ALIGNER

ALIGNER D3.4

Cybersecurity requirements structure for AI solutions

Deliverable No.	D3.4
Work Package	WP3
Dissemination Level	PU
Author(s)	Martin Karresand (FOI), Jenni Reuben (FOI)
Co-Author(s)	-
Contributor(s)	-
Due date	2023-09-30
Actual submission date	2023-10-10
Status	Final
Revision	1.0
Reviewed by (if applicable)	Daniel Lückerath (Fraunhofer)

This document has been prepared in the framework of the European project ALIGNER – Artificial Intelligence Roadmap for Policing and Law Enforcement. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101020574.

Contact:

info@aligner-h2020.eu
www.aligner-h2020.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no. 101020574.

Executive Summary

The superior data correlation and analysis capability of AI have enhanced the investigative capabilities of Law Enforcement Agencies (LEAs) in fulfilling their duties to safeguard the European society and its citizens against crimes, threats and terrorism (Westman et al. 2022). However, as recognized earlier in the ALIGNER project, besides the use cases of AI technologies for empowering LEAs, the threats posed by these technologies for the proper functioning of LEAs are two parts of the same coin. The European Agency for Cybersecurity (ENISA) has published a report on AI Cybersecurity Challenges (Malatras et al. 2020) that warns that the use of AI technologies in one's digital solutions and networks open avenues for new attack methods, and attack vectors that are previously unknown of.

As part of the ALIGNER project this deliverable proposes a model for deriving cybersecurity requirements for AI systems. It also presents the minimum requirement for how the cybersecurity requirements should be structured. The model is based on a combination of the threat landscape and AI system lifecycle presented in the ENISA Cybersecurity Challenges report (Malatras et al. 2020), the NIST AI Requirements Management Framework (NIST 2023a), the NIST Cybersecurity Framework (NIST 2023b), and the requirements formulation principles of FOI (Hansson et al. 2011; Hallberg et al. 2018) and Hull et al. (Hull et al. 2005). The model consists of the following steps:

1. ***Objectives identification****: identify the security properties the system should have.*
2. ***Survey****: map the system, its components and their interactions and the interdependencies with external systems […].*
3. ***Asset identification****: pinpoint the critical assets in terms of security that are in need of protection […].*
4. ***Threat identification****: identify threats to assets that will lead to the assets failing to meet the aforementioned objectives […].*
5. ***Vulnerability identification****: determine – usually based on existing attacks – whether the system is vulnerable with respect to identified threats*.
6. **Requirements deriving and formulation**: Derive and formulate requirements based on the vulnerabilities found.

The requirements structure is extendable, as long as the minimum requirements are fulfilled. Whether an extension is needed depends on for example complex requirements needing to be divided into several subsets, a higher focus on the priority of the requirements, or a complex set of identified assets that have to be divided into subset. The requirements should be structured with (minimum) the following fields:

1. Unique ID
2. Asset
3. Life Cycle Stage
4. Combined Priority
5. Parent ID
6. Requirement Text

This is deliverable D3.4 of the ALIGNER EU project, which complements deliverables D3.1-D3.3 by providing a method for deriving cybersecurity requirements for AI systems used by LEAs. It also provides a template for the structure of the requirements. The result of this deliverable will be used for further work in tasks T3.3, T3.4, and T4.3, where the project will screen AI technologies for their

potential (mis)use. Finally, the results from these screening tasks, using the method and requirements structure template from this deliverable, will be reported in the upcoming D5.6, D5.7, and D5.8 roadmap deliverables.

Table of contents

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
AML	Adversarial machine learning
CIA	Confidentiality-integrity-availability
DFD	Data flow diagrams
ENISA	European Agency for Cybersecurity
IEC	ISO/International Electrotechnical Commission
ISO	International Organization for Standardization
LEA	Law Enforcement Agency
MLBS	Machine learning based system
NIST	National Institute of Standards and Technology
NSCAI	National Security Commission on Artificial Intelligence
NVD	National Vulnerability Database
RMP	Risk Management Process
SRE	Software Requirements Engineering

1. Introduction

With the advancement in computational power and storage capabilities, teaching a machine to learn to perform a task without human intervention have become a reality. In this document we define the term artificial intelligence (AI) as a “software equipped with the capacity to act purposefully; in other words, AI is a software designed to take a specific action to achieve a given goal, by dealing effectively with its environment" (p. 8, Eren et al. 2022). The definition is inspired by Devillé et al. (2021). The superior data correlation and analysis capability of AI have enhanced the investigative capabilities of Law Enforcement Agencies (LEA) in fulfilling their duties to safeguard the European society and the citizens against crimes, threats and terrorism (Westman et al. 2022). In their report Westman et al., highlight that AI technologies including deep learning are used to accurately identify faces and fingerprints that may support a criminal investigation. Some examples of applications of AI technologies in automating analyses within the capabilities of LEA are;

- analysis of footages from public spaces (de Boer et al. 2017),
- anomaly detection in CCTV to detect violence in an area or city (Contardo et al. 2021),
- suspect profiling (Saif et al. 2017),
- traffic control through automatic license plate detection and vehicle identification (Luo et al. 2017),
- patrol scheduling (Chase et al. 2021),
- child pornography detection (Vitorino et al. 2018).

A literature survey (Markarian et al.,2022) presents an overview of AI technologies developed through European research funding, which are relevant to LEAs.

However, as recognized earlier in the ALIGNER project besides the use cases of AI technologies for empowering LEA, the threats aggravated by these technologies for the proper functioning of LEA are two parts of the same coin. The rapid growth of AI technologies in businesses, industries, governments, and health care have prompted the European Commission to respond with the first ever regulatory framework for the use of AI technologies (European Commission 2021). Furthermore, the European Agency for Cybersecurity (ENISA) in its published report (Malatras et al. 2020) warns that the use of AI technologies in one’s digital solutions and networks open avenues for new attack methods, and attack vectors that are previously unknown of. ENISA’s warning is substantiated by several studies including the ALIGNER project in which the threat perspective of AI capabilities is included from the beginning of the project. For example, the realization that the use of AI in LEA’s processes and tools opens up opportunities for new crime methods and techniques is reflected in the scenario 3 developed in the project. The first two scenarios reflect the AI-assisted threats at the social-level, such as social manipulation by means of fake news and deep fakes, money laundering, romance scam and weaponization of avatars, but the cybersecurity requirements work (this deliverable and further planned tasks) focuses on the threats to the AI-based tools and systems used by the LEAs for fulfilling their responsibilities.

Cyber security is an area, which will act as driver towards the trustworthy and reliable deployment of AI technologies to prevent AI-targeted attacks (i.e. subverting existing AI systems to alter their decision and prediction capabilities) and to prevent AI-supported attacks (i.e. attacks that use AI to improve the efficacy of traditional attacks). The cyber security needs of LEAs are unique because they not only concern the protection of their internal AI based systems from being attacked by malicious actors but also they need cyber solutions to defend against cyber-crimes that threaten society.

The AI Act (Madiega et al. 2023) signals a risk-based approach towards the use of AI within various applications and sectors. Therefore the same approach is explored in this deliverable through the analysis of security literature and adaptation of well-known risk management frameworks for information processing. The deliverable will provide a method to structure the cybersecurity requirements for AI use in LEAs. The cybersecurity requirements will incorporate considerations regarding, AI safety, integrity and availability, confidentiality, and privacy. The proposed structure will aid LEAs to organize cyber security requirements pertaining to the implementation of security mitigation controls and measures.

The proposed method to organize cyber security requirements for example addresses questions such as i) who is responsible to implement the requirement, ii) what is the type of the control (technical or policy), iii) which part of the development life cycle or which parts of the system is the requirement targeted at and iv) what is the prioritization of the requirement.

1.1 Gender Statement

ALIGNER partners actively safeguard gender equality and are aware of gender issues in science and technology (European Commission 2005).

ALIGNER monitors gender equality addressing biases and constraints throughout all the stages of the project as listed in Gendered Innovations 2 (Schiebinger et al. 2020).

Furthermore, the source literature and other research materials used to perform the work reported in the deliverable D3.4 is reviewed for gender bias during the internal review process following the ALIGNER gender policy (ALIGNER D1.2, Lückerath 2021)

1.2 Relation to other deliverables

This deliverable (D3.4) is related to the following deliverables and tasks;

- Deliverable D3.1 – Impact assessment of AI technologies for EU LEAS
- Deliverable D3.2 – Risk Assessment of AI technologies for EU LEAS
- Deliverable D3.3 – Taxonomy of AI supported crime
- Task T3.3 – AI technology risk assessment
- Task T3.4 – Taxonomy of AI-supported crime
- Task T4.3 – Continuous ethics and law evaluation of AI solutions

Deliverables D3.1-D3.3 are complemented by this deliverable, since it provides a method for deriving cybersecurity requirements, as well as a template for the structure of the requirements. Furthermore, the results of this deliverable will be used for further work in tasks T3.3, T3.4, and T4.3, where the project will screen AI technologies for their potential (mis)use. The results from these screening tasks using the method and template provided by this deliverable will be reported in upcoming roadmap deliverables (D5.6, D5.7, D5.8).

1.3 Structure of this report

Section 2 of the report gives the background to the work and presents the literature used in the rest of the report. Section 3 presents the proposed model that can be used to derive cybersecurity

requirements for AI systems. It also introduces the structure of the requirements introduced at the end of the section. Section 4 contains the conclusions drawn from the presented work.

2. Background

Since attackers may exploit information about AI models for attacks that disrupts the model performance, it is important to keep AI models used by LEA confidential. However, *security by obscurity* has never been an acceptable solution within the cyber security field, stated as "[s]ystem security should not depend on the secrecy of the implementation or its components" by NIST (p. 2-4, Scarfone et al. 2008). Consequently the AI systems of LEA have to be properly protected against current and future attackers. The AI concept also gives rise to new attack vectors by design, vectors not present in traditional IT systems. For example, the training data can be manipulated to modify the AI model to be trained. Attacks may even be performed by only knowing the models' responses to input data (i.e., black-box attacks) (Malatras et al. 2020; Wolff 2020).

2.1 Formulating good requirements

FOI has studied the problem of how to formulate good requirements. The basis is eight principles that each requirement should fulfil (Hansson et al. 2011; Hallberg et al. 2018):

- *form*: each requirement is written according to the same format, they all have the same structure.
- *atomic*: each requirement should be atomic, i.e. contain one and only one requirement. Typically, the requirement does neither contain the words *and* or *or*, nor any listings.
- *specified*: each requirement should contain no more or less than the information necessary for requirement to be interpreted correctly.
- *unambiguous*: each requirement should use unambiguous wording. Values should be preferred before words like *big*, *fast*, *large* and *effective*.
- *verifiable*: each requirement should be verifiable, i. e. it should be possible to verify that it has been fulfilled.
- *consistent terminology*: each requirement is formulated using a consistent terminology. Preferably, the terminology should be properly defined in the documentation of the requirements.
- *abstract*: each requirement should describe what the system should handle, i. e. it should be formulated to be independent of any solution to the problem.
- *traceable*: each requirement should be traceable, to its source, as well as to the resulting design and its implementation. Consequently, each requirement needs to be unique.

Hull et al. (2005) have also present a list of criteria for writing well-formulated requirements. As in the case of the FOI principles every requirement must fulfil the criteria to be useable. The criteria list is generally applicable, regardless of the area of application. Hull et al. give the following eight points, directly cited from the book (p. 85, Hull et al. 2005):

- *atomic*: each statement carries a single traceable element;
- *unique*: each statement can be uniquely identified;
- *feasible*: technically possible within cost and schedule;
- *legal*: legally possible;
- *clear*: each statement is clearly understandable;
- *precise*: each statement is precise and concise;
- *verifiable*: each statement is verifiable, and it is known how;

- *abstract*: does not impose a solution of design specific to the layer below.

The principles/criteria a well-written requirement has to fulfil are overlapping in the FOI and Hull et al. lists. The criteria *atomic*, *verifiable* and *abstract* exist in both lists. The criteria *unique* can be mapped to *traceable*, *precise* maps to *specified*, and *clear* maps to *unambiguously*. The criteria *feasible* and *legal* from Hull et al., together with *form* and *consistent terminology* from FOI, are unique to each list.

The principles/criteria for proper requirements formulation can be applied also on the IT security requirements of AI systems. Since the quality of the original requirements governs the outcome of the proposed requirements structure this part is key to our report.

2.2 Threats to AI systems

The National Security Commission on Artificial Intelligence (NSCAI) has published a final report on AI threats affecting the security of the USA (NSCAI 2021). Although the report is focused on the US situation, the AI related threats to society they list are globally applicable. NSCAI writes, concerning threats from AI use by LEAs, that (p. 144, NSCAI 2021):

"*AI can help automate aspects of data collection and analysis. Such methods can augment the ability of analysts or investigators to sift through and triage masses of information to establish patterns or pinpoint threats. But they also raise questions about the proper roles of machine and human analysis in these processes, including for making predictive judgments. To the extent that an AI system's functions are opaque, it may be difficult to trace and justify the computational process that led the system to make a recommendation. Determining when and how to rely on algorithms is especially pertinent to minimization and querying procedures in the IC and to building cases for law enforcement action*."

NSCAI furthermore highlight the threat from the interactive aspects and dynamic character of AI that will affect the legal certainty if not handled correctly. The problem is described as (p. 144, NSCAI 2021):

"*AI models can evolve based on changing data and interaction with other models, leading to unexpected outcomes. As a result, AI systems require more continuous testing and evaluation than prior generations of technology*."

In addition, the unintended bias that might be introduced into AI algorithms will affect the legal certainty and forensic rigor of LEAs' work. NSCAI writes (p. 144, NSCAI 2021):

"*Unintended bias can be introduced during many stages of the machine learning (ML) process, which can lead to disparate impacts in American society, a problem that has been documented in law enforcement contexts*."

A detailed description of cybersecurity related threats to AI systems can be found in ENISA's "AI Cybersecurity Challenges" (Malatras et al. 2020). Further details on their suggested threat taxonomy and list of threats based on life cycle stage are given in Section 2.6.

Papernot et al. (2016) show a taxonomy of threat models based on their complexity of the adversarial goals and the adversarial capabilities needed. The taxonomy is further used by (Yampolskiy 2019) to explain a threat model of AI based on IT security requirements.

2.3 AI cybersecurity standards

ENISA has published a study on the standardisation efforts connected to the cybersecurity of AI (Bezombes et al. 2023). They use a broad approach involving both the classical confidentiality-integrity-availability (CIA) concept of cybersecurity (Gollmann 2011), as well as the trustworthiness of AI. Since several trustworthiness requirements are affected by the CIA concept any standardisation efforts must handle cybersecurity coherently between the standardisation initiatives. However, the scope of Bezombes et al.'s work is limited to factors affecting the robustness of AI models and algorithms, as well as vulnerabilities related to them.

Bezombes et al. (2023) list different standardisation organisations' works on the cybersecurity of AI. The authors list ISO/IEC 27001 *Information security management* and ISO/IEC 27002 *Information security controls* as two general-purpose standards that are applicable to the mitigation of threats to AI systems. They also write that the ISO/IEC 9001 *Quality management system* standard is relevant to use. These three standards are applicable to all the classical cybersecurity objectives (confidentiality, integrity, and availability), except for ISO/IEC 9001 that lacks support for confidentiality, according to the authors.

Bezombes et al. (2023) identify gaps concerning to which extent the general-purpose standards should be adapted to the AI context. Identification of the gaps is important to the creation of the cybersecurity requirements structure, because the gaps point out areas where the requirements structure cannot depend on the existing standards. The gaps identified by Bezombes et al. are (p. 17, Bezombes et al. 2023):

- *Shared definition of AI terminology and associated trustworthiness concepts*
- *Guidance on how standards related to the cybersecurity of software should be applied to AI*

Bezombes et al. also find gaps regarding the need for the general-purpose standards to be complemented to also cover AI specific cybersecurity needs. They list the following gaps forcing extensions of the standards (pp. 17-19, Bezombes et al. 2023):

- *The notion of AI can include both technical and organisational elements not limited to software, such as hardware or infrastructure, which also need specific guidance*
- *The application of best practices for quality assurance in software might be hindered by the opacity of some AI models*
- *Compliance with ISO 9001 and ISO/IEC 27001 is at organisation level, not at system level. Determining appropriate security measures relies on a system-specific analysis*
- *The support that standards can provide to secure AI is limited by the maturity of technological development, which should therefore be encouraged and monitored*
- *The traceability and lineage of both data and AI components are not fully addressed*
- *The inherent features of ML are not fully reflected in existing standards*

Bezombes et al. argue that the gaps where the standards need to be complemented might be closed in the short run by ad hoc guidance/updates of the standards. However, this approach is not feasible in the end, because ad hoc solutions in combination with a standard is never a good idea. Furthermore, the ad hoc nature of the process does not guarantee a full coverage of the area, according to Bezombes et al. In addition, the fact that the AI area is still not fully mature, as new technologies are emerging, makes it currently hard to standardise (Bezombes et al. 2023).

2.4 NIST Cybersecurity Framework 2.0

The National Institute of Standards and Technology (NIST) is currently updating its Cybersecurity Framework (CF) to version 2.0 (NIST 2023b). It is meant to be used to help lower the cybersecurity risks of IT systems in a diverse set of sectors. The framework can also be used in combination with many standards and risk management processes. NIST writes (p. 8, NIST 2023b):

"*The Cybersecurity Framework provides a flexible and risk-based implementation that can be used with a broad array of cybersecurity risk management processes, such as International Organization for Standardization (ISO) 31000:2018; ISO/International Electrotechnical Commission (IEC) 27005:2022; SP 800-37, Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy; and the Electricity Subsector Cybersecurity Risk Management Process (RMP) guideline.*"

The CF is centred on the Framework Core, which consists of six functions. The functions can be further broken down into categories and subcategories representing cybersecurity outcomes. NIST describes the CF using the figure shown in Figure 1 (p. 5, NIST 2023b):

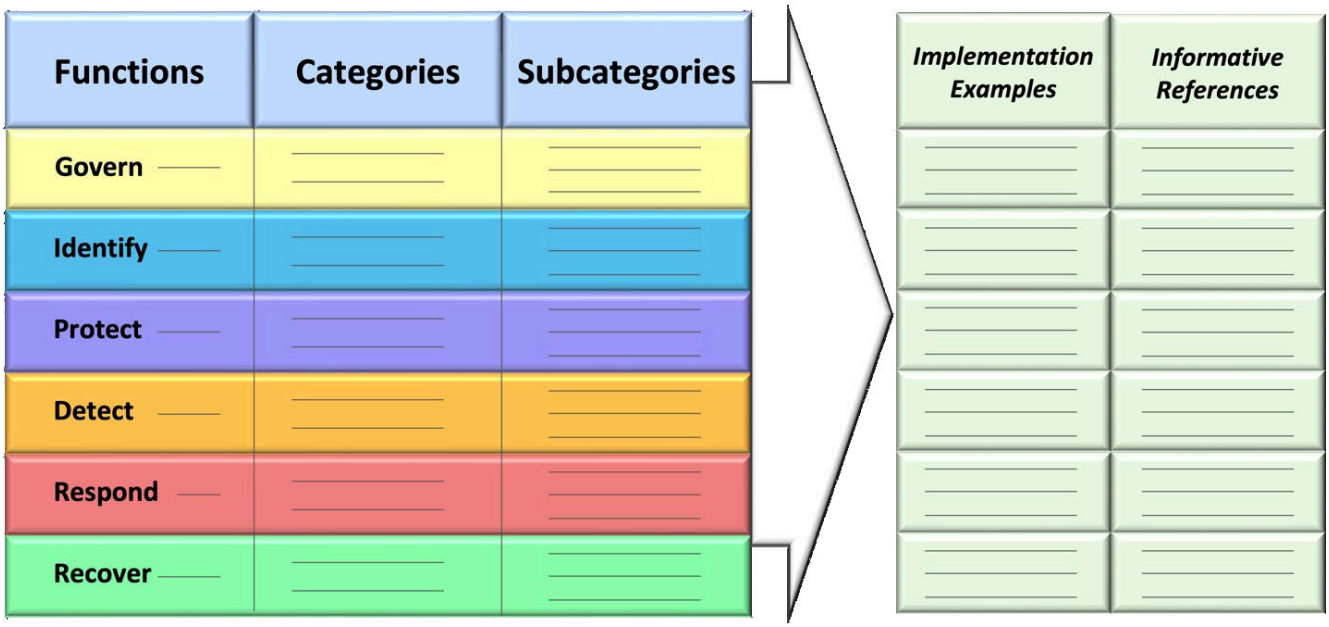

Figure 1 The core of the NIST Cybersecurity Framework (p. 5, NIST 2023b).

The six functions of the NIST CF are connected (see Figure 2) and meant to be executed in sequence due to interdependencies. The main function, Govern, is always active and used to "[*e]stablish and monitor the organization's cybersecurity risk management strategy, expectations, and policy*" (p. 5, NIST 2023b). The Govern function incorporates cybersecurity into the risk management framework of an organization (NIST 2023b).

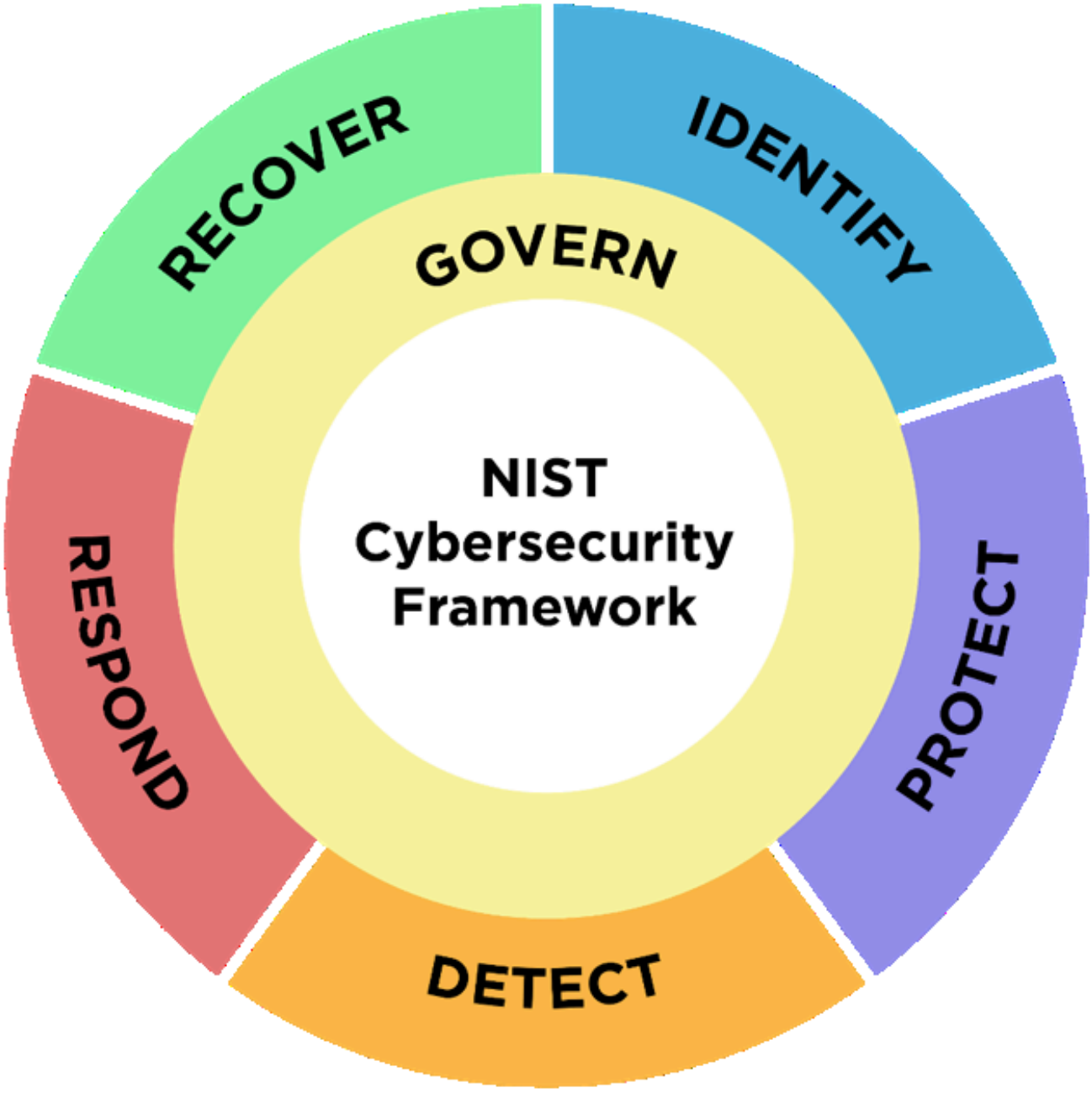

Figure 2 The NIST CF functions and their relations according to NIST (p. 6, NIST 2023b).

The Identify function "*[h]elp determine the current cybersecurity risk to the organization*" (p. 5, NIST 2023b). It is used to identify both cybersecurity risks and improvements needed to be made to mitigate the risks. The function also focuses the cybersecurity risk management work of the organization (NIST 2023b).

The Protect function "*[u]se safeguards to prevent or reduce cybersecurity risk*" (p. 6, NIST 2023b). It uses the results of the Identify function on improvements to lower the risk and impact of any cybersecurity events that might happen. The outcome of the function include both softer actions like training, as well as well as harder actions like the addition of access control hardware, backup systems, and redundant hardware to increase the resilience of the system (NIST 2023b).

The Detect function is meant to "*[f]ind and analyze possible cybersecurity attacks and compromises*" (p. 6, NIST 2023b). This means that in incorporates different types of intrusion detection and anomaly detection processes that are put in place to enable detection of ongoing attacks and incidents (NIST 2023b).

The Respond function shall "[t]ake action regarding a detected cybersecurity incident" (p. 6, NIST 2023b). The results of this function cover all stages of incident handling, for example analysis, mitigation and reporting. It is meant to enable the ability to contain or soften the effects of a cybersecurity attack or incident (NIST 2023b).

The Recover function of the NIST CF is meant to "[r]estore assets and operations that were impacted by a cybersecurity 224 incident" (p. 6, NIST 2023b). The process should be fast to reduce the effects of a cybersecurity incident and restore the system(s) to normal operation as soon as possible. It also includes allowing appropriate communication during an ongoing attack or incident (NIST 2023b).

2.5 NIST AI Risk Management Framework

NIST (2023a) has developed a risk management framework for AI (AI RMF), which covers most of the life cycle of an AI system. In the report on the framework NIST writes that (p. 39, NIST 2023a):

"Privacy and cybersecurity risk management considerations and approaches are applicable in the design, development, deployment, evaluation, and use of AI systems."

NIST presents a list of risks they have identified as being AI-specific and consequently not addressed by traditional risk assessment frameworks. The list includes the following AI-specific risks (pp. 38--39, NIST 2023a):

- *The data used for building an AI system may not be a true or appropriate representation of the context or intended use of the AI system, and the ground truth may either not exist or not be available. Additionally, harmful bias and other data quality issues can affect AI system trustworthiness, which could lead to negative impacts.*
- *AI system dependency and reliance on data for training tasks, combined with increased volume and complexity typically associated with such data.*
- *Intentional or unintentional changes during training may fundamentally alter AI system performance.*
- *Datasets used to train AI systems may become detached from their original and intended context or may become stale or outdated relative to deployment context.*
- *AI system scale and complexity (many systems contain billions or even trillions of decision points) housed within more traditional software applications.*
- *Use of pre-trained models that can advance research and improve performance can also increase levels of statistical uncertainty and cause issues with bias management, scientific validity, and reproducibility.*
- *Higher degree of difficulty in predicting failure modes for emergent properties of large-scale pre-trained models.*
- *Privacy risk due to enhanced data aggregation capability for AI systems.*
- *AI systems may require more frequent maintenance and triggers for conducting corrective maintenance due to data, model, or concept drift.*

- *Increased opacity and concerns about reproducibility.*
- *Underdeveloped software testing standards and inability to document AI-based practices to the standard expected of traditionally engineered software for all but the simplest of cases.*
- *Difficulty in performing regular AI-based software testing, or determining what to test, since AI systems are not subject to the same controls as traditional code development.*
- *Computational costs for developing AI systems and their impact on the environment and planet.*
- *Inability to predict or detect the side effects of AI-based systems beyond statistical measures*

However, there are several gaps in the abilities of the current frameworks dealing with risks, which above all do not address AI specific risks. NIST writes that the frameworks cannot (p. 39, NIST 2023a):

- *adequately manage the problem of harmful bias in AI systems; • confront the challenging risks related to generative AI;*
- *comprehensively address security concerns related to evasion, model extraction, membership inference, availability, or other machine learning attacks;*
- *account for the complex attack surface of AI systems or other security abuses enabled by AI systems; and*
- *consider risks associated with third-party AI technologies, transfer learning, and off-label use where AI systems may be trained for decision-making outside an organization's security controls or trained in one domain and then "fine-tuned" for another.*

The core of the NIST AI RMF are four functions that are used in different parts of the AI system life cycle. The core functions are broken down into categories and subcategories, which in turn are subdivided into actions and outcomes, according to NIST. However, these need not be given in the form of checklists or ordered set of steps. The idea is to use the framework to enable dialogue and actions that help develop trustworthy AI systems (NIST 2023a). The functions in the NIST framework are (p. 20, NIST 2023a):

- ***Govern****: a culture of risk management is cultivated and present;*
- ***Map****: Context is recognized and risks related to context are identified;*
- ***Measure****: Identified risks are assessed, analysed, or tracked;*
- ***Manage****: Risks are prioritized and acted upon based on a projected impact*.

The structure of the functions of the NIST framework can be seen in Figure 3 (p. 20, NIST 2023a). The

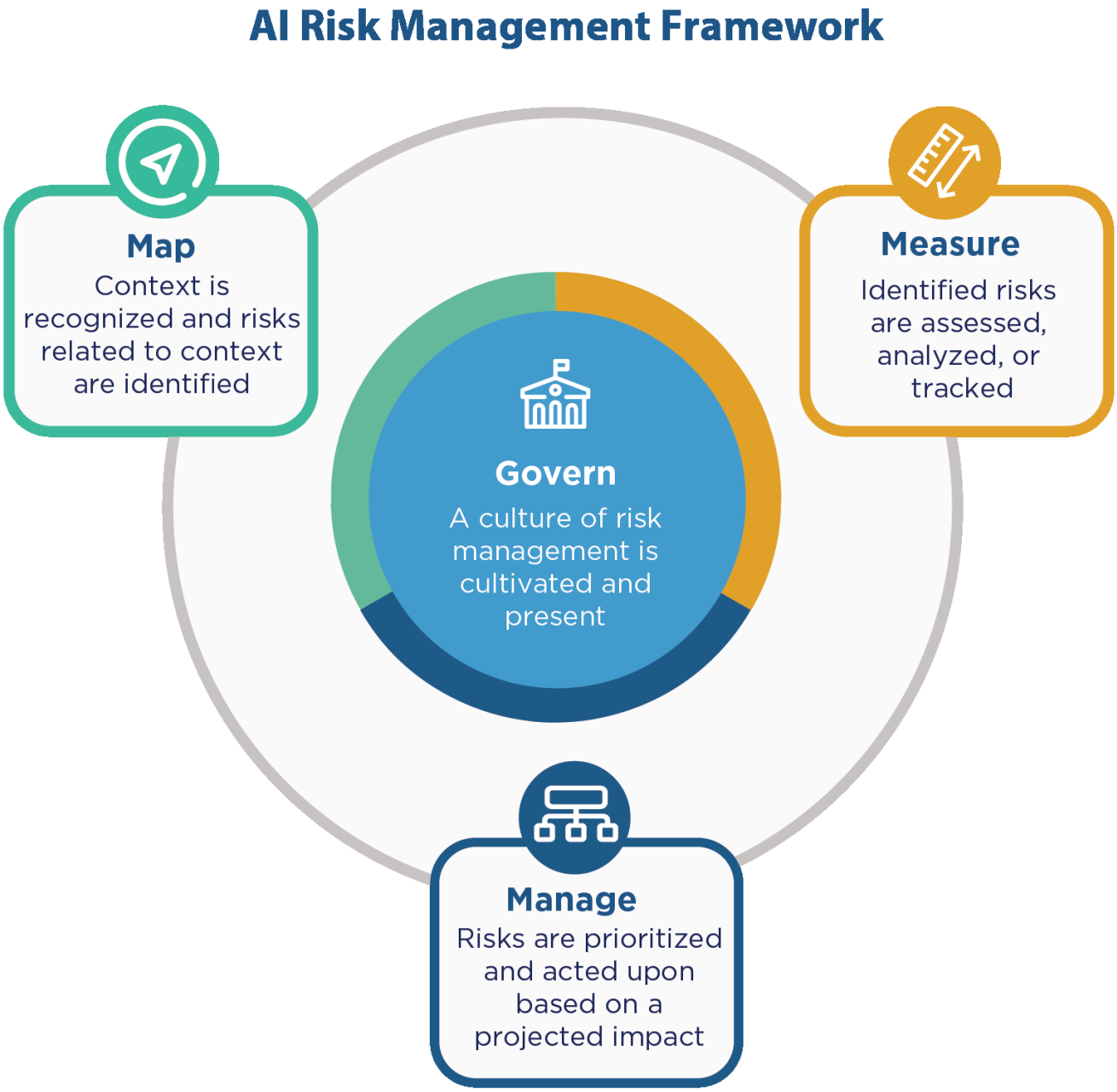

Figure 3 The structure of the functions of the NIST AI RMF (p. 20, NIST 2023a)

function Govern should be active in all stages of the AI system life cycle and is also a prerequisite for the other functions in the framework. It covers the organizational, cultural and management parts of the AI system life cycle. When the structures required by Govern function are in place the project or organization is ready for the next stages in the framework (NIST 2023a).

The Map function of the AI RMF is the first to be used after the proper implementation of the Govern function. The result of the Map stage is used as a basis for the Measure and Manage stages and therefore have a crucial role in the framework. The Map function helps an organization to increase its knowledge on risks and contributing factors affecting the organization's AI systems (NIST 2023a).

The Measure function of the NIST AI RMF is meant to "analyze, assess, benchmark, and monitor AI risk and related impacts" (p. 28, NIST 2023a). The function uses the results from the Map function,

expands them and sends them to the Manage function. Typical examples of processes executed in the Measure function are (rigorous) software testing and performance assessments (NIST 2023a).

The NIST AI RMF Manage function implements the knowledge gained in the previous steps on the risks of the system. The function should be run regularly to update the plans and documentation on the risks and how to mitigate them. The function should also have implemented processes and mechanisms to enable continuous improvement and assessment of new risks (NIST 2023a).

2.6 ENISA AI Cybersecurity Challenges

Malatras et al. (2020) present an ENISA study of cybersecurity challenges to AI. Part of the work involves creating a threat modelling methodology consisting of five steps (p. 25, Malatras et al. 2020):

1. ***Objectives identification:*** *identify the security properties the system should have.*
2. ***Survey:*** *map the system, its components and their interactions and the interdependencies with external systems […].*
3. ***Asset identification:*** *pinpoint the critical assets in terms of security that are in need of protection […].*
4. ***Threat identification:*** *identify threats to assets that will lead to the assets failing to meet the aforementioned objectives […].*
5. ***Vulnerability identification:*** *determine – usually based on existing attacks – whether the system is vulnerable with respect to identified threats.*

In the objectives identification phase the authors include both the traditional cybersecurity properties (CIA) and extend the set by *authenticity*, *authorization* and *non-repudiation*, also from the cybersecurity domain. Finally, the more AI specific properties *robustness, trustworthiness, safety, transparency, explainability, accountability*, as well as *data protection* are added to the set of properties. The non-CIA properties are mapped to CIA by the authors in the following way (pp. 25-26, Malatras et al. 2020):

- *Authenticity may be affected when integrity is compromised, since the genuineness of the data or results might be affected.*
- *Authorization may be impacted when confidentiality and integrity are affected, given that the legitimacy of the operation might be impaired.*
- *Non-repudiation may be impacted when integrity is affected.*
- *Robustness of an AI system/application may be impacted when availability and integrity are affected.*
- *Trustworthiness of an AI system/application may be impacted when integrity, confidentiality and availability are affected, because the AI system/application may be operating under corrupted data or underperforming.*
- *Safety may be affected when integrity or availability are affected, since these properties might adversely impact the proper operation of an AI system/application.*
- *Transparency may be affected when confidentiality, integrity or availability are impacted, since it hinders the disclosure of why and how an AI system/application behaved as it did.*
- *Explainability may be affected when confidentiality, integrity or availability are impacted, since it hinders the inference of proper explanations on why an AI system/application behaved as it did.*

- *Accountability may be affected when integrity is impacted, since it hinders apportioning of verified actions to owners.*
- *Personal data protection may be affected when confidentiality, integrity or availability are affected. For example, breach of confidentiality (e.g. achieved through combination of different data sets for the same individual) can lead to the disclosure of personal data to unauthorised recipients. Breach of integrity (e.g. poor data quality or "biased" input data sets) can lead to automated decision-making systems that wrongly classify individuals and exclude them from certain services or deprive them from their rights. Breach of availability, can disrupt access to one's personal data in important services, based on AI. Transparency and explainability can also directly affect personal data protection, while accountability is also an inherent aspect of personal data protection. In general, AI systems and applications may significantly limit human control over personal data, thus leading to conclusions about individuals, which directly impact their rights and freedoms. This may happen either because machine outcomes deviate from the results expected by individuals, or because they do not fulfil individuals' expectations.*

After the objectives identification phase a survey of all points of interaction between the AI system and the surrounding world. It is important that all interaction points are included, both manual (human interfaces) and automatic (external system interfaces). When the interaction points have been surveyed the assets of the system are surveyed. Malatras et al. have divided the assets into six categories, which all have to be taken into account. The categories are (p. 22, Malatras et al. 2020):

- *Data*
- *Model*
- *Actors*
- *Processes*
- *Environment/Tools*
- *Artefacts*

The ENISA list of assets is further broken down into an asset taxonomy, which can be used as a support during the asset survey phase of the ENISA model. The ENISA asset taxonomy is shown in Figure 4 (p. 23, Malatras et al. 2020). However, due to the currently rapid development of the AI ecosystem, its complexity, and scale the set of assets might change over time. The same is true for the details of the asset taxonomy, where new assets might be added and old deleted. The asset mapping phase is therefore still evolving and consequently not yet mature and stable (Malatras et al. 2020).

Figure 4 The ENISA asset taxonomy (p. 23, Malatras et al. 2020)

The next step in the ENISA model is threat identification. Malatras et al. use AI specific parts of the ENISA threat taxonomy[1]. This gives the following list of high-level threats to AI systems (the citations in the list are made by Malatras et al. and are taken from the ENISA threat taxonomy) (p. 27, Malatras et al. 2020):

[1] The complete ENISA threat taxonomy can be found at https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends/enisa-threat-landscape/threat-taxonomy/view

- *Nefarious activity/abuse (NAA): "intended actions that target ICT systems, infrastructure, and networks by means of malicious acts with the aim to either steal, alter, or destroy a specified target".*
- *Eavesdropping/Interception/ Hijacking (EIH): "actions aiming to listen, interrupt, or seize control of a third party communication without consent".*
- *Physical attacks (PA): "actions which aim to destroy, expose, alter, disable, steal or gain unauthorised access to physical assets such as infrastructure, hardware, or interconnection".*
- *Unintentional Damage (UD): unintentional actions causing "destruction, harm, or injury of property or persons and results in a failure or reduction in usefulness".*
- *Failures or malfunctions (FM): "Partial or full insufficient functioning of an asset (hardware or software)".*
- *Outages (OUT): "unexpected disruptions of service or decrease in quality falling below a required level".*
- *Disaster (DIS): "a sudden accident or a natural catastrophe that causes great damage or loss of life".*
- *Legal (LEG): "legal actions of third parties (contracting or otherwise), in order to prohibit actions or compensate for loss based on applicable law".*

The final stage of the ENISA model is mapping threats to vulnerabilities. Malatras et al. therefore lists 74 identified threats to AI system in Annex B of the ENISA threat model report. The threats in the list are accompanied by their related CIA properties and possible assets to be affected (Malatras et al. 2020).

2.7 Other threat modelling frameworks

Wilhjelm et al. (2020) propose a method to elicit security requirements for machine learning based systems (MLBSs). The method is based on two main concepts, the use of data flow diagrams (DFDs) and STRIDE classification method, which is used to identify adversarial machine learning (AML) threats. STRIDE consists of the threat categories Spoofing, Tampering, Repudiation, Information disclosure, Denial of service and Elevation of privileges. The authors give the following outline of the proposed method (p. 427, Wilhjelm et al. 2020):

- *Identify threats related to MLBSs using DFDs and STRIDE:*
  - *Develop a software model (an architectural model) for MLBSs using a DFD*
  - *Develop an AML threat taxonomy based on existing literature*
  - *Map the AML threat taxonomy to the DFD*
  - *Bridge the conceptual gaps between AML threats and STRIDE*
  - *Use STRIDE to identify AML threats related to MLBSs*
- *Rank AML threat impacts using Microsoft AI/ML Bug Bar's threat ranking approach.*
- *Elicit AML threat mitigations using Microsoft AI/ML attack library.*

Wilhjelm et al. (2020) motivate the choices they have made during the construction of the proposed model with an aim for simplicity. For example, they chose to use DFD during the first modelling stage, because it is easier to understand by developers, than other methods to identify potential threats. The use of DFD is also focused on the software itself, instead of identification of business assets or what motivates an attacker. STRIDE is designed to help software developers analyse the threat landscape of traditional IT systems and consequently has a software focus.

To map the AML threats onto STRIDE the authors categorise the AML threats into three classes based on their respective attack vector in the following way (p. 428, Wilhjelm et al. 2020):

- *Flawed data*
- *Model extraction*
- *Data Extraction*

The mapping of ML threats to STRIDE is shown in Table 1, which is based on Table III in Wilhjelm et al. (2020).

Table 1: ML-threats mapped to STRIDE (p. 429, Wilhjelm et al. 2020). FD: Flawed Data, ME: Model Extraction, DE: Data Extraction, T: Traditional Threats

Element	S	T	R	I	D	E
Data Flows		T		T	T	
Data Stores		FD		T	FD	
Processes	ME	FD/ME	FD/ME	ME/DE	FD/ME	FD/ME
External Entity	FD/ME		FD			

To rank the AML threats found in the previous steps of the method Wilhjelm et al. use the Microsoft AI/ML Bug Bar (Marshall et al. 2019). Other alternatives, such as DREAD and CVSS, are not studied because of the current immaturity of the AML field (Wilhjelm et al. 2020).

The final step in the model proposed by Wilhjelm et al. is to find mitigations for the identified threats. The authors chose to use an attack library for MLBSs created by Microsoft (Shevchenko et al. 2018). The library needs improvement and requires knowledge of ML. Furthermore, it deviates from what traditional security requirements engineers are accustomed to, making it less usable. However, it is still the best alternative, according to Wllhjelm et al.

2.8 Security patterns and severity measures

An integral part of the requirements engineering field is the use of security patterns. These are used to create secure software functions by working as templates for the programmer to follow. Van den Berghe et al. (2022) and Cordeiro et al. (2022) both present attempts to survey and collect smaller sets of security patterns into catalogues of patterns. We have not found any security patterns dedicated to AI systems.

To rate the severity of a threat to a specific asset of a system the Common Vulnerability Scoring System (CVSS)[2] can be used (FIRST 2023). The CVSS scoring system is maintained by the Forum of Incident Response and Security Teams (FIRST) and part of the information held in the National Vulnerability

[2] https://www.first.org/cvss/user-guide

Database (NVD)[3] maintained by the National Institute of Standards and Technology (NIST). The NVD can be used to search for a specific version of a software and in that way find both any vulnerabilities affecting it, as well as the severity (CVSS score) of the vulnerability.

[3] https://nvd.nist.gov/

3. A Proposed Modell for AI Cybersecurity Requirements and Their Structure

As presented in Section 2, various standards bodies such as the International Organization for Standardization (ISO) and National Institute of Standards and Technology (NIST) specifies particular risk assessment frameworks for safeguarding information security management systems[4]. Thus, organizations follow a risk assessment procedure to identify, re-evaluate and maintain cybersecurity risks associated to the organizational assets to maintain their security level. Alternatively, since the cybersecurity risks concerns systems, networks, software programs and software artifacts, organizations may employ Software Requirements Engineering (SRE) methods such as KAOS (Bertrand et al., 1998), Secure Tropos (Susi et al. 2005; Mouratidis et al. 2007), SEPP (Schmidt et al. 2011) and CORAS (Lund et al. 2011) for threat Analysis to assess and ensure cyber security protection. However, it is shown by Beckers et al., (Beckers et al. 2012) that following the software requirements engineering approaches to maintain security posture of one's organizational systems and networks fulfills the obligations required by the standards bodies, because the former and the later overlaps at the structure level.

The report by ENISA on AI cybersecurity challenges presented in Section 2.6, follows their own threat modelling process and knowledge from the experts to map out the security attack landscape of the ecosystem concerning development and deployment of AI-enabled software systems. This work is inspired by ENISA's method, but also the traditional information security risk assessment framework and traditional threat modelling method are used to study the cybersecurity risks to the AI ecosystem setup used by LEAs. The reason for the combination is that ENISA stops at the threat level, together with a lack of requirements formatting (structure) information in the studied frameworks. Thus, this section presents a mapping of cyber security risks to AI technologies ecosystems by combining the traditional risk assessment frameworks from ISO, NIST and threat modelling results from the work by ENISA, extended by the requirements formatting principles of FOI and Hull et al.

3.1 Outline of Proposed Model

The model from the ENISA AI Threat Landscape (Malatras et al. 2020) inspires the proposed model for extraction of cybersecurity requirements presented in this publication. Our model contains the following steps (the ENISA steps are printed in italics):

1. ***Objectives identification:*** *identify the security properties the system should have.*
2. ***Survey:*** *map the system, its components and their interactions and the interdependencies with external systems […].*
3. ***Asset identification:*** *pinpoint the critical assets in terms of security that are in need of protection […].*
4. ***Threat identification:*** *identify threats to assets that will lead to the assets failing to meet the aforementioned objectives […].*
5. ***Vulnerability identification:*** *determine – usually based on existing attacks – whether the system is vulnerable with respect to identified threats.*

[4] When the ALIGNER documentation on D3.2 Risk Assessment for AI Technology is published it will supersede the NIST risk assessment framework in our case.

6. **Requirements deriving and formulation:** Derive and formulate requirements based on the vulnerabilities found. The requirements should be derived as stated in Section 3.3 and follow the principles of FOI and Hull et al. presented in Section 2.1. The requirements should be structured in accordance with what is stated in Section 3.4.

The first two steps in the model are important, but nevertheless out of scope in this report. It incorporates main knowledge that should already be known to the system owner. For help in that work we recommend the NIST AI RMF (NIST 2023a) or possibly NIST CF 2.0 (NIST 2023b) when it is finalized. If the organization using our model already has implemented another cybersecurity framework it can be used for the objects identification and system survey steps if needed.

3.2 Recognize Assets, Threats and Vulnerabilities

The notion of risk can be defined as the following; Given a set of threat scenarios, the corresponding impact to the assets and the likelihood of such scenarios then the risk of the system corresponding to the set of threat scenarios is denoted as a tuple,

$$\text{Risk} \cong \{(\text{Scenario, Impact, Likelihood})\}.$$

3.2.1 Assets

Identification of assets and asset actors as the first step in the section of 4.2.1 in ISO 27001 standard on information security management systems, which aligns with the threat modelling procedure employed by ENISA. Assets are anything that are valuable to an organization because the loss of the asset will constitute a risk to the organization. At FOI, after years of research in security evaluation and information security risk assessment, we have established that clear description of one's assets plays a crucial role in the subsequent risk assessment phase (Hallberg et al. 2018). Although, various standard organizations and independent actors have developed various asset taxonomies for categorizing a software system and its environment, we refer to the asset taxonomy developed by ENISA (Malatras et al. 2020). As described in Section 2, ENISA's asset taxonomy pertains specially to the AI ecosystem. Furthermore, to understand the impact of risking the security protection of an asset, we propose impact categories and impact-level ranging from low to high. The impact categories include short-term versus long-term and discrete versus connected. For example, for the asset category machine learning models the corresponding impact category is connected and impact-level is high.

Table 2. An example of how to identify an organizational assets

Asset Id	ENISA Asset Category	Impact Category	Impact-level
A1	Models	Connected	High

3.2.2 Threats

Besides the data, storage and computational systems categories, the asset taxonomy includes processes and artefacts for example, data governance policies, patch management process, etc. Such artefacts enable an organization to identify threats that are relevant to the identified assets. To guide the LEAs in the threat identification phase, we recommend the LEAs to refer to ENISA's threat taxonomy. ENISA's AI threat landscape (Malatras et al. 2020) maps both the threats to AI systems and threats originating from AI adversarial systems to the assets associated to every phase of an AI system lifecycle.
Using the threat taxonomy we consider the threat scenarios that are relevant to one's organization. Different scenarios could be considered depending on elements such as: what devices are compromised, how knowledgeable about the system the adversary is, what type of adversary is attacking the system, what security mechanisms are in place, among others.

For assessing the risk, we also have to assess the threat likelihood. However, calculating the threat likelihood is tricky in the information risk management. In contrast to information security risks, in safety risk management that deals with risk against natural phenomena, the threat likelihood is calculated as the posterior probability of a given consequence event (e.g. earthquake). However, when it comes to cyber security risk the consequence is not a cause of nature but from an intelligent and malicious adversary, so it is difficult to objectively calculate the probability of a security event, a priori. Therefore, to assess the likelihood of the threat occurring to an organizational ICT system, we propose to look into other proxy measures such as what are the capabilities of the adversary, how severe is the vulnerabilities. For example by using the CVSS score (NIST vulnerability metrics[5]) we can assess the severity of the vulnerabilities identified in ENISA's threat taxonomy simply by mapping the vulnerability to its corresponding score CVSS scores in the National Vulnerability Database (NVD).

Perform the risk assessment using the NIST risk assessment for information security. The outcome will be a matrix for each identified threat scenarios. From the matrix pick up the high impact, high likelihood threat and gather all the threat information such as the threat actor, vulnerabilities, asset actors, etc., for deriving the cyber security requirements.

3.3 Deriving Cyber Security Requirements

In general, the Security Requirements Engineering (SRE) methods such as KAOS (Bertrand et al., 1998), Secure Tropos (Susi et al. 2005; Mouratidis et al. 2007), SEPP (Schmidt et al. 2011) and CORAS (Lund et al. 2011), are used to devise the security properties of the to-be built system by analyzing the system requirements, identification of assets as well as its corresponding actors and threats.

Security is a system property and a desirable property to protect the system from undesirable effects. Following the terminology devised by Jackson (Jackson 2001) a system constitutes of one or more machine and environment. A machine in our case is the machine learning building programs, data processing programs, automatic model validation programs, etc., whereas the environment describes the real world where the machine will be put to use. In Section 3.2 we

[5] National Vulnerability Database, Vulnerability metrics https://nvd.nist.gov/vuln-metrics/cvss

have established the components of a system in terms of assets and in this section, we focus on the desired behavior of the machine with its environment. The desired behavior are the specifications or requirements based on which the system is built within the context of the given environment. The security properties of the system that a system guarantees after it is built are the security requirements that are prioritized during the development of the system. For non-AI enabled organizational systems, the most common security properties are confidentiality (C), integrity (I) and availability (A) along with extended properties such as authenticity, authorization and non-repudiation. However, when it comes to the AI-enabled systems, ENISA specifies additional security properties besides the traditional CIA properties and they are robustness, trustworthiness, safety, transparency, explainability, accountability and data protection.

The impact of threats to the system can be explained in terms of CIA and by understanding the impact of threats on the fundamental property it become easier to map the impacts of threats to additional security properties pertaining specifically to AI-enabled systems. For example,

- Robustness – Threat to the availability and integrity of services and data affects the robustness of an AI application
- Trustworthiness – If the machine learning models within an AI application is based on poisoned data thus undermining the integrity, confidentiality and availability of the system then the trustworthiness of the AI application is compromised.
- Safety – Safety of an AI-enabled system relates to the threats to integrity or availability properties of the system.
- Transparency and Explainability – When confidentiality or integrity or availability of an AI application is compromised then that impacts the proper interpretation of why the system behave in a certain manner.
- Accountability – When integrity of the AI application is impacted then it is challenging to verify the behavior of the system to the stakeholders.
- Data protection – Breach of data confidentiality, integrity affects the data protection principles such as data minimization, data segmentation, etc.

The result of the risk analysis is the information of the threats, system and environment (the domain knowledge and the assumptions). Along with the information from the risk analysis, SRE method like CORAS states to conduct stakeholder interviews to elicit security goals for the assets owned by each stakeholder. Often the security goals refers to aforementioned security properties, but risk analysis result do not contain the mapping of the threats to the security properties. Therefore, we employ the threat modelling approach taken by the work conducted at ENISA, ENISA have mapped the threats to each assets present in an AI system architecture and then map the impact of the threats to security properties.

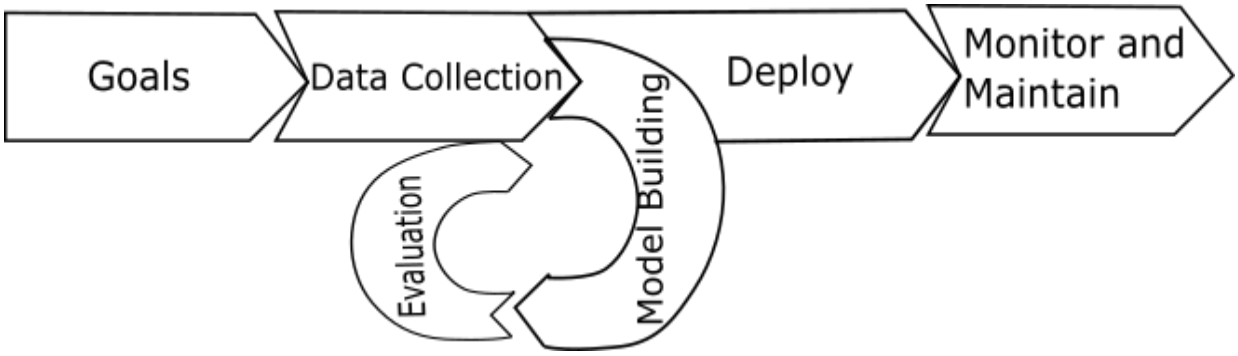

Figure 5. A conceptual model of AI development lifecycle

The security goals (i.e., security properties) formulated as security requirements from each stakeholder are collectively analyzed and prioritized in the first stage of the AI development life cycle in Figure 5. For the prioritization of the security requirements, the most common method used within the security literature (Wilhjelm et al 2020) is the Common Vulnerabilities Scoring System (CVSS). Accordingly, threats and vulnerabilities are teased out from the selected high impact risks identified through the risk analysis. Then the vulnerabilities are ranked according to the scores leading to the formulation of the system requirements for fixing the vulnerabilities and thus mitigating the risks. However, we propose to use the legal specifications such as the AI ACT, GDPR to prioritize the security requirements.

Prioritization of identified functional requirements should be based on the applicable EU legislation. Despite functional cybersecurity requirements being inherently technical, the EU is adopting or has already adopted a number of legal acts (regulations or directives) establishing (cybersecurity) legal obligations for providers or deployers of digital services. These legal acts may pertain to different domains, including cybersecurity itself (e.g., the EU Cybersecurity Act), privacy and data protection (e.g., the GDPR or the Law Enforcement Directive) or AI (e.g., the upcoming AI Act). To ultimately prioritize the implementation of cybersecurity measures in the governance systems of LEAs, the identified functional cybersecurity requirements will be confronted with the existing body of EU legislation to:

- Possibly trace back the origin of the functional cybersecurity requirement to one or more EU legal acts; and
- Distinguish functional cybersecurity requirements that LEAs must (i.e., have an absolute obligation to) or should (i.e., have a best-endeavors obligation to) implement in their governance systems.

The legal requirements have to be introduced early in the requirements derivation process and should be used to amend the severity grading of the requirements. Technical requirements might not have a severity level matching the legal requirements. If the severity level of the technical requirement is low, but the level of the legal requirement is high, the severity level has to be increased. In the opposite case, where the severity level of the legal requirement is low, but the technical requirement is high, the technical severity level should be retained. Consequently, the sum of the severity levels should be used.

3.4 The Structure of the Requirements

The requirements should be formulated in accordance with a combination of the formats given in Section 2.1. By combining the principles of FOI (Hansson et al. 2011; Hallberg et al. 2018) and Hull et al. (2005) ten principles have to be fulfilled. Due to the *atomic* principle the use of sub-requirements might be necessary. However, the increased readability of the requirements outweighs the increased number of requirements to handle in that case.

The set of requirements might be structured in the form of a tree (a graph), preferably with each of the identified assets at the highest level. The next level might either indicate the stage in the life cycle of the system (taken from Malatras et al. 2020), or the priority (severity level) of the requirement. The opposite order might also work if the number of high priority requirements are low, because then all can be addressed at once. The graph format is easily converted to a structure where each requirement is self-contained.

The exact format of the requirements cannot be given, because it depends on the studied AI system, where it is used (is it a key function within the organization, or not), to what extent it is used, if it is the only tool used for that purpose, or if there are alternative tools/processes to be used, the legal requirements in each specific case etc. The differences will require a lower or higher number of sub-requirements to be used. Consequently, the detailed format is depending on the situation. However, the structure of the requirements should at least contain the following fields:

1. Unique ID
2. Asset
3. Life Cycle Stage
4. Combined Priority
5. Parent ID
6. Requirement Text

The *Unique ID* field can for example be an Universally Unique IDentifier (UUID) or any other unique identifier for a requirement. The *Asset* field describes the affected asset. The field can be either numeric or text based, as long as it is possible to identify which asset the requirement belongs to. The *Life Cycle Stage* is taken from the ENISA AI Threat Landscape (Malatras et al. 2020) list of life cycle stages for an AI system. The *Combined Priority* field is the sum of the CVSS score of the identified vulnerability connected to the requirement and the identified legal priority. If needed the CVSS score and the legal priority can have their own fields in the structure. The Parent ID enables the use of sub-requirements if needed due to formatting issues. If the requirement does not have a parent, a zero, “0”, can be used as parent ID. The *Requirement Text* should be formulated in accordance with the FOI and Hull et al. principles in combination, giving ten principles to follow (see Section 2.1).

4. Conclusions

This report is part of the ALIGNER EU project and presents a model for deriving cybersecurity requirements for AI systems and their structure. The requirements structure should be used in latter stages of the project when the cybersecurity requirements of AI systems used by LEAs are to be found. As described in Section 1.2, this deliverable complements deliverables D3.1-D3.3 by providing a method for deriving cybersecurity requirements, as well as a template for the structure of the requirements. The result of this deliverable will also be used for further work in tasks T3.3, T3.4, and T4.3, where the project will screen AI technologies for their potential (mis)use. Finally, the results from these screening tasks, using the method and requirements structure template from this deliverable, will be reported in the upcoming D5.6, D5.7, and D5.8 roadmap deliverables.

The model for requirements derivation is based on an extended version of the ENISA AI Threat Landscape model. The extensions include selected additions of NIST procedures for risk management, both for AI, as well as general cybersecurity domains. In this way, the derived model is based on well-established models and procedures for cybersecurity risk and threat management.

The structure of the cybersecurity requirements is based on the information used in the proposed model for requirements derivation. The knowledge gained during the development of the model has been combined with a set of principles for formulation of good requirements at a general level. The principles in turn are based on work done by FOI and Hull et al. (2005). The presented requirements structure contains (at least) six fields, but can easily be extended if needed. Reasons for extension can for example be complex requirements needing to be divided into several subsets, a higher focus on the priority of the requirements, or a complex set of identified assets that have to be divided into subset.

5. References

Beckers, K., Fassbender, S., Heisel, M. & Schmidt, H. (2012). *Using Security Requirements Engineering Approaches to Support ISO 27001 Information Security Management Systems Development and Documentation*. 2012 Seventh International Conference on Availability, Reliability and Security. DOI: 10.1109/ARES.2012.35

Berghe, A. v. d., Yskout, K. & Joosen, W. (2022). *A Reimagined Catalogue of Software Security Patterns*. 2022 IEEE/ACM 3rd International Workshop on Engineering and Cybersecurity of Critical Systems (EnCyCriS). DOI: 10.1145/3524489.3527301

Bertrand, P., Darimont, R., Delor, E., Massonet, P., & van Lamsweerde, A. (1998), Grail/kaos: an environment for goal driven requirements engineering, in *Proceedings 20th International Conference on Software Engineering (ICSE)*. IEEE ACM.

Bezombes, P., Brunessaux, S. & Cadzow, C. (2023). *Cybersecurity of AI and Standardisation*. European Union Agency for Cybersecurity (ENISA).

de Boer, M. H., Bouma, H., Kruithof, M. C., ter Haar, F. B., Fischer, N. M., Hagendoorn, L. K., & Raaijmakers, S. (2017). Automatic analysis of online image data for law enforcement agencies by concept detection and instance search. In *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies* (Vol. 10441, pp. 155-168). SPIE.

Chase, J., Phong, T., Long, K., Le, T., & Lau, H. C. (2021). GRAND-VISION: An Intelligent System for Optimized Deployment Scheduling of Law Enforcement Agents. In *Proceedings of the International Conference on Automated Planning and Scheduling* (Vol. 31, pp. 459-467).

Contardo, P., Sernani, P., Falcionelli, N., & Dragoni, A. F. (2021). Deep Learning for Law Enforcement: A Survey About Three Application Domains. In *RTA-CSIT* (pp. 36-45).

Cordeiro, A., Vasconcelos, A. & Correira, M. (2022). *A Catalog of Security Patterns*. 29th Conference on Pattern Languages of Programs (PLoP).

Devillé, R., Sergeyssels, N., & Middag, C. (2021). Basic Concepts of AI for Legal Scholars. In J. De Bruyne & C. Vanleenhove (Eds.), *Artificial Intelligence and the Law* (Centrum voor Verbintenissen-en Goederenrecht, pp. 1-22). Intersentia. DOI:10.1017/9781839701047.002

Eren, E., Casaburo, D. & Vogiatzoglou, P. (2022). *State-of-the-art reports on ethics & law aspects in Law Enforcement and Artificial Intelligence*. ALIGNER D4.1

European Commission (2005). *Women and Science: Excellence and Innovation - Gender Equality in Science*. Commission Staff Working Document, SEC(2005) 370, 11 March 2005. Retrieved from https://data.consilium.europa.eu/doc/document/ST-7322-2005-INIT/en/pdf.

European Commission (2021), *Proposal for a Regulation of the European Parliament and of the Council, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Forum of Incident Response and Security Teams (FIRST). (2023). Common Vulnerability Scoring System version 3.1: User Guide. https://www.first.org/cvss/user-guide

Gollmann, D. (2011) *Computer Security*. 3rd edition. Wiley

Hallberg J., Bengtsson J. & Karlzen H., (2018), *Beskrivning av hot vid säkerhetsanalyser*, Technical Report FOI-R--4676—SE. Swedish Defence Research Agency.

Hansson, J., Granlund, H. & Hallberg, N. (2011). *Att uttrycka krav i materielmålsättningar – Formulera och granska*. Technical Report FOI-R--3250—SE. Swedish Defence Research Agency.

Hull, E., Jackson, K. & Dick, J. (2005). *Requirements Engineering*. 2nd edition. Springer.

Jackson, M., (2001), *Problem Frames: Analyzing and structuring software development problems*. Addison-Wesley.

Lückerath, D. (2021). *ALIGNER D1.2 – Project Handbook*. ALIGNER – Artificial Intelligence Roadmap for Policing and Law Enforcement. European Commission.

Lund, M. S., Solhaug, B., & Stølen, K., (2011), *Model-Driven Risk Analysis: The CORAS Approach*, 1st ed. Springer Publishing Company, Incorporated.

Luo, X., Shen, R., Hu, J., Deng, J., Hu, L., & Guan, Q. (2017). A deep convolution neural network model for vehicle recognition and face recognition. *Procedia Computer Science*, *107*, 715-720.

Madiega, T. & Chahri, S. (2023). BRIEFING -- EU Legislation in Progress -- Artificial intelligence act. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

Malatras, A. & Dede, G. (2020). *AI Cybersecurity Challenges – Threat Landscape for Artificial Intelligence.* European Union Agency for Cybersecurity (ENISA). DOI 10.2824/238222

Markarian, G., Karlovic, R., Nitsch, H., & Chandramouli, K. (Eds.) (2022). *Security Technologies and Social Implications*. Wiley.

Marshall, A., Parikh, J., Kiciman, E. & Kumar, R. S. S. (2019). *AI/ML Pivots to the Security Development Lifecycle Bug Bar*. Micosoft Security. https://learn.microsoft.com/en-us/security/engineering/bug-bar-aiml

Mouratidis, H., & Giorgini, P., (2007), Secure tropos: a security oriented extension of the tropos methodology. In *International Journal of Software Engineering and Knowledge Engineering*.

National Institute of Standards and Technology (NIST). (2023a). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. NIST. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

National Institute of Standards and Technology (NIST). (2023b). *The NIST Cybersecurity Framework 2.0*. Initial Public Draft. NIST. https://doi.org/10.6028/NIST.CSWP.29.ipd

National Security Commission on Artificial Intelligence (NSCAI). (2021). *Final report*. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

Saif, H., Dickinson, T., Kastler, L., Fernandez, M., & Alani, H. (2017). A semantic graph-based approach for radicalisation detection on social media. In *European semantic web conference* (pp. 571-587). Springer, Cham.

Scarfone, K., Jansen, W. & Tracy, M. (2008). *Guide to General Server Security*. Special Publication 800-123. NIST. https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-123.pdf

Schiebinger, L., & Klinge, I. (2020). *Gendered innovations 2: How inclusive analysis contributes to research and innovation*. Luxembourg: Publications Office of the European Union.

Schmidt, H., Hatebur, D., & Heisel, M., (2011), A pattern- and component-based method to develop secure software," in *Software Engineering for Secure Systems: Academic and Industrial Perspectives*, H. Mouratidis, Ed. IGI Global.

Shevchenko, N., Chick, T. A., O'Riordan, P., Scanlon, T. & Woody, C. (2018). *Threat Modeling: A Summary of Available Methods*. White paper, Carnegie Mellon University, USA. https://insights.sei.cmu.edu/documents/569/2018_019_001_524597.pdf

Susi, A., Perini, A., Mylopoulos, J., & Giorgini, P., (2005), The tropos metamodel and its use, In *Informatica*.

Vitorino, P., Avila, S., Perez, M., & Rocha, A. (2018). Leveraging deep neural networks to fight child pornography in the age of social media. *Journal of Visual Communication and Image Representation*, *50*, 303-313.

Westman, T., Svenmarck, P., & Chandramouli, K. (2022). *ALIGNER D3.1 – Impact Assessment of AI Technologies for EU LEAs.* European Commission.

Wilhjelm, C. & Younis, Y. (2020). *A Threat Analysis Methodology for Security Requirements Elicitation in Machine Learning Based Systems*. 2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C). DOI 10.1109/QRS-C51114.2020.00078.

Wolff, J. (2020). *How to improve cybersecurity for artificial intelligence*. Brookings. https://www.brookings.edu/articles/how-to-improve-cybersecurity-for-artificial-intelligence/

Yampolskiy, R. V. (2019). *Artificial Intelligence Safety and Security*. Taylor & Francis.