machine learning and AI relevant for the GRACE project.

Special attention is given to the nature of the data (child sexual exploitation and abuse material - CSEM), the victims (children) and cross-border exchange of information among law enforcement and data protection.

For the preparation of this Deliverable, several methods including a red-teaming exercise have been used to ensure that the identification of relevant societal and issues for consideration is going beyond state-of-the-art.

1.2. 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	Data Management Plan	Deliverable D1.3 points out the need for special attention to ethical concerns related to collecting and processing data. Deliverable D9.1 addresses ethical issues related to data collection, data processing and data management.
D2.1	Use Cases, Process and Data Flows Refinement	The use cases and the data flow described in Deliverable D2.1 along with the key technologies for the future GRACE platform provide the basis giving rise to the discussion of ethical concerns in Deliverable D9.1.
D8.1	Pilots scenario definition	Deliverable D8.1 depicts the initial pilot scenario for the GRACE platform based on story telling approach listing the technical and expertise needs. The prototype of the GRACE platform will evolve in iterative pilot executions each of which provide inputs to adjust the development based on ethical needs emanating from the ethical concerns discussed in Deliverable D9.1.

Table 1 – Relation to other deliverables – receives inputs from

Provides outputs to:

Deliv. #	Deliverable title	How the two deliverables are related
D1.4	Social, Legal, Ethical, and Privacy (SELP) guidelines for GRACE	While Deliverable D9.1 expounds the ethical foundation and all concerns regarding law enforcement's use of Big Data, Machine Learning and AI in investigations concerning CSEM, Deliverable D1.4 derives from these foundations and concerns concrete and practical guidelines for the GRACE project.
D9.3	Overall Legal and Ethical Framework	Based on the issues identified in the Ethical Report of Deliverable D9.1 and the Legal Report of Deliverable D9.2, Deliverable D9.3 will develop a framework with concrete recommendations related to the GRACE tools and platform as guidance for the development of their functionalities and the overall project including input related to guidelines for end-users.

Table 2 – Relation to other deliverables – provides outputs to

1.3. Structure of the Deliverable

This document includes the following sections:

- Section 2. on the "Ethical Standard for the Development and Operation of GRACE Platform": This section describes the challenges of online CSEM evolution which law enforcement faces today (2.1) and then presents the vision for GRACE tools and federated platform (2.2). After an initial overview of the ethical issues identified for GRACE (2.3), the main focus is on the examination against which ethical standard the development, deployment and use of GRACE tools and platform among law enforcement in the EU need to be measured (2.4), which not only identifies the "Ethical Guidelines for Trustworthy AI" as relevant standard (2.5) but also looks at the prospect how this standard may be expected to be implemented (2.6).
- Section 3. on "Ethical Issues Concerning Tools and Platform": In this section, the seven key requirements for an AI system to be trustworthy serve as key criteria for the evaluation of law enforcement's use of the GRACE tools and platform. The analysis, therefore, examines the conditions under which the tools and platform establishing the envisioned GRACE system meet the requirements of human agency and oversight (3.1), technical robustness and safety (3.2), privacy and data governance (3.3), transparency (3.4), diversity, non-discrimination and fairness (3.5), societal and environmental well-being (3.6) and accountability (3.7).
- Section 4. on "Ethical Issues Concerning Automated External Searches": The GRACE platform and tools are envisioned only for analysing and categorising and managing the data contained in the CSEM reports. Because of the investigative necessity to verify and the convenience to update the data contained in a CSEM report with fresh online content related to CSE and CSEM at some stage, this section examines whether monitoring the surface web as well as the dark web with an automated search tool for any content related to CSE and CSEM might be feasible (4.1) or whether an automated

search tool should be restricted to verifying and updating the existing content a CSEM report and could perhaps supplement an individual investigation of with fresh content regarding the CSEM report's victim(s) and suspect (4.2).

- Section 5. Conclusion: This section summarises the results elaborated in the analyses in the previous sections 2. – 4. (5.1) and points out the next steps in light of the legislative Proposal for an Ethical Framework which is already scheduled for the first half of 2021 (5.2).

2. Ethical Standard for Development and Operation of GRACE Platform

This section first describes the challenge law enforcement is faced with (see 2.1. below), then presents the vision for GRACE tools and federated platform (see 2.2. below) and examines against which ethical standard the development, deployment and use of GRACE tools and platforms among police forces in the EU need to be measured (see 2.3. below).

2.1. Challenges of Online CSEM Evolution

The online dimension of child sexual abuse offers offenders a way to interact with each other on the surface web as well as on the dark web for obtaining Child Sexual Exploitation and abuse Material (CSEM). The production, dissemination, possession and accessing of CSEM is one of the most serious forms of victimisation of children. One third of Internet users worldwide is estimated to be children and adolescents under the age of 18 years creating an expanding pool of potential victims.[1] While facilitating made-to-order services offering offenders to request the production of content according to their sexual preferences regarding the child's age, gender, race and appearance,[2] the digitisation has led new forms of Child Sexual abuse and Exploitation (CSE) to emerge like live-streaming services offering offenders access to a stream for observing and directing the abuse of a child in real time.[3] The commercialisation of online CSE has become so widespread as to indicate an emerging threat[4] and live-streaming has already become mainstream.[5]

[1] Because two-thirds of the world's nearly three billion Internet users live in the Global South where the proportion of children in the population is far higher than in the Global North, the projected growth in Internet users will include a rising portion of children. See: Livingstone/Carr/Byrne, "One in Three: Internet Governance and Children's Rights", Innocenti Discussion Paper No.2016-01, UNICEF Office of Research, 2016 Florence, p. 16 et seq.

[2] UNODC, "Study on Effects of New Information Technologies on the Abuse of Children", (2015), p. 21.

[3] UNODC, "Study on Effects of New Information Technologies on the Abuse of Children", (2015), p. 23.

[4] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 40, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

[5] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 39, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

The amount of detected online CSEM has grown exponentially over the last decade[6] and the COVID-19 crisis has caused an extra surge in the online distribution of CSEM.[7] While the digital world has created a global market for CSEM,[8] reports indicate that the EU has become the largest host of CSEM globally (from more than half in 2016 to more than two thirds in 2019).[9] Combating CSE, including the production and dissemination of CSEM, is a priority for the EU[10] and one of the three crime priorities of the 2020 Operational Action Plan for the European Multidisciplinary Platform Against Criminal Threats (EMPACT).[11]

An invaluable source in the fight against CSEM online originates from the obligatory case reporting of CSEM by social media providers in the USA to the National Center for Missing and Exploited Children (NCMEC)[12] and in Canada to the National Child Exploitation Coordination Centre (NCECC)[13]. The exponential growth of the number of these CSEM reports discovered by providers themselves or reported to them by their users shows no signs of stabilising, let alone declining.[14] When the NCMEC or the NCECC receives a case report involving foreign jurisdictions, the case report is referred to the relevant national Law Enforcement Agencies (LEAs) depending on the nationality and location of the child and offender. In addition to these referrals, national LEAs are alerted to CSEM by regular complaints requiring investigation as well as referrals from Internet Services Providers (ISPs) and hotline reports. All these reports serve at least two key purposes:

- *First*, they provide law enforcement worldwide with the evidence for investigating individual cases, identifying and rescuing victims, and prosecuting offenders, and
- *second*, they contribute significantly to preventing the re-victimisation by the continued online circulation of CSEM which has severe negative health and social consequences for the victims.[15]

The data of these reports have been instrumental for years in rescuing children in the EU from ongoing

[6] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2018, 18 September 2018, p. 33, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2018.

[7] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 36, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

[8] European Commission, “EU strategy for a more effective fight against child sexual abuse”, Communication, COM(2020) 607 final, 24 July 2020, p. 1.

[9] Internet Watch Foundation (IWF), Annual Reports of 2016 to 2019, available at: https://www.iwf.org.uk/what-we-do/who-we-are/annual-reports.

[10] European Commission, “EU strategy for a more effective fight against child sexual abuse”, Communication, COM(2020) 607 final, 24 July 2020, p. 2.

[11] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 10, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

[12] www.missingkids.org/footer/media/keyfacts.

[13] https://www.rcmp-grc.gc.ca/en/online-child-sexual-exploitation.

[14] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 41, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

[15] Kardefelt-Winther/Day/Berman/Witting/Bose on behalf of UNICEF’s cross-divisional task force on child online protection, “Encryption, Privacy and Children’s Right to Protection from Harm”, Innocenti Working Paper 2020-14, UNICEF Office of Research, 2020 Florence, p. 8; more than 2/3rds of surviving victims worry constantly about being recognized by someone who has seen images of their abuse, see: Canadian Centre for Child Protection, Survivor’s Survey 2017, Executive Summary, p. 7, available at: www.protectchildren.ca/en/resources-research/survivors-survey-results/.

abuse[16] and are vital not only for understanding the extent of the problem of CSEM online but also for how it is accessed and shared. Currently, P2P network sharing still is among the most popular ways of sharing CSEM. However, targeted distribution and sharing increasingly takes place on social networking platforms as well as via widely used encrypted communication applications such as WhatsApp, Telegram and Signal.[17] The use of end-to-end encryption by popular messaging services enables even less tech-savvy offenders to remain untraceable and is a strong argument in the debate on how to proportionately balance the protection of children from CSE online against the protection of all messaging service users' rights to privacy and data protection.[18]

In an effort to overcome the dilemma of having to choose between either protecting children against CSE or protecting the privacy and data protection of all users of a messaging service, the European Commission initiated in 2020 the creation of a technical expert process to explore solutions, which could ultimately allow companies to detect and report CSEM in end-to-end encrypted electronic communication.[19]

At the same time, the European Commission envisions the creation of a European centre maintaining a single database in the EU of known CSEM for facilitating the detection of CSEM in companies' systems, on the one hand, and to support law enforcement by coordinating and facilitating the takedown of CSEM online identified through hotlines.[20] More relevant for GRACE, this European centre is also intended to support Member States by (i) receiving reports in relation to CSE in the EU from companies offering their services in the EU, (ii) ensuring the relevance of such reports, and (iii) forwarding these reports to law enforcement for action.[21]

As a result, the number of CSEM reports can be expected to increase even more and to remain unaffected by the use of end-to-end encryption. The end of COVID-19 pandemic restrictions is expected to result in an extra increase in the number of CSEM reports.[22] A typical CSEM report contains several million images and hours of video footage amounting to up to three terabytes of data. Currently, national LEAs' response to reports on

[16] European Commission, "EU strategy for a more effective fight against child sexual abuse", Communication, COM(2020) 607 final, 24 July 2020, p. 14 et seq.

[17] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 37, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020; Kardefelt-Winther/Day/Berman/Witting/Bose on behalf of UNICEF's cross-divisional task force on child online protection, "Encryption, Privacy and Children's Right to Protection from Harm", Innocenti Working Paper 2020-14, UNICEF Office of Research, 2020 Florence, p. 9 et seq.

[18] Farid, "Opinion: Facebook's Plan for End-to-End Encryption Sacrifices a Lot of Security for Just a Little Bit of Privacy", Berkeley School of Information, News of 16 June 2019, available at: https://www.ischool.berkeley.edu/news/2019/opinion-facebooks-plan-end-end-encryption-sacrifices-lot-security-just-little-bit-privacy; Kardefelt-Winther/Day/Berman/Witting/Bose on behalf of UNICEF's cross-divisional task force on child online protection, "Encryption, Privacy and Children's Right to Protection from Harm", Innocenti Working Paper 2020-14, UNICEF Office of Research, 2020 Florence, p. 10 et seq.

[19] This technical expert process is a specific initiative under the EU Internet Forum: European Commission, "EU strategy for a more effective fight against child sexual abuse", Communication, COM(2020) 607 final, 24 July 2020, p. 16.

[20] European Commission, "EU strategy for a more effective fight against child sexual abuse", Communication, COM(2020) 607 final, 24 July 2020, p. 13.

[21] European Commission, "EU strategy for a more effective fight against child sexual abuse", Communication, COM(2020) 607 final, 24 July 2020, p. 13.

[22] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 41, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

CSEM varies widely as a consequence of capacity and resource constraints – not only at global level[23], but also within the EU[24]. Driven by the growth of self-produced material, the increase in online CSEM has reached a level at which the sheer volume of reports forces national LEAs in the EU Member States to choose between investigating one report instead of another.[25]

2.2. Vision for GRACE Platform

The core aim of the GRACE project is to deliver a European-wide platform providing significant operational value to LEAs across Europe in tackling the volume of online CSEM reports. At the moment, LEAs in some EU Member States receive referrals by the NCMEC and NCECC directly (e.g., Austria, France, Germany, Ireland, Italy, Lithuania, Netherlands, Portugal and Spain) whereas LEAs in other EU Member States receive these referrals by using Europol as the catalyst (e.g., Belgium, Cyprus, Poland and Romania). The vision of the GRACE project is to develop advanced high-level digital and analytical tools made available to LEAs via a Federated Platform which transforms their investigative capabilities into a synchronised and impactful response to the immense influx of reports.

With a new dynamic approach based on Big Data technologies supported by advanced Artificial Intelligence (AI), the solutions developed by the GRACE project will allow LEAs in the entire EU to close the technological gap with offenders. For tackling the influx of CSEM reports, the GRACE project develops Big Data solutions for data ETL[26] which will not only standardise the management of CSEM reports, but also avoid duplicate processing and enhance collaboration amongst national LEAs within the EU. The data of each report will be analysed in terms of visual, audio and text information using AI technologies to produce structured and validated information from the report's content. For this purpose, GRACE will provide novel forensic analysis tools for (i) CSEM-specific content analysis and classification, (ii) content-based geo-localisation, (iii) the creation of evidence graphs to connect cases, (iv) case prioritisation techniques and (v) predictive analysis of trends in CSE offenders' tactics. For the operational coordination of LEAs in all Member States, a Federated (Machine) Learning platform will be developed and established which will exploit available infrastructure as well as any CSEM content distributed across the entire EU.

With the help of these analytical tools, LEAs within the EU can gain the much-needed capacity to address the backlog in reports of CSEM referred to them. A semi-automated mechanism is envisioned to analyse and prioritise the content of the CSEM reports as well as to provide actionable intelligence for the protection of victims and for the apprehension of offenders. The Federated GRACE (Machine) Learning Platform is intended to create a unified learning infrastructure keeping pace with evolving trends in CSE as well as in the use of CSEM for the benefit of law enforcement across the EU without making the actual CSEM of a report available to LEAs with no jurisdiction. The workflow for CSEM reports is currently envisioned for the EU as follows:

- *External Reports:* CSEM reports from outside the EU enter the GRACE platform on a central server at

[23] Kardefelt-Winther/Day/Berman/Witting/Bose on behalf of UNICEF's cross-divisional task force on child online protection, "Encryption, Privacy and Children's Right to Protection from Harm", Innocenti Working Paper 2020-14, UNICEF Office of Research, 2020 Florence, p. 9.

[24] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 36, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

[25] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2020, 5 October 2020, p. 36, available at: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

[26] ETL = Extract, Transform, Load; referring to the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s), see: https://en.wikipedia.org/wiki/Extract,_transform,_load.

Europol where they are enriched by the GRACE tools with several categorisations and made machine readable. Each enriched CSEM report is then forwarded only to national LEAs the jurisdiction of which has been identified as relevant by the GRACE system, while a copy of the enriched report is retained in a database.

- *Internal Reports:* A national LEA participating in the GRACE platform can also be an entry point for a CSEM report. The workflow for national CSEM reports is similar to the workflow for reports from outside the EU, but it will not involve forwarding a copy of the national report to the central server at Europol. Rather, the national report is enriched locally by the same GRACE tools with the same categorisations and made machine readable after which the enriched national report is forwarded only to other national LEAs the jurisdiction of which has been identified by the GRACE system while only the extracted metadata of the national report is shared with the federated GRACE system.

The GRACE platform and tools (= the GRACE system) are envisioned only for analysing, categorising and managing the data contained in the CSEM reports. From a purely investigative point of view, however, it would appear helpful for LEAs if the GRACE platform had also some tools integrated for searching the surface web as well as the dark web. Once a CSEM report is uploaded onto the GRACE system, such tools could automatically either (i) be restricted to verify the data contained in the CSEM report and to update as well as supplement the CSEM report (see 4.1. below) or (ii) search independently of any existing CSEM report, continuously for new CSE(M) related content creating new reports of its own (see 4.2. below). Because of the investigative necessity to verify and the convenience to update the data contained in a CSEM report at some stage, it appears not unlikely that the GRACE platform may be combined with such search tools at some point in the future. The technological design of the GRACE platform cannot prevent a later combination with suitable search tools and, in that sense, will be open for being combined with such automatic search tools for investigative evidence. For that reason, it seems appropriate to include the ethical concerns related to a combination with a search tool in the analysis presented in this Deliverable D9.1 (see 4. below), even though the development and integration of such search tools in the GRACE platform is not part of the GRACE project.

2.3. Overview of Ethical Issues

The GRACE tools and the federated GRACE platform form together the complex GRACE system based on Big Data technologies and supported by AI and machine learning techniques.

In an effort to identify the key ethical issues that have to be covered in this Deliverable D9.1, a traditional PESTLE or STEP approach[27] was carried out. Based on this traditional method, law enforcement's use of the GRACE system has to serve the general principles of respect for human rights, democracy, justice and the rule of law. To achieve these principles, LEAs must work to guarantee that the design and use of the GRACE tool and platform comply with four requirements: fairness, accountability, transparency and explainability.

For the purposes of focus control, WP9 carried out a red teaming exercise to identify additional ethical concerns that need to be included in the assessment provided in this Deliverable D9.1.

Red teaming or alternative analysis is a specific method used to review plans, strategies, and hypotheses.[28]

[27] A PESTLE approach aims to carve out the Political, Economic, Social, Technological, Legal and Environmental aspects (see https://pestleanalysis.com/what-is-pestle-analysis/), while a STEP approach aims to understand four major external surrounding factors of a project: Social, Technological, Economic and Political (see https://pestleanalysis.com/step-analysis/).

[28] See: *Herman/Frost/ Kurz*, Wargaming for Leaders. 2009; *Sabin*, Simulating War, 2012; Fryer-Biggs, Building better cyber red teams, defensenews.com, 14 June 2012; *Lauder*, Red Dawn: The Emergence of a red teaming

Two teams are formed, a Red Team and a Blue Team.[29] The Red Team assumes the role of the attacker, while the Blue Team focuses on defence.[30] This method has been successfully employed by the military for decades[31] and has also been applied in civil activities for a number of years.[32] It is explicitly not restricted to acting out physical attacks. The methodology can also be used to investigate theoretical issues from different angles and with varying emphases – reaching as far as intangible constructs such as a legislative draft.[33] Red teaming can be particularly useful when developing cybersecurity strategies since the attack situation reflects the real threat situation. However, strategies are mostly developed from the defence angle. A change or expansion of perspective enables a company's own strategies to be examined more critically. Red Teaming is not limited to the military context, but it can even be utilized in the process of drafting legislation. [34]

The red teaming exercise revealed an investigative necessity as an additional source of ethical concern regarding the development of the GRACE tools and the federated GRACE platform and regarding their use in the law enforcement ecosystem: Among the first steps of any investigation is the verification of facts followed by an update of the evidence which typically includes a search for potential fresh evidence regarding the investigated suspect(s) and victim(s). Therefore, the GRACE system is likely to be combined with tools for automated searches in the surface web as well as in the dark web, at some stage. The ethical risks of LEAs integrating or combining such automated search tools are elaborated in section 4.2. below. Such automated search tools could either (i) be restricted to verify the data contained in the CSEM report and to update as well as supplement the CSEM reports data with fresh sources or (ii) search independently of any existing CSEM report continuously for new CSE related content and create new CSEM reports of its own.

Taking all ethical aspects regarding the GRACE system together and complementing them with the three additional requirements suggested in the "Ethics Guidelines for Trustworthy AI", the key ethical concerns to be assessed here can be structured as follows:

(1) human agency and oversight (see 3.1. below),

(2) technical robustness and safety (see 3.2. below),

capability in the Canadian Forces, Canadian Army Journal, Vol. 12.2, 2009; *Longbine*, Red Teaming: Past and Present, 2008; *Wood/Duggan*, Red Teaming of Advanced Information Assurance Concepts, DARPA Information Survivability Conference and Exposition, 2002. DISCEX 00 Proceedings, Vol. 2, S. 112ff.

[29] See *Wood/Duggan*, Red Teaming of Advanced Information Assurance Concepts, DARPA Information Survivability Conference and Exposition, 2002. DISCEX 00 Proceedings, Vol. 2.

[30] See *Meija*, Red Team Versus Blue Team – How to run an effective Simulation, CSO 25.03.2008.

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

[32] See *Lauder*, Red Dawn: The Emergence of a red teaming capability in the Canadian Forces, Canadian Army Journal, Vol. 12.2, 2009.

[33] See *Gercke*, "Red Teaming" Ansätze zur Effektivierung von Gesetzgebungsprozessen? Die Übertragbarkeit einer klassischen, militärischen Methodik auf Gesetzgebungsprozesse im IT-Bereich, CR 2014, page 344 et seq.

[34] See *Gercke*, "Red Teaming" Ansätze zur Effektivierung von Gesetzgebungsprozessen? Die Übertragbarkeit einer klassischen, militärischen Methodik auf Gesetzgebungsprozesse im IT-Bereich, CR 2014, page 344 et seq.

(3) privacy and data governance (see 3.3. below),

(4) transparency (see 3.4. below),

(5) fairness (see 3.5. below),

(6) societal and environmental well-being (see 3.6. below),

(7) accountability (see 3.7. below),

(8) automated search tool for individual investigations (see 4.2. below),

(9) automated search tool for CSE content (see 4.2. below),

2.4. Ethics Frameworks and Guidelines

The ethical perspective on systems using information technology has gained momentum with the rise of the market for artificial intelligence (AI). Focusing on the development of automated decision-making, sets of ethical frameworks and guidelines have mushroomed in recent years. In April 2020, the “AI Ethics Guidelines Global Inventory" run by the German initiative Algorithm Watch counted more than 160 different entries from all over the world.[35] A similar inventory offered by the Chinese initiative “Linking Artificial Intelligence Principles” (LAIP) currently lists over 60 different sets of ethical guidelines and offers suggestions on how these sets could supplement each other based on an analysis with an automated word analysis tool.[36]

All these ethical frameworks and guidelines have been issued by public, private or academic organisations and vary from mere recommendations over voluntary commitments up to binding policies depending on whether an organisation has the means to sanction non-compliance.[37] The study of whether and to what extent these frameworks and guidelines gravitate towards a universal ethical standard, has become part of scientific research.[38]

There is a valuable discussion on whether and to what extent ethical guidelines might be mere “ethics washing” aimed strategically to avoid governmental regulation and to maintain an apparent lack of monitoring.[39] The main thrust of this discussion is aimed at ethical guidelines issued by the private sector.[40]

[35] Algorithm Watch, AI Ethics Global Inventory, https://inventory.algorithmwatch.org/about.

[36] Linking Artificial Intelligence Principles (LAIP), http://www.linking-ai-principles.org and Zeng/Lu/Huanfu, Linking Artificial Intelligence Principles, paper for the AAAI Workshop on AI Safety.

[37] See categorisation by Algorithm Watch, AI Ethics Global Inventory, https://inventory.algorithmwatch.org/about.

[38] Mittelstadt, “Principles Alone Cannot Guarantee Ethical AI”, Nature Machine Intelligence, November 2019, https://ssrn.com/abstract=3391293; Cowls/Floridi, Prologemina to a White Paper on an Ethical Framework for a Good AI Society, SSRN of 19 June 2018, https://dx.doi.org/10.2139/ssrn.3198732, pp. 1-14; Floridi et al., “AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Minds and Machines” (2018), Vol. 26, p. 689 (p. 696).

[39] Benkler, Don’t let industry write the rules for AI, Nature, Vol. 569 (9 May 2019), p. 161.

[40] See e.g. Technical University of Munich, “Neues Forschungsinstitut für Ethik in der KI“ press release of 20 January 2019 available in German at: www.tum.de/nc/aktuelles/pressemitteilungen/details/35188/; Walker, “An external advisory council to help advance responsible development of AI, post on Google Blog of 26. March 2019, www.blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/.

Although the AI industry collects the data and has the expertise to integrate fairness into the design of AI systems, the key argument is that it is not for the AI industry to define the processes for investigating which concerns are real and which protective measures are effective.[41] Rather, independent organisations are better qualified, especially when subsidised by the state and subjected to the scrutiny of scientific peer review.[42] The High-Level Expert Group on Artificial Intelligence (AI H-LEG)[43] is such an independent organisation and was established by the European Commission to explore how to uphold in the face of AI the indivisible and universal values of human dignity, freedom, equality and solidarity as well as the rule of law upon which the EU is founded.[44]

2.5. Ethics Guidelines for Trustworthy AI

In April 2019, the AI H-LEG delivered "Ethics Guidelines for Trustworthy AI"[45] which aim to provide guidance in three layers of abstraction. Based on an approach founded on fundamental rights, the most abstract layer identifies as foundation four ethical principles and their correlated values that must be respected in the development, deployment and use of AI systems:

- 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.[46]

These four fundamental ethical principles and their correlated values reflect the universal four classic principles of medical ethics on which ethical guidelines for information technology seem to have converged[47]: (1) respect for human autonomy, (2) non-maleficence, (3) beneficence and (4) justice.[48] Because it is

[41] Benkler, Don't let industry write the rules for AI, Nature, Vol. 569 (9 May 2019), p. 161.
[42] Benkler, Don't let industry write the rules for AI, Nature, Vol. 569 (9 May 2019), p. 161.
[43] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.
[44] Second paragraph of the Preamble of the Charter of Fundamental Rights of the European Union (EU-Charter), 26 October 2012, 2012/C 326/02.
[45] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[46] 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.
[47] Cowls/Floridi, Prologemina to a White Paper on an Ethical Framework for a Good AI Society, SSRN of 19 June 2018, https://dx.doi.org/10.2139/ssrn.3198732, pp. 1-14; Floridi et al., "AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Minds and Machines" (2018), Vol. 26, p. 689 (p. 696).
[48] The groundbreaking work for the medical field by Beauchamps/Childress, "Principles of Biomedical Ethics" is e.g. referred to by Schöne-Seifert, "Prinzipien und Theorien in der Medizinethik", in Ach/Bayertz/Siep, Grundkurs Ethik, Vol. II, p. 9 (p. 16).

necessary to evaluate whether technology based on autonomous and automating algorithms actually respects human autonomy and serves non-maleficence, beneficence as well as justice, it is important to realise which decisions are made by this technology, what advantages and disadvantages might be involved and who is accountable and/or liable for its use, the principle of explicability had been suggested to supplement the classic principles of medical ethics in the context of technology.

When tensions arise between these fundamental ethical principles, it will be for the developers and end-users within the GRACE consortium in concert with its Ethics Board to approach such ethical dilemmas and trade-offs via reasoned and evidence-based reflection rather than intuition or random discretion.[49]

Based on the fundamental rights captured in the four ethical principles and their correlated values, the AI H-LEG elaborated a less abstract layer listing for their fulfilment seven key requirements that an AI system should meet in order to be trustworthy:

- human agency and oversight,[50]
- technical robustness and safety,[51]
- privacy and data governance,[52]
- transparency,[53]
- diversity, non-discrimination and fairness,[54]
- societal and environmental wellbeing[55] and
- accountability.[56]

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.[57]

For each of the seven key requirements to become operational, the Ethics Guidelines for Trustworthy AI provide at concrete and non-exhaustive assessment list that has to be tailored to the specific use case.[58] In

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

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

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

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

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

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

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

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

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

[58] 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.

July 2020, the AI H-LEG made available a refined and final version of this list entitled Assessment List for Trustworthy AI (ALTAI)[59]. 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.[60]

2.6. Towards A Future Ethical Framework for AI

In literature, there is a line of argument that AI ethics is, by itself, deficient in regulating behaviours and practices for proper development and deployment of AI.[61] Not only a lack of a mechanism for reinforcing its own normative claims is detected, but also a lack of mechanisms for accountability.[62] This is countered with the argument that the objective of ethics itself is neither to impose particular behaviours nor to ensure these are complied with, but rather that ethics itself is primarily a form of continuously refreshed and agile attention to reality as it evolves.[63] However, principles, norms and values seem to be, at least, an *end* of ethics and as such form a body of 'soft' moral rights and expectations beyond what is already fixed by law and regulations. Therefore, ethics seem adequately suited for screening the reality of rapid technological developments, especially in the areas of big data, AI and machine learning, and for evaluating what is state-of-the-art. The European Union appears to be heading towards establishing a legal framework for the ethical compliance of AI:

In February 2020, the European Commission suggested in its White Paper on AI to set up a prior conformity assessment for 'high risk' AI systems to verify that they comply with a range of new requirements derived from the AI H-LEG's Guidelines on Trustworthy AI.[64]

In October 2020, the European Parliament adopted a Resolution recommending that the European Commission elaborates a proposal for a new Regulation as a comprehensive European legal framework of ethical principles for the development, deployment and use of AI, robotics and related technologies including not only guiding principles but also binding requirements on high-risk systems.[65] The definitions of AI, robotics and related technologies suggested in its legislative-initiative report seem to indicate that for the European Parliament the autonomy of technological systems is more relevant than the qualification as actual

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

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

[61] Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines", *Minds and Machines* (2020) 30, p. 99; Article 19, "Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence", April 2019, p. 9, https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

[62] Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI", *Nature Machine Intelligence*, November 2019, pp. 6-8 https://ssrn.com/abstract=3391293; Article 19,"Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence", April 2019, p. 18, https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

[63] Rességuier/Rodrigues, "AI ethics should not remain toothless! A call to bring back the teeth of ethics", *Big Data & Society*, July-December 2020, p. 1 (p. 2).

[64] European Commission, White Paper on Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final, 19 February 2020, p. 23-25.

[65] European Parliament, Report with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)), A9-0186/2020, 8 October 2020.

AI for triggering a mandatory ethical compliance assessment for a European certificate of ethical compliance.[66]

The European Commission can be expected to start the European legislative process with its proposal for a comprehensive regulatory framework for AI in the first half of 2021.

3. Ethical Issues Concerning Federated GRACE Tools & Platform

Law Enforcement Agencies (LEAs) need to stay up-to-date with state-of-the-art technology in order to fulfil their role in modern society. LEAs need to keep pace with criminals utilising new technology in order to position themselves adequately for combatting emerging new forms of crime. Because LEAs have the moral obligation to society to perform their task of serving and protecting as well as reasonably possible, the use of new technologies like Big Data, AI and machine learning by LEAs can improve the efficacy and efficiency of investigations initiated by CSEM reports. At the same time, such use has an element of empowerment for all civilians reporting CSEM either to their online service provider or directly to an LEA. Section 3. focusses on using these technologies for internal processes solely within the law enforcement ecosystem, while the following section 4. considers their potential use in combination with an automated tool for external searches. Taken together, these two sections examine the full range of possible ethical concerns related to the interaction of LEAs with Big Data and techniques in the field of machine learning and AI aiming to provide guidance for the development and use of the GRACE tools and platform by LEAs as support in their fight against CSE.

Rapid technological developments raise new questions regarding their use by law enforcement. Not all of these questions are or even can be covered by the law. Nevertheless, law enforcement is not only expected to comply with existing laws and regulations, but also to adhere to what is morally right or wrong in the EU in general and each Member State in particular. This expectation of responsibility extends to the use of information technology including Big Data, AI and machine learning by law enforcement and mirrors the trustworthiness emphasised by the H-LEG as a prerequisite for people and societies to develop, deploy and use AI systems. In fact, the trustworthiness of information technology, especially of AI systems and the responsibility in using them seem to be two sides of the same coin. For LEAs and the entire law enforcement ecosystem to act responsibly means to accept moral integrity and authenticity as ideals and to deploy reasonable effort towards achieving them.[67] The four fundamental ethical principles and their correlated values identified in H-LEG's "Ethical Guidelines for Trustworthy AI"[68] provide the framework for the moral integrity that law enforcement has to continuously strive towards when using information technology. In this quest for moral integrity, law enforcement is responsible for striking a proper balance between rule and value when considering the use of information technology and especially of AI. Understanding this responsibility straightforwardly as requesting LEAs to act in a sensible and trustworthy manner,[69] the seven key requirements for an AI system to be trustworthy elaborated by the H-LEG suggest themselves as key criteria for evaluating the use of Big Data, AI and machine learning by LEAs:

[66] European Parliament, Report with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)), A9-0186/2020, 8 October 2020, Art. 4 and Art. 5(1) on p. 50.

[67] INTERPOL-UNICRI, "Towards Responsible AI Innovation", Second Report on Artificial Intelligence for Law Enforcement, 2020, p. 33 and Zardiashvili/Bieger/Dechesne/Didgnum, "AI Ethics for Law Enforcement", *Delphi 4/2019*, p. 1 (p. 2 et seq.), both citing Dworkin, "Justice for Hedgehogs", 2011, p. 111.

[68] See section 2.5. above.

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

3.1. Human Agency and Oversight

According to the first key requirement, the use of AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy.[70] 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.[71] For that purpose, the user needs to be provided with a sufficient understanding of the AI system enabling the user not only to interact but also to reasonably self-assess or challenge the AI system.[72] Such *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.[73]

The GRACE tools and platform are envisioned to analyse, categorise and prioritise the content of all CSEM reports referred to law enforcement in the EU as well as to provide actionable intelligence for the protection of victims and for the apprehension of offenders. For tackling the influx of CSEM reports, the GRACE project develops Big Data solutions, which analyse the data of each CSEM report in terms of visual, audio and text information using both traditional as well as AI technologies to produce structured and validated information from the CSEM report's content. This involves developing forensic analysis tools for (i) CSEM-specific content analysis and classification, (ii) content-based geo-localisation, (iii) the creation of evidence graphs to connect cases, (iv) case prioritisation techniques and (v) predictive analysis of trends in CSE offenders' tactics. For the operational coordination of LEAs in all Member States, a Federated (Machine) Learning platform will be developed and established which will exploit available infrastructure as well as any CSEM content distributed across the entire EU.

Delegating the analysis and prioritisation of the content of all CSEM reports to an automated system like the set of GRACE tools and the GRACE platform can be seen as a significant gain in efficiency for law enforcement. Having processed and categorised such large amounts of data automatically to aid investigation seems rather beneficial for LEAs because it enhances their productivity and workflows. 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.[74] The GRACE tools' and platform's functionality of automatically 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. Such loss of *human oversight* could be minimised by arranging for these functionalities to be processes with a Human In The Loop (HITL), but this could create a speed bottleneck and thereby defeat the benefits of introducing this functionality. Therefore, *meaningful human oversight* preferably needs to be sought either by having a Human On The Loop (HOTL) or by having a Human In Command (HIC) or perhaps both. 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

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

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

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

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

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

how to use the AI system. Considering the sheer volume of CSEM reports to be processed automatically by the GRACE tools and platform, HOTL processes seem to suggest themselves considering that most of these CSEM reports would otherwise not be processed at all.

At this very early stage of the GRACE project, the design of the tools and platform is still evolving. The design of the functionalities has to allow for *meaningful human oversight* to ensure that the GRACE tools and platform support LEAs in making faster, better, and more considerate decisions in prioritising their investigations based on CSEM reports in their fight against CSE. The higher the impact of the automated decision, the more important it is to get it right, and to treat it with respect.[75] The impact of automated decisions by the GRACE tools and platform can be expected to present CSEM reports to LEAs with jurisdiction for starting an investigation based on some built-in criteria. Comparing this with the current situation in which LEAs are simply overwhelmed by the numbers of CSEM reports and unable to prioritise their investigations based on the bigger picture, any built-in criteria applied across all CSEM reports would already appear to be an improvement. The nature of CSE as a crime suggests, on the one hand, to bear the vulnerability of the (mostly minor) victims in mind when prioritising investigations and to allow for adjustments to the automatic prioritisation in accordance with national laws and targeting goals. 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.[76] These governance mechanisms have to include a periodical review of the functioning of the system, risk management and assessment of ethical and legal compliance.

3.2. 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.[77] This ethical concern is present in every application of engineering values to the efficacy of a technological system. For an AI system to qualify as trustworthy, the algorithms have to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of the AI system and to adequately cope with erroneous outcomes. The AI system needs to be reliable, secure enough to be resilient against both overt attacks and more subtle attempts to manipulate data or algorithms themselves, and the AI system must ensure a fall-back plan in case of problems.[78] The decisions made by the AI system must be accurate or, at the very least, correctly reflect its level of accuracy, and its outcomes must be reproducible.[79]

The GRACE tools & platform are developed to automate the analysis, categorisation and prioritisation of the content of all CSEM reports in order to make each CSEM report machine readable in uniform and standardised data quality and available to the LEAs which have jurisdiction for the investigations, on the one hand, and to learn about current trends and behavioural shifts regarding CSE as soon as they evolve so that all LEAs

[75] Zardiashvili/Bieger/Dechesne/Didgnum, "AI Ethics for Law Enforcement", *Delphi 4/2019*, p. 19.
[76] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 16, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[77] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 16, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[78] Commission, "Building Trust in Human-Centric Artificial Intelligence", Communication COM(2019) 168 final, 8 April 2019, p. 4 et seq.; AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 16 et seq., https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[79] Commission, "Building Trust in Human-Centric Artificial Intelligence", Communication COM(2019) 168 final, 8 April 2019, p. 5; AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 17, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

connected via the federated GRACE platform may place their priorities accordingly, on the other.

In terms of accuracy, the decisions, categorisations and predictions provided by the GRACE tools and platform have to be correct. Subjecting the CSEM report's data on innocent victims and potential suspects to a faulty system would be unethical. The challenge for the GRACE project is how best to improve their efficacy, while maintaining and enhancing their services to society, at the same time. The guiding principle in this respect is that victims, suspects and society should not be worse off than without the automated system.[80]

Robustness concerns how well the GRACE tools and platforms can deal with novel data and situations in a constantly changing world. It is safe to assume that despite ample preventative measures, changes and errors will occur. The GRACE system must not only be robust to errors and/or inconsistencies in its design, development, deployment and use phase, but also degrade gracefully in extraordinary situations including adversarial interactions with malicious actors. How to mitigate the impact of the evolving technological environment, errors and inconsistencies in the GRACE system, including the impact of erroneous attributions by tools based on machine learning or AI, will be a major challenge throughout the GRACE project.

Considering the envisioned pivotal role in coordinating law enforcement's response to CSEM reports, the GRACE system has to be protected against vulnerabilities allowing them to be exploited by adversaries. Security measures need to prevent unauthorised entities as much as possible from gaining access to the GRACE tools or platform and/or tampering with the data of the CSEM reports. Attacks may target 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.[81] The GRACE project has to elaborate processes and measures both, to prevent and mitigate the damage of successful attacks. The security measures for the GRACE system need to be tailored not only to the GRACE system's complexity, but also to safeguarding its potentially high risk for CSE victims and suspects.

The safety of the GRACE system's operators and affected individuals has to be a priority within the GRACE project. Safety means ensuring that the entire GRACE system does what it is supposed to do with no harm to people or resources. The ultimate aim of the GRACE project is to develop tools and establish a federated platform to help prioritise and coordinate investigations based on enriched CSEM reports. For this purpose, the data of the vulnerable and exploited victims specific to these CSEM reports need to be categorised into metrics. To increase the GRACE system's accuracy some real victim's data may have to be processed. Once the GRACE tools and platform are deployed and in use by LEAs, there is an element of (re)victimisation in the context of law enforcement managing all content data of CSEM reports. These victims might be aware that their abuse has been discovered and viewed by law enforcement and they are unable to determine whether or under what conditions their content data (e.g. imagery) is used to support law enforcement activity. There is a fundamental apprehensiveness about how law enforcement's concern with victim identification and offender apprehension can supersede and invalidate the needs, wishes and interests of CSEM victims.[82] Because the GRACE tools and platform are intended for merely serving to automate and harmonise the necessary evaluation of CSEM reports for police investigations, the GRACE system not only stays within the margins of what seems absolutely necessary for acquiring evidence for prosecuting offenders and for first time identification of victims, but ultimately also reduces the number of times the data of a CSEM report are accessed by a human officer. In addition, although a copy of each enriched CSEM report is retained in a central database, the content data of these CSEM reports are only accessed by the LEAs involved in the actual investigation, while only their abstract categorisations are fed into the federated GRACE learning platform for the identification of trends and patterns. This envisioned procedure would keep the dangers of

[80] Dechesne/Didgnum/Zardiashvili/Bieger, "AI Ethics at the Police", White Paper, March 2019, p. 22 et seq.
[81] AI H-LEG, "Ethics Guidelines for Trustworthy AI", 8 April 2019, p. 16, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
[82] Taylor/Holland/Quayle, "Typology of Paedophile Picture Collections", The Police Journal, Volume 74 (2004, p. 97 (p. 100).

(re)victimisation at bay and, at the same time, reduce the exposure of human officers to actual CSEM.

In this context, processes for verification and validation appear essential for a proper development and evaluation of the GRACE tools and platform as well as the functionalities of the entire system. While verification is the process of checking that development at each step happens in accordance with the specifications and that there are no defects, validation considers the impacts of benefits and potential risks to evaluate whether the developed system actually saves time, helps catch criminals, improves the services of law enforcement to society. [83]

3.3. Privacy and Data Governance

As an integral part of human dignity, privacy is intertwined with the principle of prevention of harm and includes the dimension of data protection.[84] Both, the right to privacy and the right to the protection of personal data are at the core of the Police Directive[85] and the General Data Protection Regulation[86] (GDPR) and must be guaranteed by AI systems throughout their entire lifecycle.[87] The legal concerns emanating from the application of these two EU secondary laws and their surrounding wealth of relevant legal rules and regulations for the GRACE project as well as to the deployment and use of the GRACE tools and platform is elaborated and examined in Deliverable D9.2 Legal Report as the second Deliverable of WP9. Deliverable D9.2 will address the relevant data retention rules as well as the balance between the need for large quantities of training data and the principle of data minimisation.

The ethical dimension of preventing harm from privacy necessitates adequate data governance aiming to ensure the quality and integrity of the data used as well as monitoring the data's relevance, processing and access protocols.[88]

The GRACE tools and platform are envisioned to analyse, categorise and prioritise the content of all CSEM reports referred to law enforcement in the EU. At the time of their referral to law enforcement in the EU, the content data of the CSEM reports is already complete and their quality has to be preserved, at the very least. The GRACE tools are intended to improve the data quality of a CSEM report by enriching them with several categorisations and making them machine readable. Each enriched CSEM report is then forwarded to the LEAs in whose jurisdiction the investigation falls. The entire GRACE system has to ensure that the integrity of

[83] Formal verification by logically proving as well as practical verification by testing including unit testing, integration testing, system testing, stress testing and other forms of testing, see Dechesne/Didgnum/Zardiashvili/Bieger, "AI Ethics at the Police", White Paper, March 2019, p. 24.

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

[85] 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.

[86] 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.

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

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

each original CESM report's content data and its enriched metadata is ensured.

Because the CSEM reports have already been referred to law enforcement before the GRACE tools and platform analyse and prioritise their data, the categorisation and distribution of the enriched CSEM reports do not affect any areas of society in which individuals can reasonably expect to enjoy their privacy. Rather, the future GRACE system is merely intended to coordinate and improve the management of CSEM reports solely within the law enforcement ecosystem. Only if the GRACE system were to include tools for external search tools for the surface web and the dark web, the use of the GRACE system by LEAs would potentially affect where individuals can reasonably expect to be private. How this may influence the balance between privacy, on the one side, and law enforcement's duty to maintain order and guarantee security in society, on the other, is elaborated in section 4. below.

3.4. 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.[89] 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.[90] 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.[91] As a vital component in building trust in AI systems, the ethical dimension of transparency can be distinguished in (i) about what, (ii) to whom and (iii) how much and to what end transparency should be provided.[92]

3.4.1. Transparency About What?

Transparency can and needs to be provided for any aspect of an AI system including the overall goal of using AI in a specific context, the decisions selected for automation, the type of machine learning model and the data used, the features in a dataset and the sensitive individual attributes considered by the AI system.[93]

Focussing transparency on the question how an AI system arrives at a certain outcome requires predominantly technical properties of the AI system itself including the sourcing, the usage of training data as well as the processes of development and implementation.[94] The GRACE project needs to meet these

[89] 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.

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

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

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

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

[94] 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,

transparency requirements and to address the following aspects especially:

- *Involved Parties:* It is important to document all parties involved in the financing, management, development, operation and maintenance of the GRACE system identifying especially who contributed what (algorithm, data, process, etc.).
- *Goal:* It is vital to explain why the AI techniques are the best tools for what they are meant to achieve disclosing their requirements for the GRACE system and their exact scope of operation.
- *Design:* 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.
- *Operational Use:* A vision needs to be presented for how the GRACE system can be operated in practice by the personnel of LEAs.
- *Accuracy:* It is vital to indicate the accuracy of each tool and functionality and how an LEA can use it.
- *Data Hygiene:* While the input data for the future GRACE system are the CSEM reports, it is necessary to point out the data used for training, validation and testing.
- *Implementation:* It is important to provide a description of how the entire GRACE system works as a whole and in its parts.
- *Automated Decisions:* 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 most 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.

3.4.2. Transparency to Whom?

The transparency of an AI system very much depends on the group for which transparency needs to be provided.[95] This creates a need for the GRACE project to produce explanations to all groups potentially in touch with the GRACE tools and platform and/or affected by its use.

The GRACE project will develop Big Data solutions which analyse the content data of CSEM reports in terms of visual, audio and text information using AI technologies to produce structured and validated information about the CSEM report's content. During the development phase, the group of all developers involved will need technical documentation for all systems (including high-level information about their data) to be integrated into the GRACE platform or with which the GRACE platform is intended to interoperate. The envisioned end-users of the GRACE system, in contrast, are individual officers at LEAs who will only need sufficient information about how exactly to operate a GRACE tool and the GRACE platform based on a simple understanding of how the entire GRACE system works. These individual end-users are part of a particular unit and organisation within the law enforcement ecosystem which also need to be provided sufficient insight into the use of the GRACE system according to the level of confidentiality along the chain of authorisation. The law enforcement organisation may routinely be audited by government watchdogs which need to be provided sufficient explanations about the functionalities of the GRACE system for their evaluation of how responsibly

March 2019, p. 11; Samek/Wiegand/Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", 2017, arXiv/1709.08296.

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

the audited law enforcement organisation performs its role. Because the GRACE tools and platform aim to provide actionable intelligence for the protection of CSE victims and for the apprehension of CSE offenders, sufficient understanding without extensive technical detail will ultimately need to be provided for all parties involved in judicial proceedings including judges, prosecutors and defence lawyers.

3.4.3. How Much Transparency?

The danger of bad actors finding ways to manipulate the GRACE system raises the question of how much transparency is sufficient for which group. In this respect, transparency and explicability are a gradual matter catering to the level of understanding needed by the group it is provided for.[96]

For transparency within the law enforcement ecosystem, auditability of the GRACE system should be ensured by providing traceability mechanisms, which document the methods used for its development. The auditability of the GRACE system requires documentation of testing methods especially for explicability, privacy, fairness, performance, safety and security.

Ultimately, transparency concerning the reasons for AI-generated decisions amounts to explicability and primarily serves to maintain *meaningful human oversight* over the decisions an algorithm makes. Such *meaningful human control* is necessary to trace moral accountability for the outcomes of machine learning algorithms back to human beings. However, the ethical value of *meaningful human oversight* may not be confused with the *epistemic value* explicable AI might provide.[97] Therefore, there is a strong argument that only AI-generated decisions capable of causing harm require explicability.[98] It has been argued that there is also a catch-22 element in requiring explicability for any potentially harmful AI-generated decisions. This catch-22 element would lie in the fact that the very explanation enabling a human to check the acceptability of the considerations used for an AI-generated decision, actually required the human to already know which considerations should be used which, in turn, then rendered the use of machine learning AI questionable.[99] In this line of argument, the decisions made and suggested by the GRACE system may not qualify as decisions capable of causing harm because the GRACE tools and platform are envisioned to analyse, enhance and prioritise the content of all CSEM reports only after they have been referred to law enforcement for investigation. 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 for human officers at LEAs to make. However, the catch-22 element already appears less convincing when the human check of the considerations used by an AI-generated decision only needs to identify ethically unacceptable considerations. 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. Therefore, the decisions generated by the GRACE tools and platform need to provide traceability mechanisms for transparency, because law enforcement is envisioned to rely on the GRACE system.

[96] Dechesne/Didgnum/Zardiashvili/Bieger, “AI Ethics at the Police”, White Paper, March 2019, p. 14.

[97] Robbins, “A Misdirected Principle with a Catch: Explicability for AI”, Minds and Machines (2019) 29, 495 (501 et seq.).

[98] Robbins, “A Misdirected Principle with a Catch: Explicability for AI”, Minds and Machines (2019) 29, 495 (508).

[99] Robbins, “A Misdirected Principle with a Catch: Explicability for AI”, Minds and Machines (2019) 29, 495 (510).

3.5. 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[100] and calls for inclusion and diversity throughout an AI system's entire life cycle.[101] Automated decisions may not be taken based on discriminatory or unjust attributes.[102]

Building considerations of fairness and non-discrimination into an automated system poses a problem because scalable automated methods to detect and combat discriminatory decision-making require clear-cut rules or quantifiable thresholds. In contrast, the notion of fairness and non-discrimination has historically been specified contextually according to the details of the case and defined in European jurisprudence by judicial intuition, not statistics.[103] The (judicial) interpretative flexibility is not a 'bug' of the notion of fairness and non-discrimination, but rather intentional and essential. Therefore, the technical perspective is vital in providing statistical evidence as well as developing tools for the detection of bias and measuring fairness, while the concept of "contextual equality" needs to be guaranteed and exercised by the judiciary, legislators and regulators.[104] In order to combine these strengths of both, the technical as well as the ethical (and legal) community, it has been convincingly suggested that AI systems need ideally be designed with an 'early warning system' for automated discrimination which produces consistently the types of statistical evidence necessary for a human (and ultimately judicial) detection of unfairness and discrimination.[105]

The development and the design of the GRACE tools and platform need to incorporate measures that together amount to an effective 'early warning system' for unfairness and discrimination. The GRACE project involves developing forensic analysis tools for (i) CSEM-specific content analysis and classification, (ii) content-based geo-localisation, (iii) the creation of evidence graphs to connect cases, (iv) case prioritisation techniques and (v) predictive analysis of trends in CSE offenders' tactics. These tools are developed and trained with non-CSE specific data and biases could easily slip in through selections in the training data or in the tool's design. Each specific tool may either involve or lead to a trade-off concerning fairness and non-discrimination. Once the GRACE tools and platform are deployed and used by law enforcement, the GRACE system's behaviour and results have to be monitored closely for potential changes due to the input of real CSEM report content data. The notion of fairness and non-discrimination requires that the GRACE system will be rigorously audited continuously.

3.6. Societal and Environmental Well-Being

In line with the principles of fairness and prevention of harm, society at large and the environment should

[100] Art. 20 EU-Charter.

[101] 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.

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

[103] Wachter/Mittelstadt/Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI", 3 March 2020, arXiv/2005.05906, p. 1 (p. 44).

[104] Wachter/Mittelstadt/Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI", 3 March 2020, arXiv/2005.05906, p. 1 (p. 46).

[105] Wachter/Mittelstadt/Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI", 3 March 2020, arXiv/2005.05906, p. 1 (p. 47).

also be considered as stakeholders throughout an AI system's life cycle.[106] 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.[107]

The GRACE tools and platform are intended to exploit as much already available infrastructure as possible for the operational coordination of LEAs across Member States. This environmentally friendly approach needs to be maintained when considering resource usage and energy consumption in their development phase and after deployment.

At the level of an individual end-user, the GRACE tools and platform are envisioned as a support system for the individual officer at an LEA offering suggestions about how to prioritise the overwhelming influx of CSEM reports. This kind of assistance is likely and aimed to improve the job satisfaction of an individual LEA officer as an end-user of the GRACE system because the stressful and time-consuming task of analysing, categorising and prioritising CSEM reports is automated at a speed, scale and level of complexity that defy human understanding.

The positive effect at the level of an individual end-user is likely to be mirrored by the effect of law enforcement actually using the GRACE system because the overwhelming influx of CSEM reports would become more manageable and duplicated investigative efforts would significantly be minimised, if not eliminated. Most importantly, however, society at large would potentially benefit from more victims being rescued from their ordeal and CSE offenders apprehended. This would help reducing the risk of re-victimisation for these victims, at the same time, and could also send a strong message to potential future CSE offenders.

3.7. Accountability

The requirement of accountability is closely linked to the principle of fairness and demands mechanisms to be put in place to ensure responsibility and accountability for an AI system and its outcomes throughout the entire AI system's life cycle.[108] 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.[109] 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.[110] 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 the societal level.[111]

The GRACE project has integrated ethics into its project plan in Deliverable D1.4 as well as in Work Package

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

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

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

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

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

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

WP9 and has also established an external Ethics Board for the ethics review of all tasks and deliverables contributing to the development of the GRACE tools and platform. While this Deliverable D9.1 presents the results of the ethical assessment of LEA use of Big Data and AI in Task T9.1, Task 9.3 will develop overall legal and ethical recommendations as guidance for the use of all functionalities of the GRACE tools and platform, because operating in line with ethical and legal standards is a top priority of the GRACE project. Deliverable D1.4, on the other hand, derives from the assessed ethical foundations and concerns concrete and practical guidelines for the development of these functionalities during the GRACE project.

The necessary oversight of the development process requires each tool and functionality as well as the entire GRACE system to be reviewed (auditability) and every decision concerning the technical design to be explained and justified (explicability) on a technical level. Each GRACE tool's behaviour and that of the GRACE platform also needs to be able to verifiable and reproducible in all situations (reproducibility).

4. Ethical Issues Concerning Automated External Searches

The GRACE platform and tools are envisioned only for analysing and categorising and managing the data contained in the CSEM reports. From a purely investigative point of view, however, among the first steps of an investigation is the verification of facts followed by an update of the evidence which typically includes a search for potential fresh evidence regarding the investigated suspect(s) and victim(s). Therefore, it would appear helpful for LEAs if the GRACE platform could, at some stage, be combined with some tools for searching the surface web as well as the dark web. Once a CSEM report is uploaded onto the GRACE system, such tools could automatically either (i) verify the data contained in the CSEM report and update as well as supplement the CSEM reports data with fresh sources (see section 4.1. below) or (ii) separately search even for new CSE related content creating new CSEM reports of its own (see section 4.2. below). Because of the investigative necessity to verify and the convenience to update the data contained in a CSEM report at some stage, it appears technologically viable and beneficial for LEAs that the GRACE system may be combined with such search tools at some point in the future. The technological design of the GRACE system cannot prevent a later combination with suitable search tools and, in that sense, will be open for being combined with such automated search tools for investigative evidence. For that reason, it seems appropriate to include the ethical concerns related to the combination with an automated search tool, even though the development and integration of such search tools in the GRACE platform is not part of the GRACE project.

4.1. Automated Search Tool for Individual Investigations

This section examines the potential ethical concerns related to an automated search tool restricted to the preparation of an individual investigation based on a CSEM report. Taking the evidence contained in the CSEM report which has been analysed, categorised and prioritised by the GRACE system as a starting point, such an automated search tool could verify the CSEM report's evidence, update it and supplement the evidence with additional information about fresh sources, accounts etc. for the victim(s) and suspect(s). The ethical concerns regarding the use of such an automated search tool for the detection of CSEM or CSE related content seem to nestle around the key question of whether a CSEM report establishes sufficient evidence for law enforcement to arouse a reliable initial suspicion on which to base an automated investigation on.

While a CSEM report referred to law enforcement presents sufficient evidence for human officers at LEAs to start a proper investigation, the 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. However, human officers not only have limited workload capacity and speed, but also are morally obliged to verify and supplement certain parts of the evidence in any case (e.g. matches sources of voiceprint, social media accounts etc. contained in a CSEM report), because of the time elapsed since the original report to an Internet Service Provider (ISP) – especially considering the increasing backlog of investigating CSEM reports.

If the automated search tool was restricted to these foreseeable and easily identifiable parts of the evidence which inevitably require verification, updating and supplementation, then the workload for human officers would be reduced and the efficiency of law enforcement would gain, especially if the automated search tool indicated the degree of matching for each result. This would then shift the focus on the accuracy of the search results in order to fulfil the ethical requirement of technical robustness. If the automated search tool was restricted to the inevitable verifications, updates and supplementation and the search results were accurate, then the efficiency gain for law enforcement would render the use of such an automated search tool unproblematic. The inevitability of these verifications, updates and supplementations may even suggest integrating such an automated search tool into the GRACE system and incorporate its results into the envisioned GRACE tools for the prioritisation of CSEM reports.

4.2. Automated Search Tool for CSE Content

If an automated search tool was not restricted to the preparation of an individual investigation based on a CSEM report, then the maximum range for such a general automated search tool would be to constantly monitor the surface web as well as the dark web for any potential content related to CSE and CSEM. This section 4.2. discusses the relevant ethical concerns regarding law enforcement using an automated search tool for CSE content. The use of such an automated search tool may be intended to automate investigations for leads that currently need to be carried out by humans. For that purpose, such an 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 an LEA.

4.2.1. Algorithms and Machine Learning

In order to allow and fulfil these intended functions, the automated search tool has to employ algorithms. Although an "algorithm" may formally be defined as a purely *mathematical construct*[112], lay usage of the term "algorithm" also includes the implementation of the mathematical construct into technology and an application of the technology configured for a particular task[113]. A fully configured algorithm incorporates the abstract mathematical structure that has been implemented into a system for analysis of tasks. Whereas a strict wording would have to distinguish between constructs, implementations and configurations, for the discussion of ethical issues in this deliverable generically referring to "algorithm" will suffice. Algorithms in this sense will not only be found in the configuration of the automated search tool itself but also in the configuration of the external search engines like Google, Bing, DuckDuckGo, Torch, Ahmia and others which the automated search tool might take advantage of and incorporate into its operations.

Replacing a human operator of an investigation at least to a significant extent by an algorithm has the

[112] Hill, "What an algorithm is", *Philosophy & Technology* [2016] 29 (1), p. 35 (p. 47).

[113] See Turner/Angius, "The Philosophy of Computer Science" in: Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017), available at: https://plato.stanford.edu/archives/spr2017/entries/computer-science/.

advantage that the analysis is augmented already by the scope and scale of data and rules involved. The way in which an algorithm makes sense of streams of data and determines features relevant to a given decision outperforms any human operator and involves a qualitatively different decision-making logic applied to larger inputs. The new scale of analysis and the complexity of decision-making are already ethically challenging and this challenge is increased by the opacity of the work by employed algorithms. Traditionally, algorithms operate on decision-making rules which are defined and programmed individually "by hand" (e.g., Google's PageRank algorithm), but increasingly rely on machine learning capacities which are also referred to as "predictive analytics"[114] and "artificial intelligence"[115] because these algorithms are capable of learning.[116] Also, an automated search tool would ultimately be envisioned to have capacities of machine learning. Machine learning generally means that the algorithm defines the decision-making rules to handle new inputs independently of a human operator.[117] Such learning capacities grant algorithms a degree of autonomy the impact of which ultimately remains uncertain. As a result, tasks performed by machine learning algorithms are not only difficult to predict beforehand but also difficult to explain afterwards and this uncertainty might inhibit the identification and redress of ethical challenges.[118]

Against this background, this section identifies the following major ethical challenges for the use of an automated search tool for CSE content:

(1) the quality of evidence produced by an automated tool (see 4.2.2. below),

(2) fairness and non-discrimination (see 4.2.3. below),

(3) negative effects on societal trust and cohesion (see 4.2.4. below),

(4) monitoring effectiveness (see 4.2.5. below),

(5) the (un)suitability of criteria as selectors (see 4.2.6. below) and

(6) the avoidance of chilling effects (see 4.2.7. below).

4.2.2. Quality of Evidence

The first major ethical challenge posed by an automated search tool which is influence by decision-making algorithms concerns the quality of evidence produced by the algorithm. It seems appropriate to divide this challenge into the following three components:[119]

- the (in)conclusiveness of evidence (see 4.2.2.1 below),
- (in)scrutability of evidence (see 4.2.2.2 below) and
- the risk of potential bias (see 4.2.2.3 below).

[114] See Siegel, "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die", 2016.

[115] See Domingos, "The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake the World", 2015.

[116] Tutt, "An FDA for algorithms", *Administrative Law Review* [2017] 69, p. 83 (pp. 94).

[117] Matthias, "The resonsibility gap: Ascribing responsibility for the action of learning automata" *Ethics and Information Technology* [2004], 6, p. 175 (p. 179).

[118] Mittelstadt/Allo/Taddeo/Wachter/Floridi, "The ethics of algorithms: Mapping the debate", *Big Data & Society* [2016] 3 (2), p. 1 (p. 3).

[119] Mittelstadt/Allo/Taddeo/Wachter/Floridi, "The ethics of algorithms: Mapping the debate", *Big Data & Society* [2016] 3 (2), p. 1 (p. 4) referring to the quality of evidence as "inconclusive", "inscrutable" and "misguided".

4.2.2.1. (In)Conclusiveness of Evidence

Algorithmic decision-making and data mining rely on inductive knowledge and correlations identified within the data examined. The evidence produced by an algorithm does not establish any causality. The necessary search for causal links is complicated by the phenomenon that correlations based on a sufficient volume of data could increasingly be seen as sufficiently credible to direct actions without first establishing causality.[120] Acting upon mere correlations may ethically be legitimate but requires a higher threshold of evidence to justify actions with ethical impact. The risk is that algorithmic categories signal certainty, discourage alternative explorations and create a coherence among disparate objects.[121] This leads to the danger of having individuals described via too simplified models.[122] This risk as well as this danger appear to be manageable by the fact that the search results of the automated search tool will be evaluated by an officer of a LEA especially if such officer has been trained to take both, the risk of false certainty and the danger of (over)simplification into account.

4.2.2.2. (In)Scrutability of Algorithm's Functionality and Rationale

The scrutability of evidence presents an essential ethical concern and addresses the transparency and opacity of an algorithm. The primary components of transparency are accessibility and comprehensibility of information, but information about the functionality of algorithms is often poorly accessible. Proprietary algorithms are kept secret either for the sake of competitive advantage[123] or of national security[124]. The transparency of an algorithm therefore involves tensions between several ethical ideals which have to be brought into an acceptable balance.

The transparency of an algorithm is further complicated by machine learning algorithms which are even more difficult to interpret and comprehend as they move along their learning process.[125] It is argued that the opacity of machine learning algorithms inhibits oversight. According to one scholar algorithms are opaque in the sense that the recipient of an algorithm's output rarely has any concrete sense of how and why a particular classification has been arrived at from inputs.[126] The opacity in machine learning algorithms appears to be a product of the high-dimensionality of data, complex code and changeable decision making

[120] Hildebrandt, "Who needs stories if you can get the data?, *Philosophy & Technology* [2011] 24 (4), p. 371 (pp.378-380).

[121] Ananny, "Toward an ethics of algorithms: convening, observation, probability and timeliness", *Science, Technology, & Human Values* [2015] 41 (1), p. 93 (p. 103).

[122] Barocas, "Data mining and the discourse on discrimination", p. 2 under section 2.3 on "faulty inferences".Available at: https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf.

[123] Glenn/Montieth, "New measures of mental state and behavior based on data collected from sensors, smartphones, and the internet", *Current Psychiatry Reports* [2014] 16 (12), p. 1 (p. 6).

[124] Leese, "The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union", *Security Dialogue* [2014] 45 (5), p. 494 (p. 502).

[125] Burell, "How the machine thinks: understanding opacity in machine learning algorithms" *Big Data & Security* [2016] 3 (1), p. 1 (p. 4); Hildebrandt, "Who needs stories if you can get the data?, *Philosophy & Technology* [2011] 24 (4), p. 371 (pp.378-380); Leese, "The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union", *Security Dialogue* [2014] 45 (5), p. 494 (p. 502); Tutt, "An FDA for algorithms", *Administrative Law Review* [2017] 69, p. 83 (pp. 94).

[126] Burell, "How the machine thinks: understanding opacity in machine learning algorithms" *Big Data & Security* [2016] 3 (1), p. 1 (p. 1).

logic.[127] Therefore, it is further argued that meaningful oversight in algorithmic decision-making appears impossible when the machine has an informational advantage over the human operator.[128]

Even concerning algorithms operating on individually "hand-written" decision-making rules it is argued that such algorithms can still be highly complex and practically inscrutable despite their lack of machine learning.[129] Especially when algorithms are developed by large teams of engineers over time, they cannot be divorced from the conditions under which they are developed and this means that algorithms need to be understood as relational, contingent, contextual in nature, framed within the wider context of their socio-technical assemblage.[130] Nevertheless, algorithmic processing contrasts with traditional human decision-making because the rationale of an algorithm may well be incomprehensible to humans which renders the legitimacy its decisions difficult to challenge.[131]

Against this background, algorithmic decision making hardly appears transparent and opacity seems to prevent meaningful risk assessment. In the context of an elaborate automated tool, it would therefore currently appear rather unethical to have any action triggered by this tool other than further scrutiny of the assembled evidence by a human officer. Especially when the officer is aware of the ethical risks and dangers involved with the use of algorithms, the problem of scrutability seems suitably kept at bay.

4.2.2.3. Risk of Potential Bias

Within the literature reviewed for this Deliverable the automation of human decision-making may not be justified by an alleged lack of bias in algorithms.[132] An algorithm's design and functionality reflects the values of its designer(s) and intended uses, if only to the extent that a particular design is preferred as the best or most efficient option.[133] Because the development of an algorithm involves many choices between several possible options, the values of the algorithm's author(s) are woven into the code which in effect institutionalises those values.[134] Without knowledge of the algorithm's development history, it is most difficult to detect latent bias in an algorithm.[135]

In the context of an automated search tool, it is also relevant that the outputs of algorithms require

[127] Burell, "How the machine thinks: understanding opacity in machine learning algorithms" *Big Data & Security* [2016] 3 (1), p. 1 (p. 6).

[128] Matthias, "The resonsibility gap: Ascribing responsibility for the action of learning automata" *Ethics and Information Technology* [2004], 6, p. 175 (pp. 182).

[129] Kitchin, "Thinking critically about and researching algorithms", *Information, Communication & Society* [2017] 20 (1), p. 14 (pp. 20 et seq.).

[130] Kitchin, "Thinking critically about and researching algorithms", *Information, Communication & Society* [2017] 20 (1), p. 14 (pp. 18).

[131] Mittelstadt/Allo/Taddeo/Wachter/Floridi, "The ethics of algorithms: Mapping the debate", *Big Data & Society* [2016] 3 (2), p. 1 (p. 7).

[132] Kitchin, "Thinking critically about and researching algorithms", *Information, Communication & Society* [2017] 20 (1), p. 14 (pp. 18); Newell/Marabelli, "Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datafication'", *The Journal of Strategic Information Systems* [2015] 24 (1), p. 3 (p, 6).

[133] Kitchin, "Thinking critically about and researching algorithms", *Information, Communication & Society* [2017] 20 (1), p. 14 (pp. 18).

[134] Macnish, "Unblinking eyes: The ethics of automating surveillance", *Ethics and Information Technology* [2012] 14 (2), p. 152 (p. 158).

[135] Hildebrandt, "Who needs stories if you can get the data?, *Philosophy & Technology* [2011] 24 (4), p. 371 (p.377).

interpretation. Concerning behavioural data, the correlations presented by the algorithm might come to reflect the interpreter's unconscious motivations, socio-economic determinations and geographic or demographic influences.[136] Therefore, a LEA officer evaluating the evidence presented by the automated search tool has to be trained and aware that meaning is not self-evident in statistical models and that the explanation of any correlation requires additional justification. Different metrics make visible aspects of individuals and/or groups that are not otherwise perceptible.[137] Consequently, it may not be assumed that the LEA officer's interpretation of the evidence correctly reflects the perception of a targeted individual or group rather than the biases of the interpreter.

4.2.3. Fairness and Non-Discrimination

Whereas bias is a dimension of the decision-making process itself, an algorithm also creates the risk of leading to unfair discrimination based on an algorithm's profiling. The algorithm infers a pattern by means of data mining and thereby constructs a profile[138] that inevitably leads to discrimination if based on biased evidence decision-making process. An individual is comprehended based on connections with others identified by the algorithm, rather than based on actual behaviour.[139]

In the context of an automated search tool, the risk of discrimination may emanate from a selector which is unreasonably based on prejudice about the likely characteristics of CSE offenders. The choice of selectors might be too broad so that they single out a group of people on the basis of a trait that is not correlated with CSE. Alternatively, the selectors might disproportionately identify communications of particular kinds of groups or individuals as suspicious who would then suffer from such indirect discrimination.

There appear to be four overlapping strategies for preventing such discrimination in general:[140]

(1) Controlled distortion of training data;

(2) Integration of anti-discrimination criteria into the classifying algorithm;

(3) Post-processing of classification models and

(4) Modification of predictions and decisions to maintain a fair proportion of effects between protected and unprotected groups.

[136] Hildebrandt, "Who needs stories if you can get the data?, *Philosophy & Technology* [2011] 24 (4), p. 371 (p. 376).

[137] Lupton, "The commodification of patient opinion: The digital patient experience economy in the age of big data", *Sociology of Health & Illness* [2014] 36 (6), p. 856 (p. 859).

[138] So the broad definition by Hildebrandt/Koops, "the challenges of ambient law and legal protection in the profiling era", *The Modern Law Review* [2010] 73 (3), p. 428 (p. 431).

[139] Newell/Marabelli, "Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datafication'", *The Journal of Strategic Information Systems* [2015] 24 (1), p. 3 (p. 5).

[140] Romei/Ruggieri, "A mulitdisciplinary survey on discrimination analysis", *The Knowledge Engineering Review* [2014] 29 (5), pp. 582-638 as cited in Mittelstadt/Allo/Taddeo/Wachter/Floridi, "The ethics of algorithms: Mapping the debate", *Big Data & Society* [2016] 3 (2), p. 1 (p. 8).

4.2.4. Negative Effects on Societal Trust and Cohesion

The ethical acceptability of counter-CSE measures may be costly due to its effects on society at large. Especially surveillance measures more often than not affect individuals without any criminal record at a time when no crime has (yet) been committed. This raises the moral risk of social trust and cohesion being eroded by uses of technology.

First, there is the citizen's trust in the policing authorities which could be weakened by what is perceived as excessive and ethically problematic use of technology.[141] In a democracy, citizens are supposed to be allowed unobserved space in which to conduct their relationships and governments and its agents are exposed to the scrutiny of those who are ruled which is a condition of the justified exercise of democratic power. Covert action by state institutions like LEAs could encourage citizens to doubt a central promise of democracy namely that the will of the people will be carried out by the people's institutions. Covert surveillance measures make it difficult to realize whether the will of the people is being done or not. However, in exceptional cases also covert action can be justified especially if the measure is taken to prevent great and imminent harm to citizens and it is reasonable to believe that there is no public alternative available to LEAs at the time when they have to act. Further, even such covert actions require informing and getting permission from bodies under democratic control before an LEA engages in such covert operation.

Second, there is the citizen's right to be trusted as an expression of the more general presumption of innocence. This right to be trusted as an innocent and norm-abiding citizen could be tainted if a surveillance measure is based on the premise that everybody is untrustworthy implying to some extent a presumption of guilt instead of presuming people innocent in the absence of evidence to the contrary. The right to be trusted is based on the ethical consideration that failure to presume people innocent of norm-breaking behaviour is incompatible with respect for them as moral agents.[142] However, the claim that failure to actively trust equates to active mistrust is a fallacy because it mistakenly assumes that trust and distrust are the only two trust-related attitudes it is possible to adopt. In fact, they appear to exist at opposite ends of a spectrum of attitudes.[143] It would be equally unreasonable to require police to treat all individuals for whom there is no individually incriminating evidence of wrongdoing as if there existed evidence of their innocence with respect to the criminal law. But asserting a right to be trusted implies that that is precisely what morality does require.

4.2.5. Monitoring an Automated Tool's Effectiveness

An automated search tool meant to be used for counter-CSE activities may include large-scale collection of data by LEAs which is performed covertly. In such a scenario, the automated tool appears as highly intrusive technology especially because it could be applied to people against whom there is either no evidence of wrongdoing at all or merely less than compelling evidence. Although the potential harm of an act of CSE is very high, at the time an LEA uses the automated search tool neither the probability of this harm is established to be high nor a very likely source for it might be established. As a consequence, the effectiveness of using the automated tool appears in doubt while the probability of intruding deeply and unjustifiably into the lives of individuals who are not involved in terrorism seems rather high.

Against this background, LEAs using such an automated search tool will have to monitor the tool's effectiveness so that they will have an evidence base from which to draw for future decisions about its use in

[141] English, Terrorism: How to Respond, 2009, p. 141.

[142] Duff, "Who must presume whom to be innocent of what?" (2013) *Netherlands Journal of Legal Philosophy* 42(3), p. 170.

[143] Ullmann-Margalit, "Trust out of distrust", (2002) *Journal of Philosophy*, 99(10), p. 532.

operations.

4.2.6. (Un)Suitability of Criteria as Selectors

Whether enriched by the GRACE tools and platform or not, a CSEM report creates sufficient reasonable suspicion against individuals for law enforcement to justify their investigative efforts to verify and broaden the evidence base for this initial suspicion with the help of an automated search tool (see section 4.1. above). In contrast, a key ethical question regarding an automated search tool for the general detection of CSEM or CSE related content is whether and how such automated search tool could establish without any prior suspicion any reliable evidence for a reasonable suspicion against an individual to be somehow involved with acts of CSE. Such an automated search tool may be envisaged to search, crawl and monitor online spaces and forums for content relevant for CSE activities. In this respect, the automated search tool could be viewed as searching for patterns supposedly characteristic of CSE also in spaces and forums with perfectly innocent online activities.

One possibility of focusing an automated search tool on genuinely pertinent information could be the automated filtering of information on the basis of selectors which would then automatically exclude the vast majority of material from intrusive further inspection by an LEA. However, there are various difficulties intertwined with the use of selectors:

First, there is the possibility of choosing particular keywords as selectors. An ideal selector keyword in this regard would be a word that is known to be used exclusively in the context of CSE activities, perhaps something like a secret codeword. The next best would be words providing reasons for suspicion based strongly on evidence. Other possible keywords appear to be rather terms used by large numbers of people for almost any reason. A keyword may not be discriminatory because it has to be indicative only of suspicious CSE activity. A single keyword appears difficult, if not impossible, to define in this regard, but also a set of several keywords may not contain a discriminatory keyword because the use of certain words is not only ambivalent but also more likely to be used by certain cultural groups and in that respect discriminatory. It follows from all this, that a keyword used as a selector for an automated search tool would have to 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 who are forensically aware and aim to avoid the use of incriminating language.

Second, the names of specific groups might serve as good selectors: It is perhaps possible to find names of specific groups providing a stronger evidence base for CSE purposes. Such specific group names, however, could either gain prominence in the press for whatever reason or the specific group is involved in legitimate political life.

Third, it may appear that the name of an individual known to be a CSE offender could be a good selector. However, searches for names of such individuals are quite likely to produce an unacceptably large number of false positives. Behind each false positive is an innocent individual who coincidentally shares their name with a suspicious individual. An error can hardly be neutral and most likely directs suspicion and scrutiny at members of cultural groups because names are to a large extent culturally inherited.

In short, an automated search tool filtering information on the basis of selectors would involve intrusions that seem more than unlikely to pass the test of proportionality, let alone meet the key ethical requirements of accuracy (technical robustness), non-discrimination and fairness.

4.2.7. Avoidance of Chilling Effects

The use of covert surveillance has to be accounted for in democratic societies. The mere knowledge about the use of covert surveillance tools by LEAs may lead citizens to a certain wariness as to how such tools may be employed. Fearing the inconvenience of being detained on a certain suspicion and then released without any conviction, citizens might be disinclined from engaging in otherwise perfectly legitimate online activities.[144] This frame of mind may further be intimidated by the power gained by the surveillant over the surveilled and may potentially cause an undesirable self-censorship leading to a loss of spontaneity when online. Such "chilling effects" are at odds with democratic values and practice. Drawing on the presumption of innocence, no reason should be given for such chilling effects.

One potential reason involving such effects is stigmatisation as criminally suspicious. On an individual level, being stigmatised as having failed to maintain the moral standards of the community can be humiliating. Such humiliation is not intended when following up on suspicion, but it can often be a side effect of such suspicion even if the interference by an LEA is proportionate and well-founded. If such individuals perceive the suspicion to be an unjust implication of wrongdoing, then this may create knock-on social costs by not only eroding their trust in LEAs but also reducing their willingness to cooperate with LEAs.[145] Stigmatisation may also make those affected feel alienated which could damage their self-confidence. When particular groups of people who share salient traits (e.g., religion or race) are stigmatised as suspicious, this may either intensify already existing prejudices against them or even create new prejudices. [146] All these potential social costs are likely to increase along with the importance of the crime one is suspected of having committed. These harms of stigmatisation and their chilling effects impose a moral duty on LEAs to refrain from stigmatising people as criminally suspicious without any good enough reason. It appears generally accepted that evidence linking a specific individual to a particular future or past crime is sufficient ground for treating the individual as a suspect and inflicting on them the costs of such treatment.

For the use of an automated search tool by LEAs, it would seem a valid evaluation that the stronger the evidence is, the more justified appears the use of highly stigmatising measures of suspicion. Further, as long as an LEA follows procedures that ensure that the surveillance will be stopped as soon as it becomes clear that insufficient evidence exists for continued suspicion, the measure could be defended as proportionate and ethically legitimate. 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.

[144] See Stoicheff, "Under Surveillance: Examining Facebook's Spiral of Silence Effects in the Wake of NSA Internet Monitoring", *Journalism & Mass Communication Quarterly* [2016] 93 (2), pp. 296–311; the Washington Post reported about this study under the Headline "Mass surveillance silences minority opinions", *Washington Post*, March 28, 2016. See also the PEN American Center's study "Chilling Effects: NSA Surveillance Drives U.S. Writers to Self-Censor" published November 12, 2013 presenting Research conducted by The FDR Group (thefdrgroup.com/).

[145] See section „non-reporting of discrimination" in: Fundamental Rights Agency of the EU (FRA), (Report) Respect for and protection of persons belonging to minorities 2008-2010, pp. 38 and 39.

[146] See chapter Race Law and Suspicion, in: Kennedy, Race, Crime and the Law, 1997; Lever, "Why racial profiling is unjustified", (2004) *Philosophy and Public Affairs*, 32(2).

5. Conclusion

5.1. Summary

This Deliverable D9.1 has provided an overview and analysis of all potential ethical concerns related to use of Big Data, Machine Learning and AI in the law enforcement ecosystem with regard to investigations concerning CSEM.

In section 2., the ethical standard for the development and operation of the GRACE tools and platform (the GRACE system) has been identified. The key challenge for law enforcement is that the sheer number and volume of online CSEM overwhelms the resources at national LEAs in EU Member States so that they not only have to choose investigating one CSEM report instead of another, but are also faced by an enormous backlog of CSEM reports (see 2.1. above). The vision for the GRACE project is to develop tools and a platform that automatically analyses, categorises and prioritises the content data of CSEM reports (see 2.2. above). Considering the complexity of potential ethical issues (see. 2.3. above) and the vast number of possible ethical frameworks and guidelines, the independence and scientific qualification as well as scrutiny enjoyed by the H-LEG (see 2.4.) are strong arguments for selecting the "Ethics Guidelines for Trustworthy AI" layered in four universal ethical principles and seven key requirements as "gold standard" for the analysis of ethical concerns regarding the GRACE system (see 2.5 above). This "gold standard" can be expected to be woven into the European Commission's Proposal for a comprehensive regulatory framework for ethics and AI which is currently scheduled for the first half of 2021 (see 2.6 above.)

In section 3., the ethical issues concerning the GRACE tools and the federated GRACE platform are measured against the seven key requirements established by the "Ethics Guidelines for Trustworthy AI". For the purpose of meaningful human agency and oversight, processes with a Human On The Loop (HOTL) and a Human In Command (HIC) suggest themselves to be woven into the functionalities of the GRACE system (see 3.1. above). For the benefit of technical robustness and safety, all functionalities of the GRACE system have to work accurately and have to be protected against exploitation of their vulnerabilities, on the one side, and keep the dangers of a victim's (re)victimisation at bay (see 3.2. above). While the requirements of privacy and data protection are elaborated in the "Legal Report" of Deliverable D9.2, the ethical dimension of adequate data governance requires not only to ensure the quality and integrity of the data used, but also to monitor the data's relevance, processing and access protocols (see 3.3. above). The requirement of transparency can only be fulfilled by tailoring each explanation provided for a particular aspect of the GRACE system to the specific context which determines the level of technical detail and the degree of simplification (see 3.4. above). The requirement of non-discrimination and fairness necessitates the integration of measures, which together serve as an efficient 'early warning system' for unfairness and discrimination in each GRACE tool and in the entire GRACE platform (see 3.5. above). The requirement of societal and environmental well-being demands the GRACE project's environmentally friendly approach of exploiting law enforcement's already existing infrastructure as much as possible, as well as the significant improvement of job satisfaction at the level of the individual end-user and a significant efficiency gain for law enforcement's management of CSEM reports for the benefit of society at large (see 3.6. above). For meeting the requirement of accountability, the responsibility for an action, choice or decision including its satisfactory justification must always be assigned to a moral agent and a legal person (see 3.7. above).

In section 4., the potential ethical concerns related to an automated search tool are examined. While the combination with an automated search tool restricted to the preparation of an individual investigation based on a CSEM report seems ethically acceptable (see 4.1. above), the ethical concerns regarding law enforcement using an automated search tool for the detection of CSE content seem to outweigh its desired benefits (see 4.2. above).

5.2. Evaluation

The ethical concerns analysed in this Deliverable D9.1 render 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 degree of automation provided by the GRACE tools and platform especially demands substantial human oversight and continuous human involvement.

At this very early stage of the GRACE project, the design of the tools and platform is still evolving. Deliverable D8.1 depicts the initial pilot scenario for the GRACE platform based on story telling approach listing the technical and expertise needs. The first pilot execution will provide more detailed inputs to adjust the development based on ethical needs derived from the ethical concerns discussed above. This process is expected to be iterative up to the final prototype developed by the GRACE project.

5.3. Future Work

Based on the assessment provided in this Deliverable D9.1, the Deliverable D1.4. will derive from the ethical foundations and concerns assessed here concrete and practical guidelines for the development of these functionalities during the GRACE project. Task T9.3. will then develop overall legal and ethical recommendations as guidance for the use of all functionalities of the GRACE tools and platform and present them in Deliverable D9.3. Operating in line with ethical and legal standards is a top priority of the GRACE project.

ANNEX I - GLOSSARY AND ACRONYMS

Term	Definition / Description
AI	Artificial Intelligence
CSE	Child Sexual abuse and Exploitation
CSEM	Child Sexual abuse and Exploitation Material
ETL	Extract, Transform, Load
ISP	Internet Service Provider
LEA	Law Enforcement Agency
P2P	Peer-to-Peer

Table 3 - Glossary and Acronyms