SPARTA

D7.1

AI systems threat analysis mechanisms and tools

Project number	830892
Project acronym	SPARTA
Project title	Strategic programs for advanced research and technology in Europe
Start date of the project	1st February, 2019
Duration	36 months
Programme	H2020-SU-ICT-2018-2020

Deliverable type	Report
Deliverable reference number	SU-ICT-03-830892 / D7.1 / V1.1
Work package contributing to the deliverable	WP7
Due date	July 2020 – M18
Actual submission date	2nd June, 2021

Responsible organisation	TEC
Editor	Erkuden Ríos
Dissemination level	PU
Revision	V1.1

Abstract	This document reports the threat modelling and analysis landscape for AI systems, and describes the AI Threat model developed in SPARTA to establish a common language and understanding of the threats against AI systems together with the key threat attributes and terms. The deliverable also describes the Knowledge Base developed in SPARTA that collects the initial body of knowledge on AI Threats. The report also offers a thorough study on the challenges of AI systems compliance with GDPR.
Keywords	AI threat analysis, AI threat model, threat taxonomies for AI, AI threat knowledge base, GDPR compliance in AI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 830892.

Editor

Erkuden Ríos (TEC)

Contributors (ordered according to beneficiary numbers)

Manon Knockaert, Sophie Everarts de Velp (UNamur)

Mohammad Reza Norouzian (TUM)

Carmen Palacios, Cristina Martínez (TEC)

Raúl Orduña, Xabier Etxeberria, Amaia Gil (VICOM)

Marek Pawlicki, Michal Choras (ITTI)

Reviewers (ordered according to beneficiary numbers)

Daniel Meyer, Alexandra Kobekova, Marc-Philipp Ohm (UBO)

Mario Reyes (EUT)

Executive Summary

Due to the increasing adoption and pervasiveness of Artificial Intelligence (AI) systems in many industry domains, means for protecting these systems from potential security and privacy threats are necessary, together with mechanisms to mitigate the impact of attacks in case they occur. In this endeavour, the first step is to provide AI system developers with means for better understanding the potential incidents and attacks their systems may suffer, with a focus on those that are particular to AI due to the nature of the assets and processes involved in AI system development and operation.

Being a new area of research and standardisation attempts, this document reports the results of the analysis of the various classifications and studies on AI threats that are recently available in the literature or still under construction in on-going standardisation initiatives. Building from a comprehensive stocktaking, the deliverable proposes an AI threat modelling method to capture all the necessary information about specific threats against AI system components and classifies them according to several aspects such as data attacker goal, target asset, data analysis pipeline phase, etc.

The document describes the AI threat model developed in SAFAIR Program together with the Knowledge Base implemented to capture the body of knowledge analysed, which follows the organisation and taxonomy of the AI threat model.

The AI threat model developed establishes a common language and understanding of the threats against AI systems together with the key threat attributes and terms. These concepts have been used as guidance and common baseline in WP7 SAFAIR tasks.

The report offers a thorough study on the challenges of AI systems compliance with General Data Protection Regulation (GDPR) and explains how the legal framework requirements impact the design and operation of the AI systems that process personally identifiable information.

Therefore, the report advances in the principles of the European *AI strategy* and *The Ethics Guidelines for Trustworthy Artificial Intelligence (AI)* report issued by the High-Level Expert Group on Artificial Intelligence (AI HLEG) launched by the European Commission in June 2018.

It is important to note that the focus of the document is on Machine Learning (ML) systems and not on Artificial Intelligence systems in general (such as expert systems, reasoners, fuzzy systems, etc.) since this is the work scope of SAFAIR Program tasks. Therefore, when referring to AI systems and threats in the document we refer to those related to ML.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Scope and Purpose

The objective of this deliverable is to report the Artificial Intelligence (AI) system threat modelling methodology that has been elaborated in the SPARTA SAFAIR Program task T7.1 “Threat modelling for AI systems”. The document analyses existing AI system threat taxonomies and information capturing models and builds on top of them to advance the state of the art in identification and analysis of threats over AI systems and their components. The ultimate goal is to aid in the creation and design of secure and reliable AI systems that consider potential threats against security and privacy properties.

The focus of the deliverable is on Machine Learning (ML) systems, and other types of Artificial Intelligence systems (such as expert systems, reasoners, fuzzy systems, etc.) are not studied, since ML this is the work scope of SAFAIR Program. Therefore, when referring to AI systems and threats in the document we refer to those related to ML.

The document collects example threat models for selected target AI systems that will be tested in the other tasks of SPARTA SAFAIR Program through validation of designed defensive mechanisms.

The current growing adoption of AI and its expected pervasiveness in IT systems have increased the necessity to identify potential threats against AI systems, in terms of attacks and situations that endanger availability, data integrity, or data confidentiality. Looking beyond initial publications on adversarial examples fooling machine learning systems, there is a need for an approach that tackles the threats against AI systems and their constituent parts from early design to try to effectively prevent risks. Other security threats such as confidential data leaks and privacy threats to citizens like undesired disclosure of personal data may also appear in data processing. In order to handle all these threats, the task T7.1 aims to design a systematic method and supporting tool for identification and analysis of vulnerabilities of AI systems. The risk analysis and management in AI systems will be the target of this task in relation to providing support to capture the threat knowledge that can be analysed and reused in other phases of the AI system development process and operation.

The focus of SPARTA SAFAIR is on enabling Secure and Reliable AI systems of the future, and therefore, the T7.1 work is oriented towards aiding in the identification of security and privacy issues that may happen in AI systems as well as in the study of attacks against AI system assets, components, or procedures. As shown in Figure 1, three main relationships can be defined between Artificial Intelligence (autonomous computing systems) and cybersecurity threats: 1) securing AI from threats (incidents and attacks) that are particular to AI systems, 2) AI as security mechanism, i.e. the use of AI for defending other systems, and the 3) AI as attacking mechanism, i.e. the use of AI for threatening other systems. The threat modelling task in SPARTA relates only to the first aspect and studies potential vulnerabilities and attack vectors in AI systems and their constituent parts, with a focus on those specificities of AI that are not found in other systems not using machine learning or deep learning.

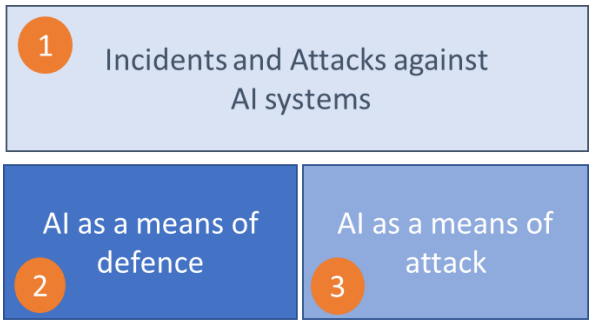

Figure 1: The AI relationships with cyber threats

The present deliverable contributes to reach the requirements of *technical robustness* and *privacy-respectfulness* of AI systems, according to their definition in *The Ethics Guidelines for Trustworthy Artificial Intelligence (AI)* report published by the High-Level Expert Group on Artificial Intelligence (AI HLEG) [1]. The contribution of SAFAIR is with respect to the following aspects:

Technical robustness:

- **Resilience to attack and security.** AI systems should be free from vulnerabilities and corruption or abuse vectors by adversaries both at software and hardware levels. The software level protections avoiding such issues and any potential unintended use of the AI system are the main focus of this deliverable. Threats against AI systems and possible countermeasures are modelled in the SAFAIR AI Threat model and they are collected in the accompanying Knowledge Base.
- **Accuracy**. It refers to the capability of AI systems to correctly perform the function they are intended to, i.e. classification, decision, judgement, etc. over the inputs. Some of the threats and attack techniques studied in this report and captured in the SAFAIR AI Threat Knowledge Base target the accuracy of the ML models.
- **Reliability and Reproducibility**. AI systems are reliable when they are always available when expected, and when they work as planned in the design, without unintended results or behaviour. Reproducibility of AI systems means that the system shows the same behaviour under the same working conditions. These two properties are the target of some of the threats and attack techniques studied herein and captured in the SAFAIR AI Threat Knowledge Base.

Privacy and data governance:

- **Privacy and data protection**. As any other system under the need to comply with GDPR, AI systems should protect personal data and ensure privacy of data subjects. Protections of AI systems against potential means of unfairness and discrimination against data subjects are the focus of SAFAIR task T7.4, and will also be part of the AI Threat model defined herein.
- **Quality and integrity of data**. The accuracy and quality of the training data, test data and results data are essential characteristics of AI systems. As it will be seen, many of the attacks against AI studied in this document relate to the use of these data as threat vectors. Appropriate data management plans and procedures are core to AI system engineering process which covers the whole life-cycle from design to operation. Additionally, permanently maintaining data integrity is a strong security requirement of AI. Threats to data integrity are studied and classified in the SAFAIR AI Threat model and Knowledge Base.
- **Access to data**. Since data is the source of the AI system behaviour, data access control mechanisms and protocols are a must in AI systems, particularly when the data processed is personal data. Data access attacks are also studied in the present report.

While the focus of T7.1 is placed in AI system security, i.e. reliability in terms of availability as well as data assets integrity and confidentiality, other tasks in SAFAIR are dedicated to explainability (and transparency) and fairness of AI described by "*The Ethics Guidelines for Trustworthy Artificial Intelligence (AI)"* report, tasks T7.3 and T7.4 respectively. We refer the interested reader to deliverable D7.2 *Preliminary description of AI system security mechanisms and tools* that collects the results from those tasks.

The D7.1 document establishes a common language and understanding of the threats against AI systems together with the key threat attributes and terms, with a particular emphasis on attacks against AI systems. These concepts have been used as guidance and common baseline in WP7 SAFAIR tasks.

Finally, this deliverable offers a wide view on the analysis of threats against AI systems by including the study of challenges and risks of AI systems concerning the compliance of privacy and data protection principles coming from GDPR.

The content of the present deliverable is closely related to the deliverable D7.2 *Preliminary description of AI system security mechanisms and tools*, and while the former provides a theoretical approach on how to analyse vulnerabilities and the threats against the AI systems, the latter addresses a more practical approach on how to tackle them. Therefore, both documents are aligned and complementary and the recommendation is to read both.

1.2 Methodology

The threat analysis of AI systems carried out to elaborate this report started in January 2019 with a deep stocktaking of the relevant existing literature, reports, white papers, surveys, legislation, initiatives and other research projects around taxonomies, models and methods for both AI systems and AI system threats. The main exponents of studied sources due to their internationally recognised expertise and impact are presented in the following state of the art section.

As it is described later in the state-of-the-art analysis, the present effort of task T7.1 in SPARTA coincides in time with some international standardisation initiatives on AI system threat classifications and studies of potential security solutions, such as the NIST-IR-8269, ENISA Big Data taxonomy and the AI threat ontology by ETSI Securing Artificial Intelligence ISG, to name a few. Therefore, this task has continuously surveyed the progress and collaborated with some of these initiatives through the participation of some of the T7.1 partners in the working groups.

Similarly, as a means to address the increasing interest on understanding GDPR implications for AI systems and the necessity of many AI systems of fulfilling GDPR privacy principles, the study of the AI threats was complemented with a thorough analysis of the articles of the GDPR explaining the challenges and clarifications with respect to GDPR compliance for AI systems.

As a result, the task has produced a report that offers a holistic view on how information about threats against AI specific components can be captured for a comprehensive understanding of the risks to which AI systems are exposed to.

1.3 Structure of the document

The structure of the document is as follows:

- Chapter 1 is the current section presenting the objectives, scope and structure of the document.
- Chapter 2 presents the state of the art on terminology for AI system assets as well as AI threats and threat taxonomies. The chapter summarises the result of the stocktaking of attack examples reported in the literature and collects different methods for threat modelling and analysis applicable to AI systems.
- Chapter 3 documents the threat modelling methods and knowledge base tool for AI systems proposed in SPARTA.
- Chapter 4 presents the main GDPR compliance challenges in AI and documents the analysis of potential privacy attacks and incidents that need to be considered in these systems.
- Chapter 5 presents the conclusions of the report.

1.4 Relation with other tasks in SPARTA

The contents of the present report have a close relationship and alignment with other two main tasks in SPARTA:

- T-SHARK (Full Spectrum Situational Awareness) - Task 4.3 “ALL Data based threat intelligence” Challenges Contest (M01-M36; Task Lead: CESNET): focused on building CSTI on aggregated data from internal and external sources: structured and unstructured, open

access, internet and other data of public availability as well as closed or confidentiality rules governed data. The knowledge base proposed in D7.1 Chapter 3 supports the CSTI of AI systems and may be considered a source that contributes to the enhancement of security and privacy by design in AI systems.

- HAII-T (High-Assurance Intelligent Infrastructure Toolkit) - Task 6.5 Privacy-by-design (M01-M24; Task Lead: BUT). The privacy principles and requirements studied in D7.1 Chapter 4 contribute to the identification of privacy-by-design and privacy-by-default needs of AI systems and therefore are complementary to HAII-T privacy techniques supporting the privacy-aware design of AI systems.

Chapter 2 State of the art of Threat Analysis for AI Systems

2.1 Introduction

As in any other system, the analysis of potential threats against AI systems would start from the analysis of the system constituent parts and assets to protect, identifying their vulnerabilities and the possible ways these components could unintentionally fail or suffer attacks from malicious adversaries. In this Chapter, we present the state of the art on terminology approaches and taxonomies for both AI system assets (in Section 2.2) as well as threats against AI systems (in Section 2.3).

The Chapter (in Section 2.4) also gathers an extensive collection of examples of attacks against AI systems reported in the literature explaining the techniques used in them. They are a good starting point to better understand the potential flaws and attack vectors that AI systems may have. The Chapter then (in Section 2.5) explains the key concepts and methods for threat modelling and analysis applicable to AI systems. The conclusion Section 2.6 summarises the key findings of the Chapter 2.

2.2 Taxonomies of AI systems and their assets

As the first step to understand and analyse threats against AI systems, this section explores the types of AI systems and the assets within them which are the potential targets of the attacks. Vulnerabilities in these assets will lead to external or internal attacks as well as may bring security and privacy flaws of the AI systems. Protecting these assets is fundamental to build trustworthy AI systems and prevent incidents that may put the system at risk.

2.2.1 AI asset taxonomies

In this Section, we identify and analyse existing main taxonomies of assets in AI systems, differentiating the potential targets of attacks against AI system infrastructure. The aim is to help in understanding which are the assets to protect in AI and the major differences between AI systems and other IT systems with respect to assets.

First, it is interesting to note that AI systems are usually named *Big Data systems* in the literature, due to big data technologies support in many cases the execution of AI algorithms. Therefore, big data taxonomies often include concepts and terms of AI-based systems such as data analytics algorithms. .

As major taxonomies, we have selected the EU-level relevant *ENISA big data taxonomy* together with the internationally recognised *NIST Big Data Reference architecture* and the *Cloud Security Alliance's (CSA) big data taxonomy*. The Section ends with other taxonomies in the literature which are holistic enough to serve the purpose of a comprehensive asset catalogue.

2.2.1.1 ENISA Big Data asset taxonomy

The ENISA Big Data taxonomy has been developed by the ENISA Threat Landscape (ETL) Group in the *ENISA Big Data Threat Landscape and Good Practice Guide, Jan 2016* [2]. This taxonomy focuses on assets to protect in the scope of Big Data Information and Communication Technology (ICT) systems, which could be abstract, virtual, physical or human assets. The taxonomy classifies the assets in a hierarchical manner, and distinguishes between five main categories shown in Figure 2.

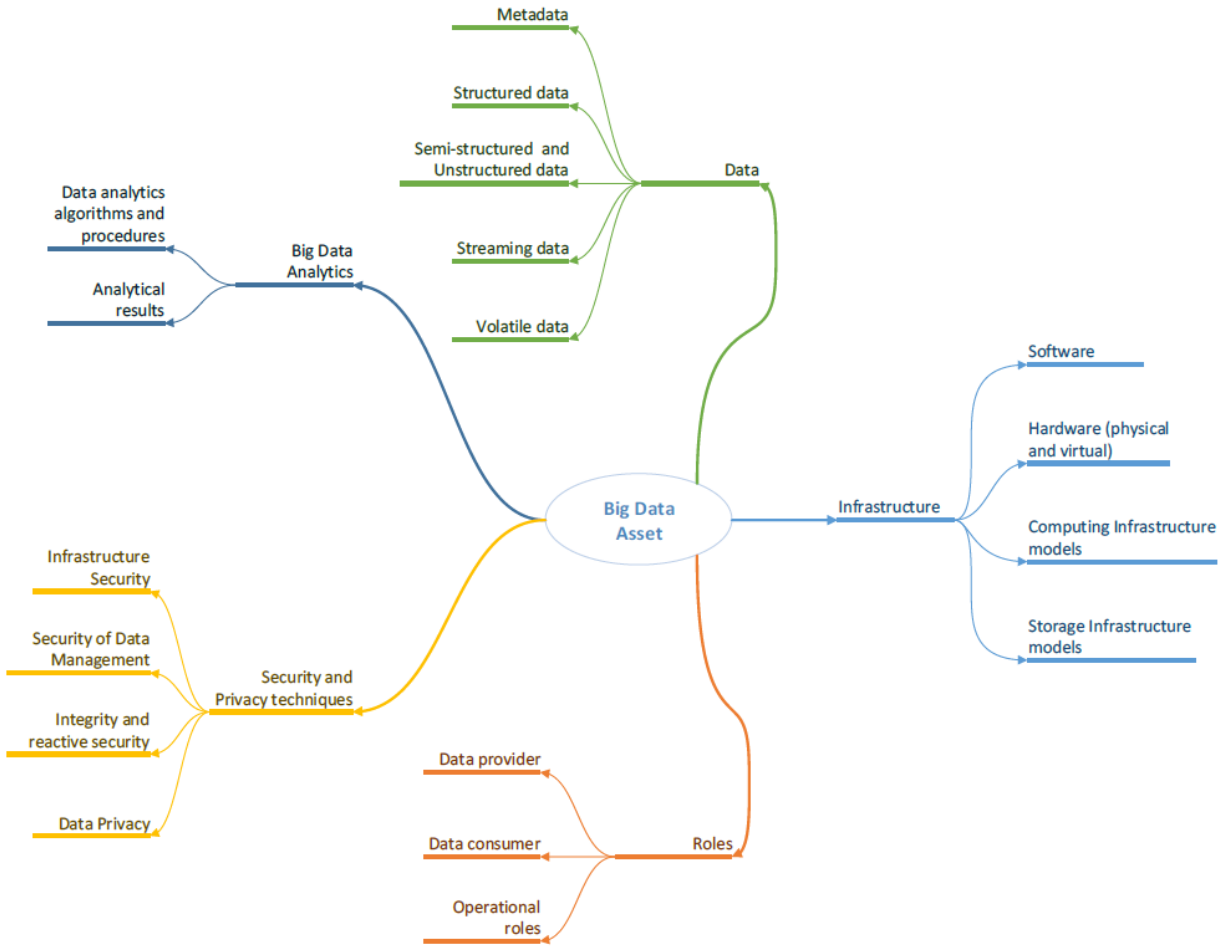

Figure 2: ENISA Big data asset taxonomy (asset group and asset types) [2]

Making a zoom on the main type of assets associated with AI systems as we deal with them in SAFAIR, the taxonomy identifies a *Big Data Analytics* category (see Figure 3) which includes the following sub-categories:

- *Data analytics algorithms and procedures*, which includes assets of the analytical process that may be carried out by an AI system, including algorithm code, model parameters and configuration, etc.
- *Analytical results*, which assimilate to AI system output, be it *textual* or *graphical* outcome.

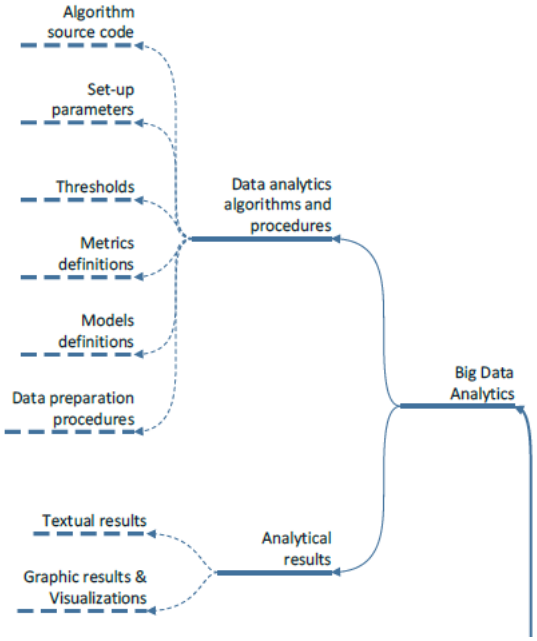

Figure 3: ENISA Big data asset taxonomy – Big Data Analytics category [2]

2.2.1.2 NIST Big Data Reference Architecture

The *NIST conceptual model of Big Data architecture* (NBDRA) described in [3] differentiates five logical functional components in a Big Data system connected by interoperability interfaces, i.e., services, (see Figure 4). *Management* and *Security and Privacy* are transversal layers (fabrics) to all the five functional components.

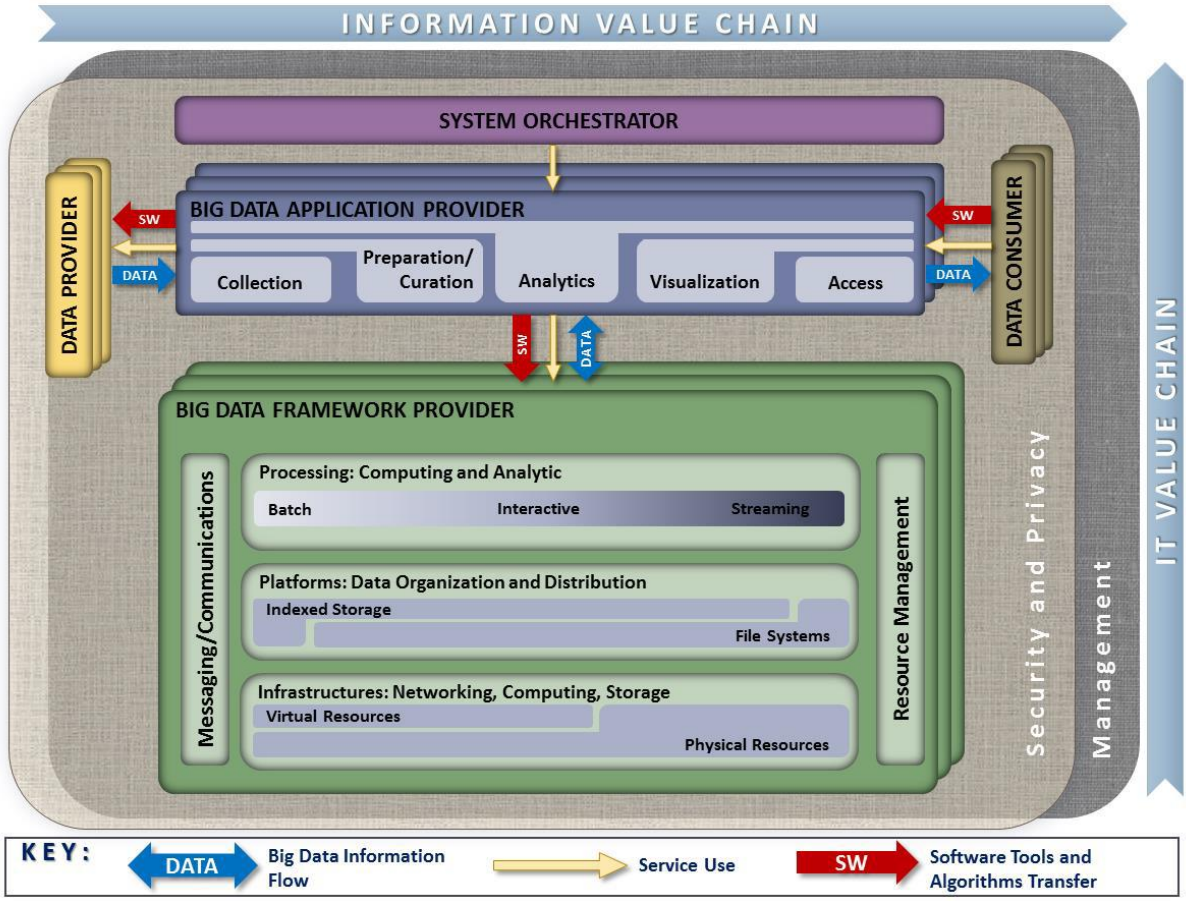

Figure 4: NIST Big Data Reference Architecture (NBDRA) [3]

The NBDRA proposes two main axes along which the architectural components are distributed: the *Information Value Chain* (horizontal axis) and the *Information Technology Value Chain* (vertical axis). These two axes represent the main value chains of the Big Data systems.

2.2.1.3 Cloud Security Alliance (CSA)'s Big Data Taxonomy

The Big Data working group of CSA published in 2014 a comprehensive report on Big Data taxonomy entitled *Big Data Working Group Big Data Taxonomy* [4]. As shown in Figure 5, the classification they proposed consists in six dimensions around the nature of the data processed: *data*, *compute infrastructure*, *storage infrastructure*, *analytics*, *visualization*, and *security and privacy domains*. The main objective of this taxonomy was to identify the types of infrastructures for data computation and storage together with data analytics techniques and security and privacy frameworks existing at that time.

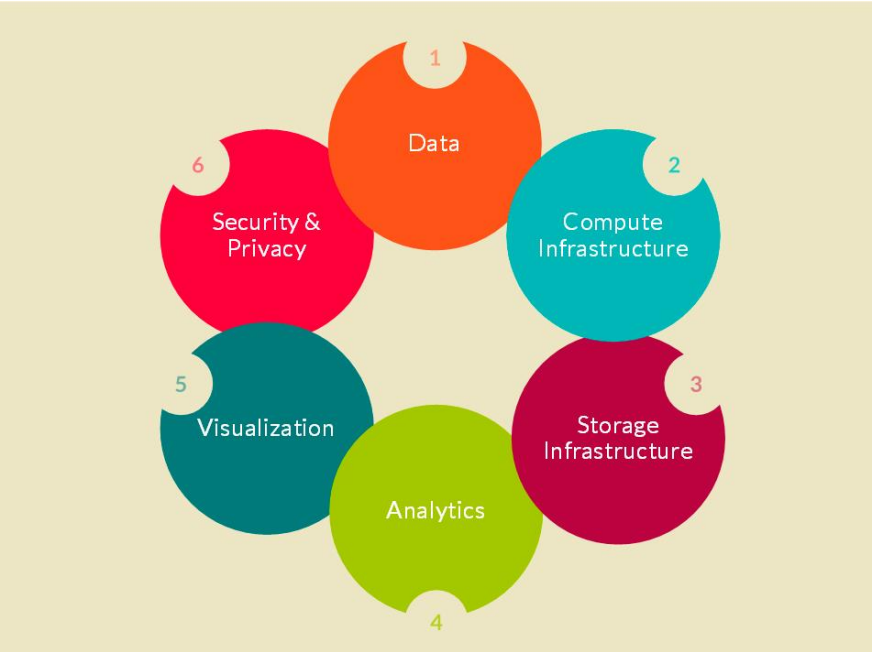

Figure 5: The 6 dimensions of Big Data Taxonomy by CSA [4]

In the Analytics domain of the taxonomy, the CSA classifies Machine Learning algorithms or learning models by Algorithm Type (see Figure 6), differentiating four major classes: *Supervised Learning*, *Unsupervised Learning*, *Reinforcement Learning* and *Semi-Supervised* Learning. It is interesting to note that this classification of learning approaches cannot be found in ENISA taxonomy [2] while it is very relevant since the learning type implies that some particular assets may exist or not in the AI system (e.g. labelled data).

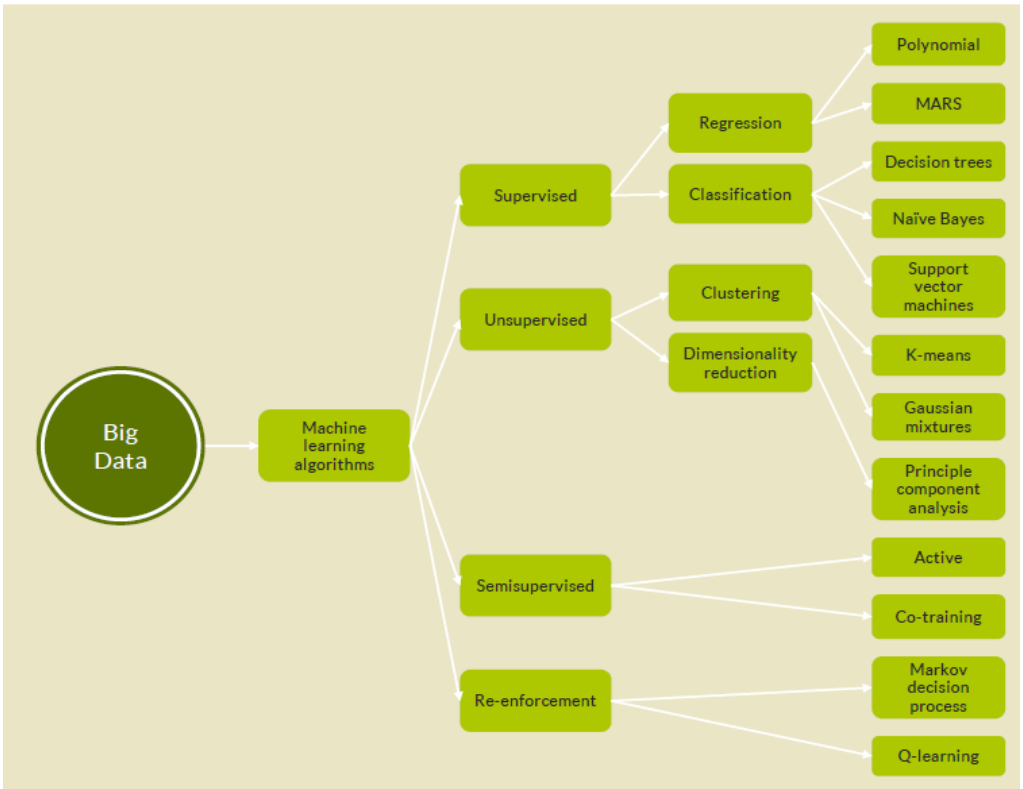

Figure 6: Analytics dimension classified by Algorithm type by CSA [4]

2.2.1.4 The Software Engineering Institute's (SEI) Machine Learning asset taxonomy

The Software Engineering Institute (SEI) of Carnegie Mellon University, in their report of *Comments on NIST IR 8269 (A taxonomy and terminology of adversarial machine learning)* [5] of February 2020, provided a complete representation of a ML process, as shown in Figure 7.

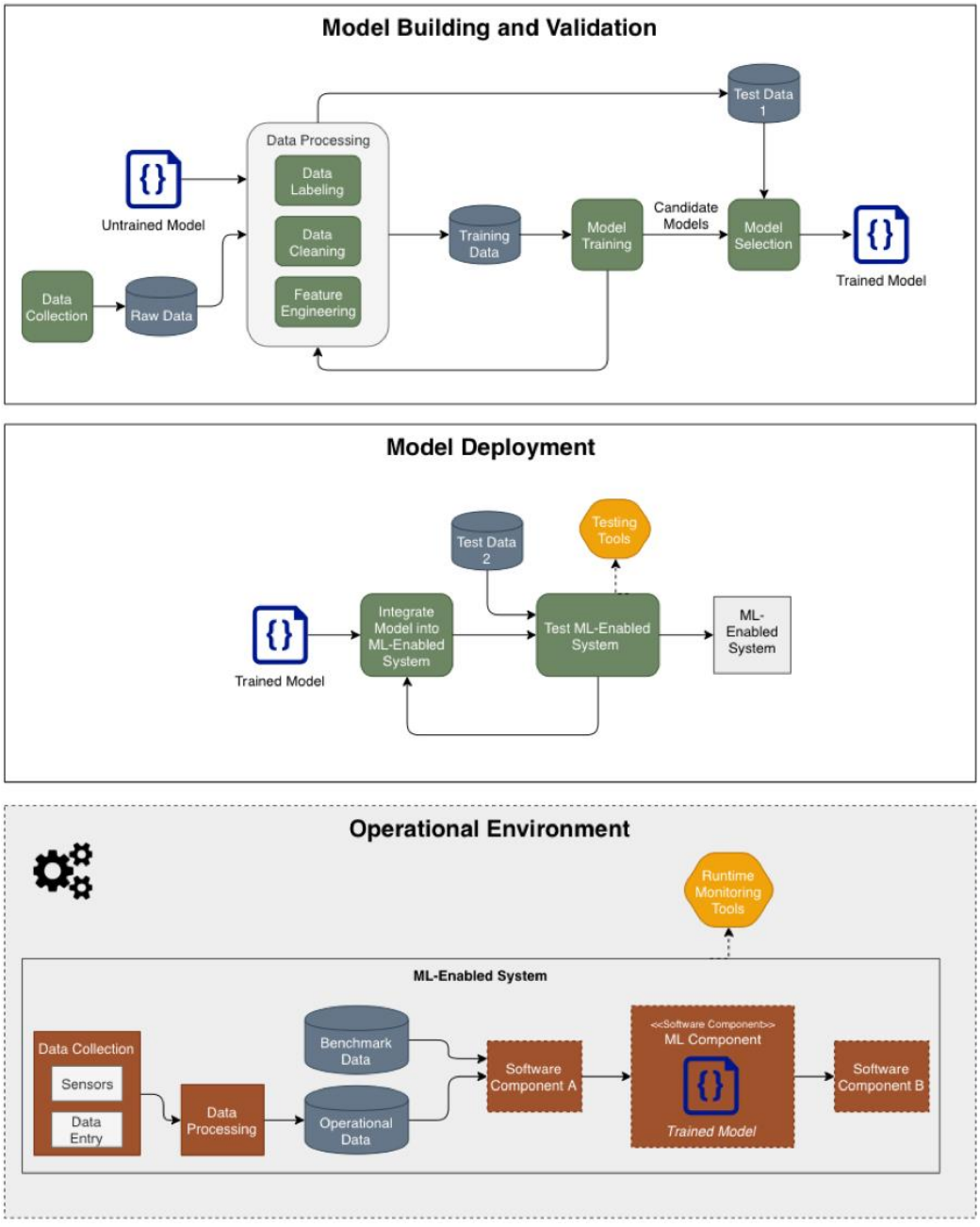

Figure 7: A Machine Learning process by SEI [5]

As it can be seen in Figure 7, according to SEI, four major asset types are involved in the ML process, as follows:

- *Machine Learning Model* (documents with brackets in Figure 7): The ML process starts with an *Untrained Model* as input, which is trained for the objective task of the ML system. The resulting adjusted model is the *Trained Model*. While *Untrained model* is usually of no relevance for the adversaries, the *Trained Model* is the target of multiple attacks as it contains the core of the functioning of the ML system. Usually, the Trained Model needs frequent updating in operation which would be represented as loops in the ML process.

- *Data* (grey-blue datasets in Figure 7): The *Raw Data* is the dataset used at training phase to build the Trained model. At training step, the Raw Data is usually pre-processed and then split into *Training Data* and *Test Data 1*. While Training data is the part of the dataset used to train the model, the Test Data 1 is the data set used to validate the model, i.e. verify that the model is functional and can produce results that generalise beyond the Training Data. Test Data 2 is used by developers at Trained Model deployment step for the validation of the Trained Model software implementation. As explained by SEI, Test Data 2 is often overlapping data items of the Training Data or Test Data 1. Finally, at operation phase, the *Operational Data* represents the input data for the model at operation, and the *Benchmark Data* is used to verify at any moment that the system behaves as designed.
- *Processes* (green boxes in Figure 7): are activities carried out in the ML overall process, often with a human directly involved.
- *Supporting tools* (orange hexagons in Figure 7): Two major supporting tools for the ML process are identified. First, the *Testing tool* used to test the ML implementation at deployment step and second, the *Runtime Monitoring tool* which is used to monitor the performance of the ML system at operation.
- *Software components* (burnt-orange coloured boxes in Figure 7): In general, the ML model, the data and the ML activities themselves are usually supported by software components such as the software component that run the ML model at operation.

The novelty of the asset taxonomy by SEI above is that it results from the holistic view of the overall ML process and includes not only data but also other main assets to be protected in the ML system lifecycle. Please note that although SEI does not mention it, as for any other system, all the software components identified run in hardware elements or components that need to be considered as system assets that are potential attack targets.

2.2.1.5 Other relevant taxonomies

Demchenko et al. [6] produced in 2014 a comprehensive study of the architecture components of the Big Data Ecosystem (BDE), which is defined as the "*whole complex of components to store, process, visualize and deliver results to target applications*" [6], and therefore, the ecosystem tries to put order in data types, models and infrastructure elements involved along the Big Data lifecycle.

The authors also proposed a new extended/improved Big Data technologies definition, based on the Gartner definition [7], where analytics relate to the processing of data and information to extract value from high-veracity sources and where the high levels of volume, velocity and variety of data require new data models along the whole data lifecycle.

As illustrated in Figure 8 and Figure 9, the infrastructure components proposed for the BDE support a data lifecycle composed of four major steps (top of Figure 8). Big Data analytics within the ecosystem include four types (Real-time, Interactive, Batch, and Streaming analytics), three different data processing methods (refinery, linking and fusion), and four major types of analytics applications (link analysis, cluster analysis, entity resolution, and complex analysis).

A *Federated Access and Delivery Infrastructure* component provides network connectivity as well as federated access control to consumers of data services (see right hand of Figure 8 and Figure 9).

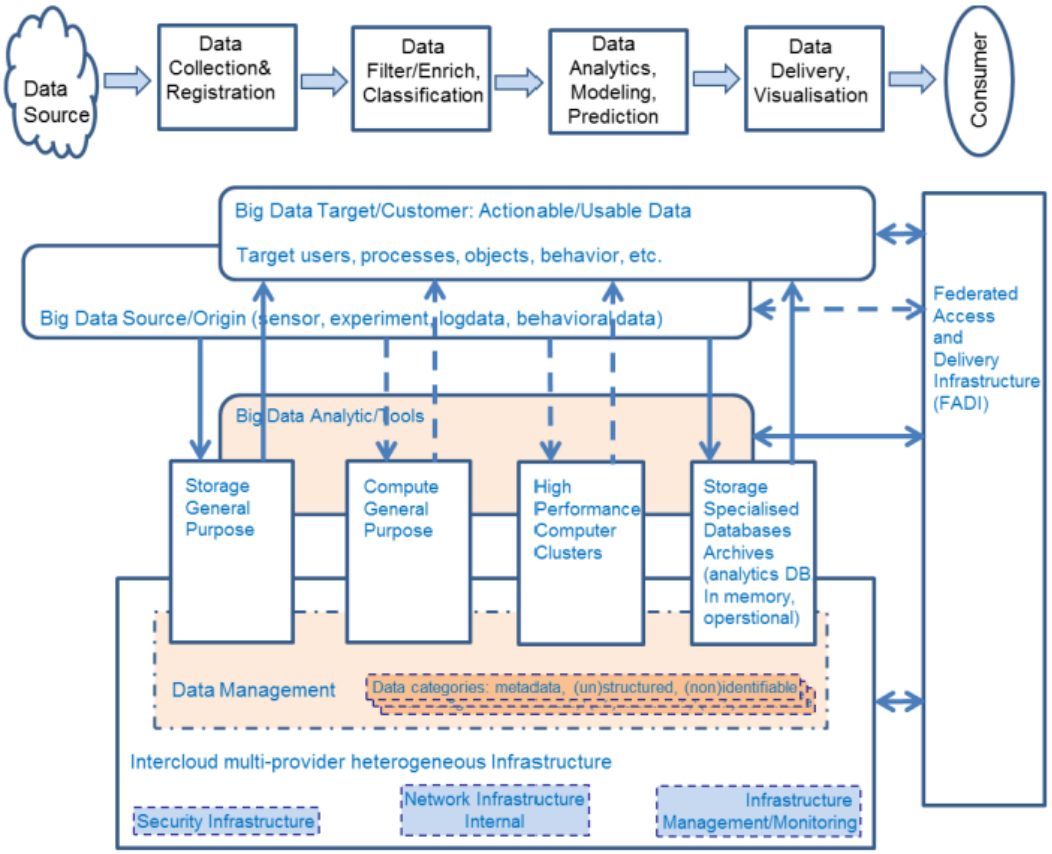

Figure 8: General Big Data Infrastructure components [6]

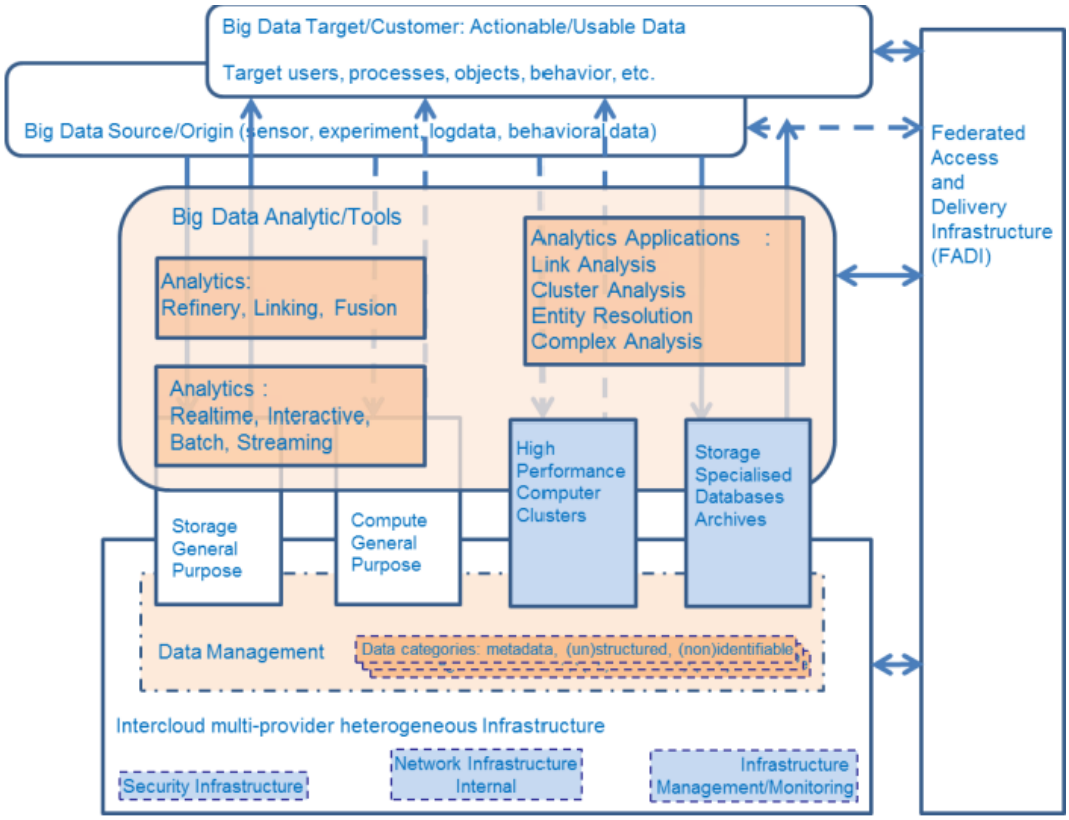

Figure 9: Big Data Analytics Infrastructure components [6]

2.2.2 AI system taxonomies

Thanks to the lately intensive research in Machine Learning and the growing adoption of AI in smart revolutionary applications, there are multiple types and implementations of systems that integrate these technologies in their architectures. In this section we try to describe the landscape of classifications of such systems, which approach them from different perspectives and paradigms. In general, AI systems can be classified by their objective or application, by their learning purpose and by the algorithm(s) used. In the following we provide a summary of the most relevant classifications found in the literature towards the objective of creating the SAFAIR AI threat model.

2.2.2.1 Taxonomy according to AI application paradigm

Golstein [8] defined a taxonomy that is based on the machine learning model task. There are four main groups, as illustrated in Figure 10, namely: *Classification, Continuous Estimation, Clustering* and *Skill Acquisition*.

- *Classification*: This group is formed by machine learning models which classify the input data. Moreover, the machine learning of this group has discrete prediction. Some typical examples of Classification models are *multi-layer perceptrons*.
- *Continuous Estimation*: These machine learning models estimate the relationship between variables. Indeed, the models do not give a specific result, just an approximation. The most known Continuous Estimation machine learning models are *regression models*.
- *Clustering*: Cluster analysis or *clustering* consists in creating clusters or groups of data points from the input dataset such that similarity between data points in the same group is higher than between data points from different groups, i.e. clustering basically classifies objects (data points) on the basis of similarity between them.
- *Skill Acquisition*: This type of machine learning is based on the evolution and the human learning ideas, the learning model modifies its skill, improving little by little. The models take actions concerning the environment in which they are and many of these environments are usually formulated as *Markov decision* processes.

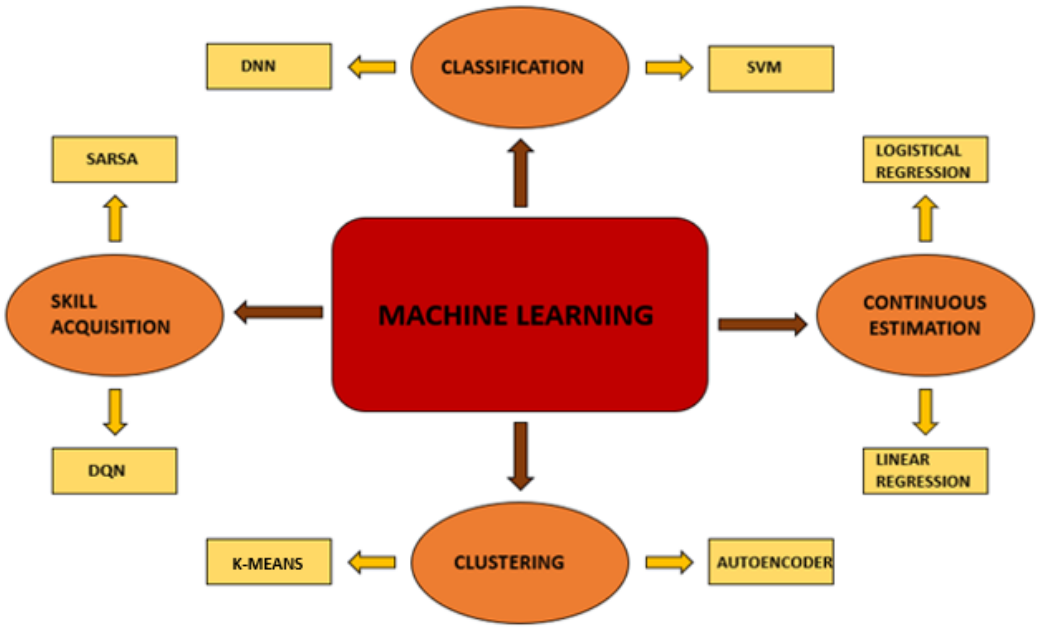

Figure 10: Taxonomy according to AI application paradigm

2.2.2.2 Taxonomy according to AI learning paradigm

This taxonomy is based on the learning method of each machine learning model, as illustrated in Figure 11 below. There are four principal groups, namely *Unsupervised*, *Supervised*, *Semi-supervised* and *Reinforcement learning* [9][10][11].

- *Unsupervised Learning*: The algorithms in this category are trained without the need for target values, as they learn the hidden structure of input data. Commonly known algorithms include *clustering* (which group samples by similarity) and *dimensionality reduction* (which combine features to obtain a reduced number of new ones with same information).
- *Supervised Learning* uses input data and target value to train a model. In case of having categorical labels (classes), *classification* algorithms are used, and if the target variable has continuous values, *regression/prediction* algorithms are used.
- *Semi-supervised Learning:* These algorithms usually have as input a small amount of labelled data together with a large amount of unlabelled data.
- *Reinforcement Learning*: The algorithms of this category are used to create models that are optimized to yield the maximum cumulative reward of actions that change states of an environment.

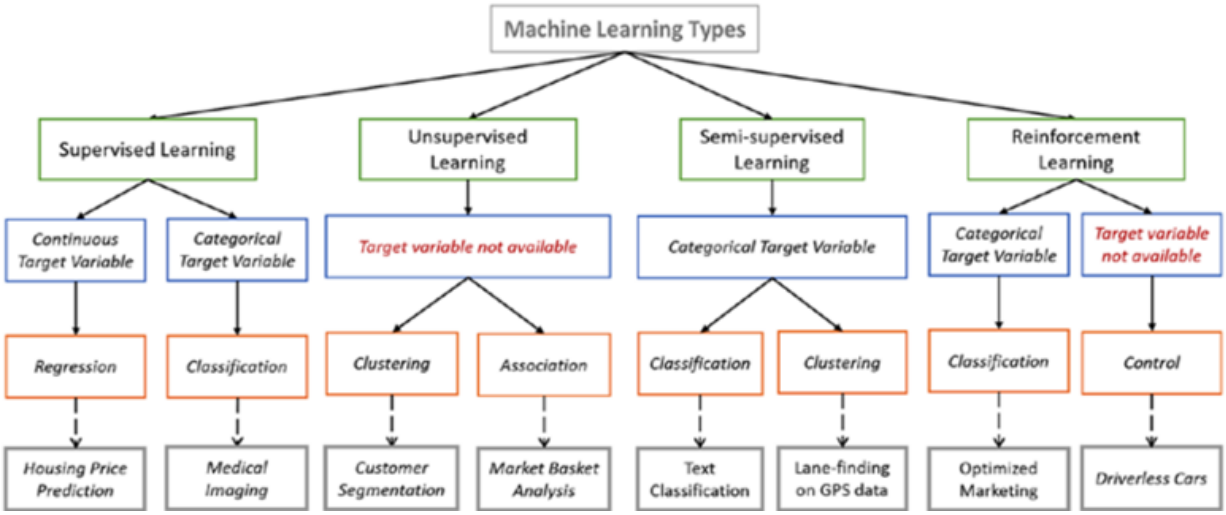

Figure 11: Taxonomy according to AI learning [12]

An AI learning paradigm-based classification can be found in the *Big Data taxonomy* published in 2014 by the Big Data working group of CSA [4]. This taxonomy includes a dimension named "*Analytics*" that comprises the classification by Algorithm type, as illustrated in Figure 6 (see Section 2.2.1.3). In this dimension, the machine learning algorithms are classified in the same categories as the ones seen in section above.

Another implementation of the AI learning paradigm can be found in the *Mindmeister taxonomy*, which groups machine learning algorithms by machine learning families [13]; in other words, the algorithms are grouped depending on the mathematical properties or structure they use for the learning task in the following categories (see Figure 12):

- *Bayesian*: The Bayesian algorithms allow to encode prior knowledge of the data source independently of the information that can be learned from data.
- *Decision Tree*: These algorithms are decision tools based on learning conditional statements allowing to predict the label value or class. The path between roots and leaf represents the conditions that input features satisfy.

- *Dimensionality reduction*: These algorithms obtain a new set of principal variables by transforming and combining high-dimensionality features of input data.
- *Instance based*: These algorithms compare, by means of distances and similarities, instances of new observations to decide upon with instances of the Training dataset.
- *Clustering***:** In these algorithms data samples are grouped based on the idea that samples from the same group share characteristics that also separate them from other groups.
- *Regression*: In these algorithms the model estimates the relations between input variables and targeted continuous values.
- *Rule system*: These algorithms can learn a set of relational rules in order to make a prediction.
- *Regularization*: Learning from data carries a risk of using the noise of the training data as essential part of input variable that should be learned by the model. Regularization algorithms help to guarantee that the learned model is more flexible and avoids overfitting issues.
- *Neural networks*: This category is inspired by the functionality of the brain, a collection of connected nodes that are called neurons and each connection allowing the transmission of signals between neurons.
- *Ensemble*: These methods are used to combine multiple machine learning algorithms to obtain a more accurate prediction.
- *Deep learning*: These algorithms are multi-layered neural networks that are capable to learn by themselves using large amount of data.

Figure 12: AI algorithms taxonomy mind map [13]

The taxonomy based on the AI learning paradigm is the most common taxonomy in machine learning. It is a good way to start classifying machine learning models, because it has no overlaps and gives a lot of information apart from the learning process. For example, the four types identified

above (*Unsupervised*, *Supervised*, *Semi-supervised* and *Reinforcement learning*) characterize very different analyses over data and preparation processes. For instance, while unsupervised learning does not require data to be classified or labelled, supervised learning requires large amounts of data to be correctly labelled beforehand. Reinforcement learning, however, does not require data at all. Moreover, each learning paradigm could be divided in two subsets. Briefly, we could divide unsupervised learning in *clustering* and *dimensionality reduction*. It is also possible to divide supervised learning in *regression* and *classification*. Finally, we could divide reinforcement learning in *Markov* and *evolution*.

Observing this taxonomy, we realize that it is very complete and groups many types of ML algorithms. However, some overlaps with other classifications appear for example in Clustering group of AI application-based taxonomy in previous section and Bayesian learning from AI algorithm-based taxonomy in section below. Because of that, it is reasonable to think that this taxonomy can be used as the baseline taxonomy for AI systems classification with some refinements, as the ones we provide in SAFAIR, which are explained in Chapter 3.

2.2.2.3 Taxonomy according to AI algorithm paradigm

This taxonomy is based on the structure or algorithm used by the machine learning model [10]. There are five main groups, namely: *Connectionist, Evolutionary, Bayesian, Analogy* and *Symbolist* (see Figure 13).

- *Connectionist*: These algorithms such as *backpropagation* are inspired on how the human brain works and creates connections.
- *Evolutionary* learning that tries to resemble how the natural selection in animals works. The *genetic programming* is the most well-known algorithm in this family.
- *Bayesian* learning is a probabilistic inference which tries to deal with uncertainty. The master of these algorithms is *Bayes' theorem* and its derivates.
- *Analogy*: Learning based on analogy aims at identifying similarities between inputs and thereby inferring other similarities. The *support vector machine* is one of the best-known analogists' algorithms.
- *Symbolist*: In symbolist learning intelligence is gained through manipulation of symbols and expressions made of them, similarly as it is done in mathematics. Typical examples are *inverse deduction* algorithms.

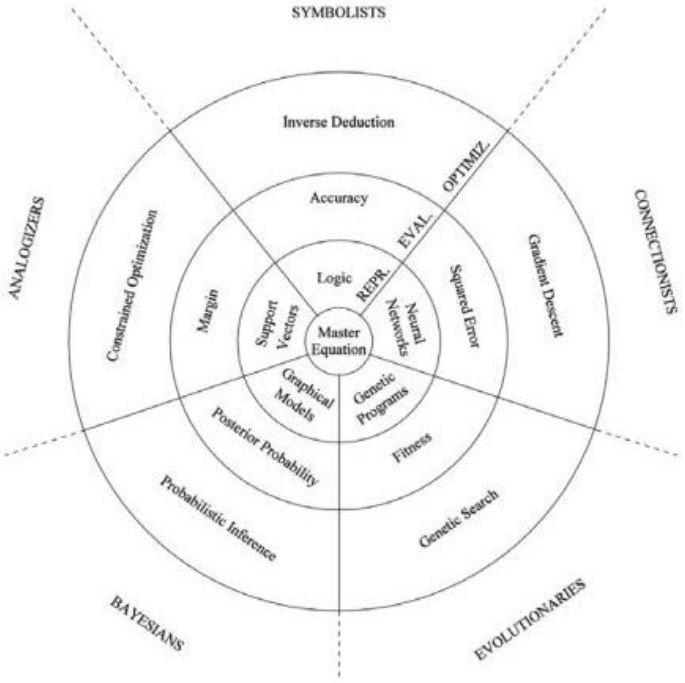

Figure 13: Algorithm paradigm diagram [10]

This taxonomy based on the algorithm/structure of machine learning models is an interesting form to classify machine learning models that avoids overlaps since every existing model can be classified undoubtedly into one of the categories. However, for machine learning threats, the learning paradigm is possibly more appropriate since machine learning threats are not dependent on the model structure but on the type of model training performed. For example, adversarial attacks are applicable to both SVM and DNN [14] which are separated in this taxonomy, while in learning paradigm taxonomy are not.

2.3 Taxonomies of Threats against AI systems

In this section we describe the main existing AI system threat taxonomies, i.e. taxonomies of threats available in the relevant literature or under work in international standardisation initiatives such as ENISA, NIST, ETSI, etc. All the taxonomies presented are particularly addressing AI since the classified threats are particular to AI system components and assets. Examples of threats are provided when explaining the threat attribute taxonomies, and a more exhaustive description of the landscape of attacks against AI is provided in Section 2.4.

A threat is defined as "*any circumstance or event with the potential to adversely impact an asset through unauthorized access, destruction, disclosure, modification data, and/or denial of service*" [85]. Hence, in the cybersecurity domain, system threats can be defined as any potential security or privacy incident (not caused in purpose) or attack (caused in purpose) against any system asset.

When it comes to threats against AI systems and their assets presented in previous sections, a significant number of relevant efforts are being carried out to understand and classify these threats. The most relevant ones are introduced below.

2.3.1 ENISA Threat Taxonomy

The ENISA Threat taxonomy has been developed by the ENISA Threat Landscape (ETL) Group in the *ENISA Big Data Threat Landscape and Good Practice Guide, Jan 2016* [2]. This taxonomy focuses on cybersecurity threats against ICT assets and builds upon previous ENISA research and reports on the subject. Figure 14 shows an overview of the taxonomy. The taxonomy is very complete

since it also includes threats to physical assets as well as disasters from both natural and human causes.

The ENISA Threat Taxonomy [2] comprises eight high level security threat groups ranging from intentional attacks (physical and logical) to unintentional damage or loss of data or assets. It is interesting to note that malicious activity and abuse comprehend the attacks from adversaries to cause damage on victims' ICT systems. Other important threats in this taxonomy refer to legal issues where non-compliance with GDPR could be enclosed.

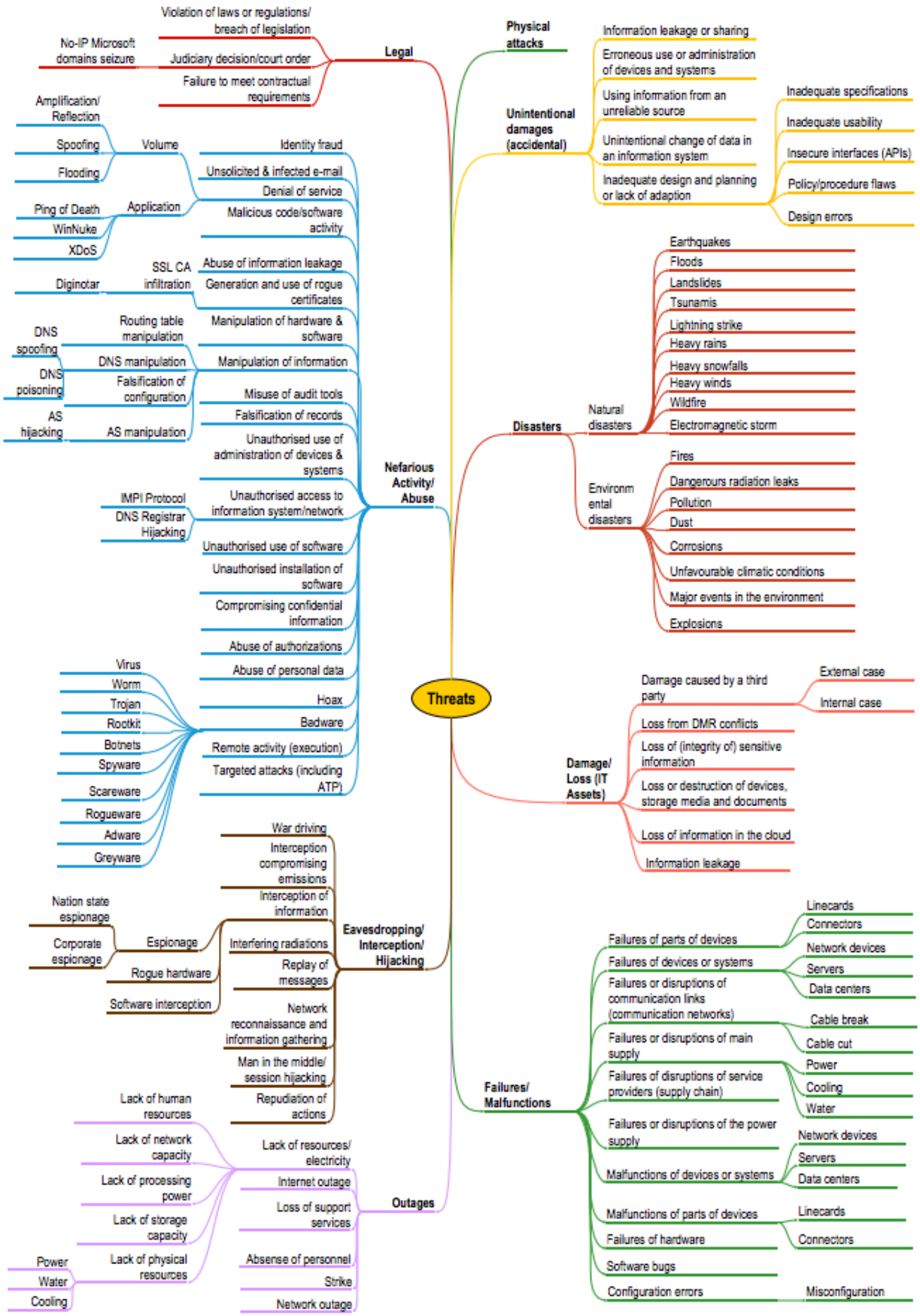

Figure 14: ENISA Threat taxonomy [133]

2.3.2 NIST IR 8269 Draft - AML taxonomy

The *NIST IR 8269 Draft A Taxonomy and Terminology of Adversarial Machine Learning (AML)* [15], from now on NIST IR 8269 AML, report introduces a taxonomy and a glossary of 104 terms of AML. The report focuses on AML techniques in the context of improving the design of the ML algorithms towards making them more robust against these attacks, particularly intentional attacks against the system, with no focus on unintentional system design flaws or data biases. Intentional attacks considered in the taxonomy include adversarial manipulation or tampering with training data, and adversarial exploitation of model sensitivities so as the accuracy in classification and regression are negatively impacted.

The taxonomy of NIST IR 8269 AML includes attacks, defences and consequences in AML, and mainly differentiates attack and defences techniques by ML pipeline phases, i.e. by training and testing (inference) phases of ML operations, as illustrated in Figure 15 and Figure 16 below.

Figure 15: NIST IR 8269 AML attack taxonomy [15]

As it can be seen in Figure 15, the attacks in NIST IR 8269 AML are classified into the following three categories:

- *Targets*: The attack *targets* are classified by NIST by phases in the ML pipeline, including the *Physical Domain* of input sources or sensors, the *Digital Representation* for pre-processing, the *Machine Learning Model* itself, or the *Physical Domain* of output actions.
- *Techniques*: Adversarial techniques may apply to the *Training* or *Testing* (*Inference*) phases of an AI system. Attacks during the *Training* phase aim at acquiring or influencing the training data or the model itself. The objective of attacks during the *Testing* (Inference) phase is to either produce adversarial examples as inputs capable to dodge correct output classification by the model (*Evasion* attacks), or to infer information about the model itself or the training data used to train the model (*Oracle* attacks).

- *Knowledge*: Threats to ML systems may also be classified according to the information the adversary has about the target model. In *Black Box* attacks, the adversary lacks all knowledge about the model except input-output *samples* of training data, or input-output pairings gathered from the use of the target model as an *Oracle*. In *Grey Box* attacks, the adversary has only partial knowledge or information about the model, such as hyperparameter values, training method, or training data. In *White Box attacks*, the adversary has complete knowledge of the model and its characteristics.

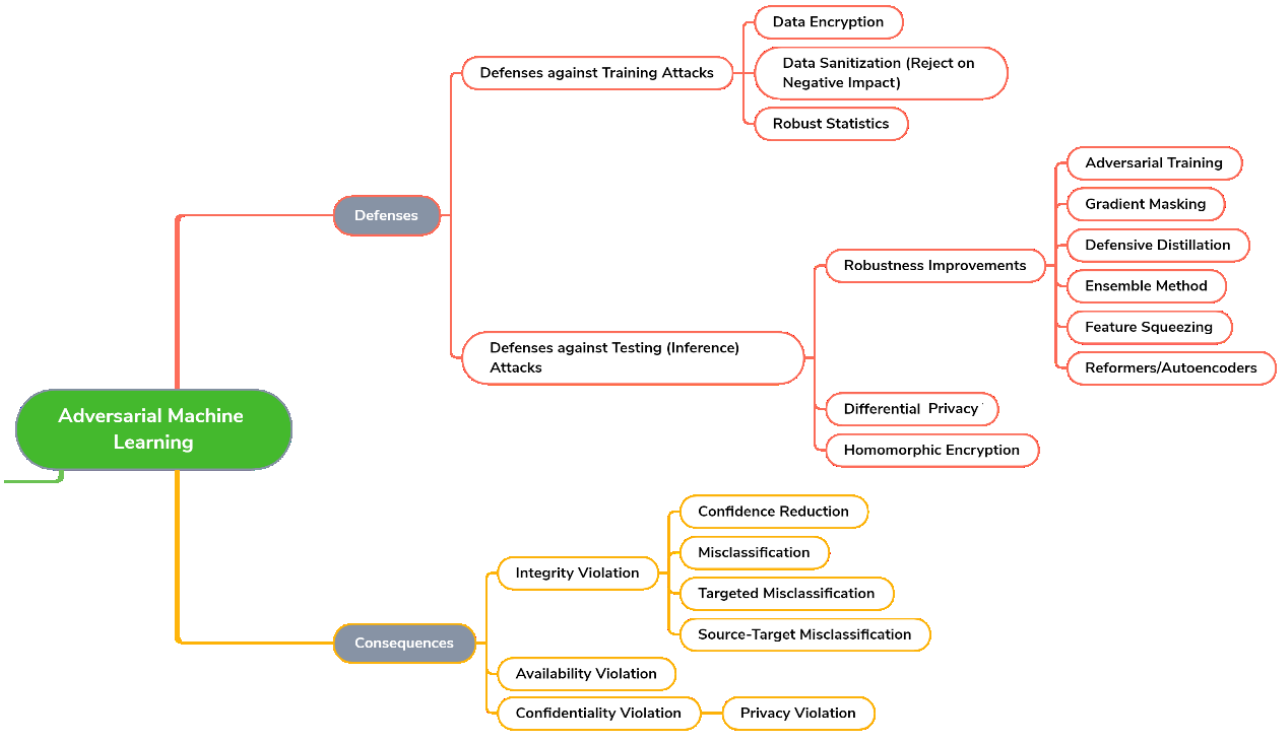

Figure 16: NIST IR 8269 AML defences and consequences taxonomies [15]

The defences in NIST IR 8269 AML are classified into two major categories (see Figure 16):

- *Defences against Training Attacks,* including countermeasures against both Data Access and Poisoning attacks on Training data. Defences against Data Access rely on traditional access control mechanisms such as *Data encryption*. Defences against Poisoning Attacks include *Data Sanitisation,* and *Robust Statistics*.
- *Defences against Testing (Inference) attacks* are usually deployed in the Training phase that precedes Testing (Inference).

The threat consequences in NIST IR 8269 AML are classified into four categories (see Figure 16):

- *Integrity Violations***,** in which the inference process is undermined resulting in ML confidence or classification accuracy reduction.
- *Availability Violations*, which render the model's output or action unavailable or unusable.
- *Confidentiality Violations*, consisting in extraction or inference of information about the model or data. Model information inference attacks are for example *Extraction attacks* or *Oracle attacks*. Data confidentiality attacks include *Inversion attacks* and *Membership Inference attack*s.
- *Privacy Violations,* which are considered a type of *Confidentiality Violation* where the adversary obtains personal information from the training data or the model.

2.3.3 ETSI SAI Threat taxonomy

The recently created ETSI Industry Specification Group (ISG) Securing Artificial Intelligence (SAI) is a new initiative aimed at developing technical specifications to aid multiple industry domains in tackling with threats to AI systems, which may come from the use of AI itself or traditional methods.

The ISG SAI work roadmap includes the development of technical standards to identify and defend from threats to and from AI systems, as well as the use of AI as a means to enhance security. The group planned outcomes include best practices and recommendations to mitigate threats at least in those domains where they may have greatest impact. Tecnalia, SAFAIR partner coordinator of T7.1, is currently contributing to ETSI SAI efforts and will benefit T7.1 with the progress and advances of this group.

The work items of the ETSI ISG SAI are currently being defined, and it is planned that one of them will be the AI Threat Ontology deliverable which is closely related to the SPARTA deliverable D7.1. The ETSI ISG SAI report will seek to align terminology across the different stakeholders and multiple industries. However, the current progress of the AI threat ontology definition is too preliminary to reflect it in this deliverable D7.1.

Therefore, the D7.1 will continuously liaison with the work of the ETSI ISG SAI AI Threat Ontology with the aim to benefit from these results to improve as appropriate if needed the threat model defined in this deliverable.

2.3.4 Q. Liu et al. threat taxonomy

Adversarial goals can be clearly described using both the expected impacts and the attack specificity of security threats. Based on the work by Fredrikson et al. [16], *Q.* Liu et al. [17] suggested a three-axis taxonomy for security threats against ML shown in Figure 17 below.

Therefore, *security threats against machine learning techniques* can be analysed from three different aspects: the threat influence on the ML technique, their impact on the security property affected (integrity, availability and/or privacy) and the attack own specificities.

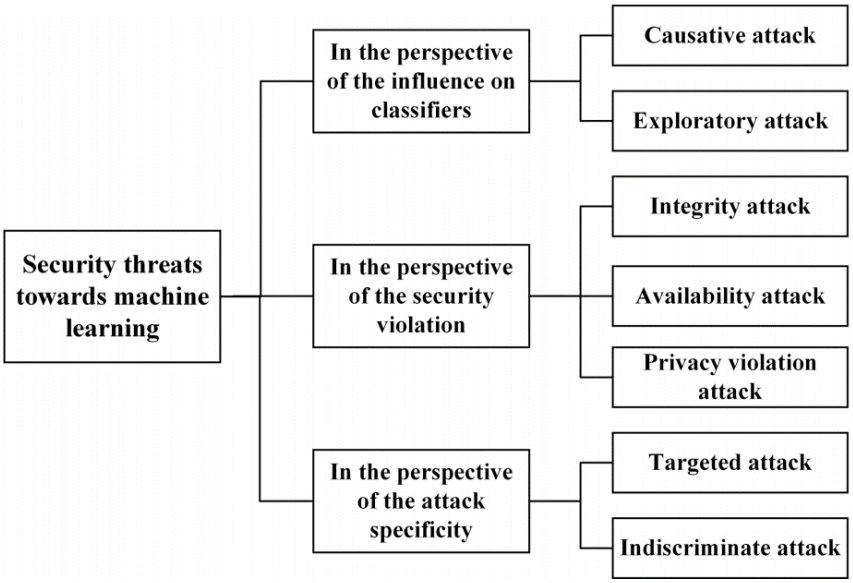

Figure 17: Q. Liu et al. taxonomy of security threats against ML [17]

The authors in [17] also proposed an extended threat and defences classification depicted in Figure 18.

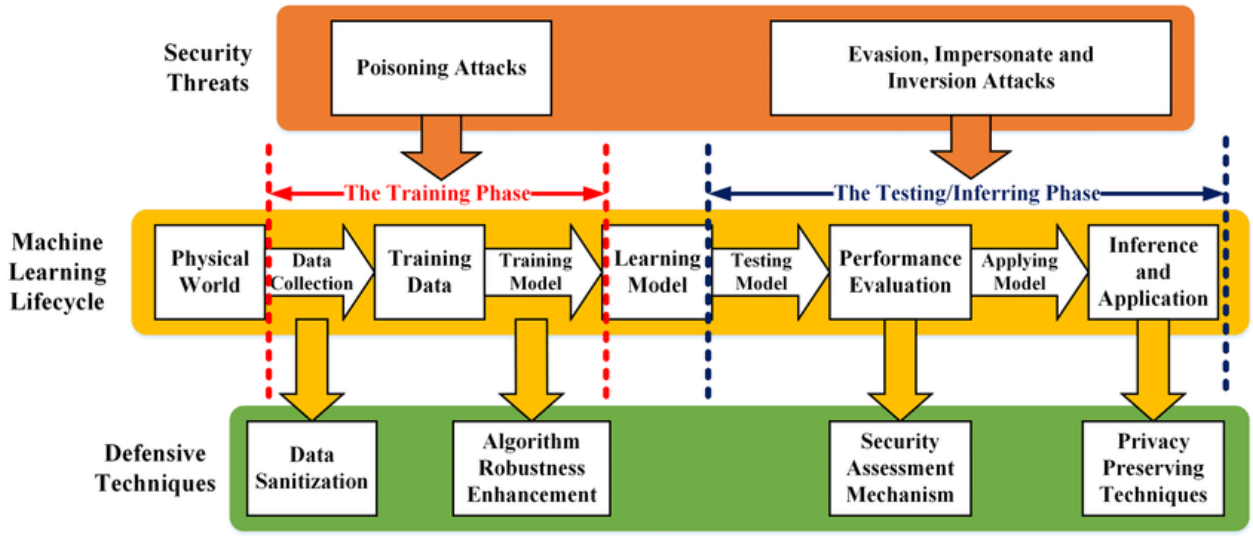

Figure 18: Q. Liu et al. taxonomy of security defences of ML [17]

As it is shown in the picture, current defensive techniques of machine learning are categorized into *data sanitization* and *improvement of algorithm robustness* in the Training phase, being *security assessment mechanisms* and *privacy preserving techniques* applied in the Testing or Inferring phase.

The extended threat taxonomy proposed by Q. Liu et al., discusses the categorical characterisation of the threats against ML according to different perspectives that are explained in the following subsections.

2.3.4.1 Taxonomy based on influence on classifiers

This taxonomy classifies different threats against ML in the perspective of which phase of the machine learning system they attack, that is, dividing the machine learning system in different phases we can classify attacks according to the phase they disturb. A possible division could be the division in Training and Testing phase. This taxonomy divides threats in two groups (see Figure 17):

- *Causative attack*: In this type of attacks, the adversaries have the capability of changing training data inducing parameter change and modifying the performance of machine learning models. One of the most famous examples in this group is the threat named *Poisoning attack*.
- *Exploratory attack*: Instead of focusing in training phase, this type of attacks aims to cause misclassification through the weaknesses developed during the training phase. One of the most studied examples in this group is the *Adversarial attack*.

2.3.4.2 Taxonomy based on security violation

This taxonomy classifies threats against ML according to the type of security property attacked by them. For example, the adversary can try to infer private information through the model. This taxonomy divides threats in three groups (see Figure 17):

- *Integrity attack*: The threats of this group try to achieve an increase of the false negatives in the machine learning model's results.
- *Availability attack*: In this type of attacks, the adversaries try to achieve an increase of the false positives in the machine learning models.
- *Privacy violation attack*: The adversaries that use these attacks aim to obtain private information from one or more individuals from training data and learned models.

2.3.4.3 Taxonomy based on attack specificity

This taxonomy classifies threats against ML focusing on its attack specificity, that is, each threat will be classified according to how many samples it attacks. There are different ways to classify threats this taxonomy, but one of the most common one consists in classifying them in two groups (see Figure 17):

- *Targeted attack*: These attacks are focused on reducing the performance of classifiers for certain samples.
- *Indiscriminate attack*: These attacks cause the classifier or machine learning model to fail in a broad and indiscriminate range of samples.

2.3.5 Taxonomy according to the impacted security property

Auernhammer et al. [14] defined a taxonomy classifying the threats according to the security property which they corrupt. This taxonomy, that is applicable for machine learning threats, includes six groups of threats:

- *Confidentiality attack***:** These attacks compromise the confidentiality of (optionally, private) data processed by the machine learning model, of the results of the model, or even of the model itself.
- *Integrity attack***:** These attacks target to modify or destroy information relative to the data, the machine learning model or the model result, with the aim to corrupt the inference process result and impacting its robustness and reliability.
- *Availability attack***:** The adversaries that use these attacks aim to achieve the parameters of a machine learning model. These parameters could be weights, hyperparameters and so on. Moreover, if the adversaries obtain every parameter, they would have full access to the model.
- *Reliability attack*: These attacks modify the behaviour of the machine learning model in order to corrupt the ability of the ML model to correctly output the intended prediction in its specification, usually with the aim that the model outputs the result desired by the adversary. While all attacks at training phase impact both integrity and reliability, at deployment phase attacks that only impact reliability may succeed, i.e. when the ML components themselves are not changed. [14]
- *Authenticity attack:* These attacks threaten the capability of the machine learning system to be genuine, verifiable and trustworthy. They usually target environmental components surrounding an ML component (such as components implementing cryptographic signatures) which are the ones supporting the authenticity capabilities in the system.
- *Accountability attack*: These attacks target to compromise the capability of the system to trace uniquely the actions that the entities in the system perform and when they are carried out.

For the last two groups of this taxonomy, *Authenticity* and *Accountability*, the authors did not find any attacks which can be classified into these categories yet, even though they expect that in the future adversaries may devise how to perform such attacks.

2.4 Examples of attacks against AI systems

In this section we identify some examples of attacks against AI systems. These attacks have been classified into four groups: *data access* attacks, *poisoning* attacks, *evasion* attacks and *oracle* attacks.

2.4.1 Data Access Attacks

According to NIST IR 8269 AML [15], in Data Access attacks the adversaries get access to all or part of the training data set and use it to create a substitute model which could then be used to check whether crafted inputs would eventually deceive the model prior to using them in attacks in the Testing (Inference) phase [5].

The aim of this type of attack is to gain access to Training data for the purpose of knowing them not changing them, which is the purpose of the poisoning attacks (see Section 2.4.2). In this case the interest is to protect the training data to avoid its later use as an attack vector for other poisoning attacks.

The techniques used to gain access to Training data are similar to other access attacks to datasets in other types of systems, i.e. they are not particular for AI systems. In fact, a traditional access control problem arises during the training phase to protect the confidentiality and privacy of the data and the model, which needs to consider the extent of the adversary's permissions and ability to access the hosting system of the data and the model [31].

2.4.2 Poisoning Attacks

The results of the state-of-the-art analysis in the poisoning attacks' domain are presented below. We have also published this analysis in a peer-reviewed article [135] acknowledging SPARTA.

"Recently it has come to attention that skilfully crafted inputs can affect artificial intelligence algorithms to sway the classification results in the fashion tailored to the adversary needs [18]. This new disturbance in the proliferation of Machine Learning has not yet been extensively researched, and thus the awareness of the challenge is adequately infrequent. At the time of writing this document a variety of vulnerabilities have been uncovered [18].

With the recent spike of interest in the field of securing ML algorithms, a myriad of different attack and defence methods have been discovered, no truly safe system has been developed however, and no genuinely field-proven solutions exist [19].

The solutions known to this point seem to work for certain kinds of attacks, but do not assure safety against all of them. In certain situations, implementing those solutions could lead to the deterioration of ML performance [18].

Poisoning of Support Vector Machines

There are a couple of known poisoning attacks featured in the literature.

In [20] a method utilising the intrinsic properties of *Support Vector Machines* is introduced. The overarching idea is that an adversary can craft a data point that significantly deteriorates the performance of the classifier. The formulation of that data point can be, as demonstrated by the authors, defined as the solution of an optimisation problem with regards to a performance measure. Thus, gradient ascent is used to identify local maxima of the error surface. The authors of [20] introduce a model that analyses label flipping attacks on support vector machines (SVM) in binary classification, which they call adversarial label noise. In their paper, the authors evaluate two major attack strategies – random label flips and adversarial label flips. Random flips are simply an accidental noise, hence the name, which influences a given percentage of data. The second instance features and adversary seeking the maximisation of classification error on testing data. The testing data has not been tampered with. The authors note that the challenge of finding the worst possible mix of label flips is not a straightforward one. The labels that are flipped the earliest are the ones that carry non-uniform probabilities according to the SVM trained on the clear dataset. The classes chosen for the flips are the ones classified with a high confidence, this should result in a significant impact on SVM accuracy.

In [24] the authors investigate poisoning attacks carried by an attacker with full knowledge of the algorithm. The assumption is that the adversary aims to poison the model with the minimum amount of poisoning examples. The attacker function is defined as a bi-level optimisation problem. The

authors notice that this function is similar to machine teaching, where the objective is to have maximum possible influence over the subject by carefully crafting the training dataset. The authors point to the mapping of teacher to attacker and from student to the AI algorithm. The paper thus offers economical solutions to the bi-level optimisation problem present in both fields. Essentially, the authors suggest that, under certain regulatory conditions, the problem can be reduced with the use of Karush-Kuhn-Tucker theorem (KKT) to a single-level constrained optimisation problem. Thus, a formal framework for optimal attacks is introduced, which is then applied in 3 different cases - SVM, Linear Regression and logistic regression.

Manipulation of Naïve Bayes-Based Spam Filters

The authors of [21] express in their paper an illustration of how an attacker could manipulate ML in spam filtering by meddling with the data to either subject the user to an ad or stop the user from receiving genuine communication. The efficiency of those attacks is illustrated. The authors use an algorithm called *SpamBayes* for their research. SpamBayes takes the head and body of a message, tokenizes them and scores the spam to classify it as either spam, ham or unsure. With this established, the paper presents a dictionary attack, in which the algorithm is subjected with an array of spam e-mails containing a set of words that are likely to be present in genuine communication. When those are marked as spam, the algorithm will be more likely to flag legitimate mail as spam.

This particular attack comes in two variations: a procedure where the attack mail contains simply the whole dictionary of the English language, called the 'basic dictionary attack' and a more refined approach, where the attack is performed with the use of a message containing word distribution more alike the users message distribution, along with the colloquialisms, misspellings etc. In this particular case the authors propose a pool of Usenet newsgroup posts. The other evaluated attack is geared towards blocking a specific kind of e-mail - a causative targeted availability attack, or a focused attack. In this scenario the adversary spams the user with messages containing words that are likely to appear in a specific message. With SpamBayes retrained on these messages it is then predisposed to filtering a distinct, genuine communication as spam. This could eliminate a competing bid of a rival company, for example. Including the name of a rival company in spam e-mails, their products or the names of their employees could achieve that objective. The authors indicate that using the dictionary attacks can neglect the feasibility of a spam filter with only 1% of retraining dataset controlled, and a masterfully crafted focused attack can put a specific message in the spam box 90% of the time.

Poisoning of Deep Neural Networks

In [22] the authors investigate a poisoning attack geared towards targeting specific test instances with the ability to fool a labelling authority, which they name 'clean-label' attacks. Their work does not assume knowledge of the training data but does require knowledge of the model. It is an optimisation-based attack for both the transfer-learning and end-to-end DNN training cases. The overall procedure of the attacks, called by the authors 'Poison Frogs' is as follows: the basic version of this attack starts with choosing the target datapoint, then making alterations to that datapoint to make it seem like the base class. A poison crafted that way is then inserted into the dataset. The objective is met if the target datapoint is classified as the base class at test time. Arriving at a poisonous datapoint to be inserted into the training set comes as a result of a process called 'feature collision'. It is a process that exploits the nonlinear complexity of the function propagating the input through the second to-last layer of the neural network to find a datapoint which 'collides' with the target datapoint but is also close to the base class in the feature space. This allows the poisoned datapoint to bypass the scrutiny of any labelling authority, and also remain in the target class distribution. The optimisation is performed with a forward-backward-splitting iterative procedure.

The work described by [23] evaluates a poisoning procedure geared towards poisoning multi-class gradient-decent based classifiers. To this end the authors utilise the recently proposed back-gradient optimization. Back-gradient optimisation offers a lighter and more reliable way to arrive at the solution to the optimisation problem of poisoning attacks. Borrowed from energy-based models and hyperparameter optimisation, this approach allows for a replacement of one of the optimisation problems with a set of iterations of updating the parameters. The authors introduce an attack procedure to poison deep neural networks taking into consideration the weight updates, rather than

training a surrogate model trained on deep feature representations. They demonstrate the method on a convolutional neural network (CNN) trained on the well-known MNIST digit dataset, a task which requires the optimisation of over 450000 parameters. They find that deep networks seem more resilient to poisoning attacks than regular ML algorithms, at least in conditions of poisoning under 1% of the data. The authors also conduct a transferability experiment in which they conclude that poisons crafted against linear regression (LR) algorithm are ineffective against a CNN, and poisons crafted against a CNN have a similar effect on LR as random label flips. A more comprehensive assessment of the effects of poisoning attacks crafted against deep neural networks with the use of the back-gradient algorithm is necessary.

In [25] the idea of watermarking, also known as the DAWN attack, has been investigated to use data poisoning as a defence against model extraction attacks. The algorithm embeds a watermark into a subset of queries it receives to poison surrogate models and allow for claiming the ownership of intellectual property in case of model extraction.

In [26] the authors propose a way of bypassing the gradient calculation by partially utilising the concept of a Generative Adversarial Network (GAN). In this approach an autoencoder is applied to craft the poisoned datapoints, with the loss function deciding the rewards. The data is fed to a neural network, and the gradients are sent back to the generator. The effectiveness of their method is tested thoroughly on the well-known MNIST and CIFAR-10 datasets. The chosen architecture is a two-layer feed forward neural network with recognition accuracy of 96.82% on the MNIST dataset, and for CIFAR-10 a convolutional neural network with two convolutional layers and two fully connected layers, with the accuracy of 71.2%. For demonstrative purposes, one poisoned datapoint is injected at a time. The authors conclude that the generative attack method shows improvement over the direct gradient methods and stipulate that it is viable for attacking deep learning and its datasets, although more research is required.

A targeted backdoor attack is proposed in [27]. The premise of the method is to create a backdoor to an authentication system based on artificial intelligence, allowing the adversary to pass the authentication process by deceiving it. The poisoning datapoints are created specifically to force an algorithm to classify a specific instance as a label of the attackers' choice. The authors propose a method that works with relatively small poison samples and with the adversary possessing no knowledge of the algorithm utilised. This claim is backed up by a demonstration of how inserting just 50 samples gets a 90% success rate.

Proposed Evaluation Frameworks

In [28] a framework for the evaluation of security of feature selection algorithms is proposed. The framework follows the outline defined by [29] in which the authors evaluate the attacker's goal, the extent of the adversary's knowledge of the workings of the algorithm, and their capability in data manipulation. The goal of the malicious user is either targeted or indiscriminate, and it aims to infringe on one or more of the well-known infosec triad items: availability, integrity or privacy.

The specific acquaintance of the adversary with the workings of the system can be one of the following [28]:

- knowledge of the training data (partial or full) - a situation where the attacker has access to full or a segment of the data, or is capable of gathering a surrogate dataset from identical distribution in the ideal situation
- knowledge of the feature representation (partial or full) - where the adversary understands what features and in what way are extracted from the dataset
- knowledge of the feature selection algorithm - the adversary can understand the feature selection algorithm, the feature selection criteria or/and the selection subset
- perfect Knowledge (worst case scenario) - the attacker has full knowledge of all the preceding characteristics
- limited Knowledge - a more realistic scenario, where the attacker can possess some knowledge of all the characteristics

With regards to the attackers' capability, in the case of causative (poisoning) attacks, the attacker can usually influence just a subsection of the training set. The adversary has to bear in mind that the labelling process varies in different use cases, with the use of honeypots and anti-virus software giving setting the constraints of the datapoints that have to be crafted in the malware detection example.

The authors evaluate the robustness of 3 widely used feature selection algorithms: *Lasso, Ridge Regression* and the *Elastic Net* against poisoning attacks with regards to the percentage of injected poisoned data points. The results show that poisoning 20% of the data inflates the classification error 10-fold. In addition to the influence on classification, the authors notice that even with a minute amount of poisoning samples the stability index drops to zero. This means that the attacker can influence feature selection.

The topic of defining the attacker model has been thoroughly investigated in [30], where the FAIL model is proposed. The FAIL model describes the knowledge of the adversary on one of four dimensions:

- *Feature knowledge* - what is the extent of the adversary's knowledge of the feature subset
- *Algorithm knowledge* - how much can the adversary understand about the algorithm to craft adversarial attacks
- *Instance knowledge* - what does the adversary know about the training set
- *Leverage* - what can the adversary modify

The authors of [30] then propose the *StingRay* attack and test its performance in scenarios of limited feature knowledge or limited instance knowledge etc. providing the first experimental comparison of the effect constraining these characteristics have on the effectiveness of poisoning attacks. The adversary's behaviour is affected by the extent of the knowledge the agent possesses of the algorithm's architecture. In literature this level of acquaintance is categorized as *black box* and *white box* [31][18].

While white box attacks presuppose full knowledge of the attacked algorithm, black box strikes are performed with no preceding knowledge of the model [31].

Conclusion

Based on the conducted literature review, the attacks could be characterised in the following way: to be effective, the analysed attacks rely on either modification of the feature vector and/or label, or on the insertion of entirely new datapoints [20][21][22][23][24][25][26][27][28]. This presupposes the knowledge of the feature space and the knowledge of the training data distribution. All the investigated methods need to input or modify multiple datapoints for the poisoning attack to be effective [20][21][22][23][24][25][26][27][28][30], however in a special case (transfer learning) for [22], the attack can be effective with just one inserted datapoint.

Regarding the effect of the attack two subclasses can be described:

- *Non-Targeted Poisoning*, attacks which aim to deteriorate the overall performance of the classifier [20][21][23][24][26][28]
- *Targeted Poisoning*, attacks that aim to control the behaviour of a classifier on one specific test instance [21][22][23][25][27][30].

Additionally, the attacks could also be crafted in a way that would be able to bypass human inspection, these are referred to as the *clean-label attacks* – attacks that causes a model to misclassify a targeted sample after being trained on some corrupted data [22][30].

Most of the attacks assume perfect knowledge [20][21][22][23][24][25][26][27][28] with the exception of *StingRay* [30], which is tested in multiple scenarios and [28]. Finally, the attack scenario presented in [21] could work for other algorithms, but the one investigated in the paper is specifically *SpamBayes*." (Cited from our own paper [135] resulting from the work in SPARTA).

2.4.3 Evasion Attacks

The most common attack to machine learning systems occurs during inference stage and is called evasion. The goal of *evasion attacks* is to modify a single malicious sample causing it to be misclassified as legitimate, remaining stealthy or mimicking some desirable behaviour [32]. For instance, a malicious agent can compromise a system under the supervision of an intrusion detection system (IDS) by encoding network packet payloads in such a way that it evades the system defences [33].

To achieve such goals, different degrees of knowledge about the targeted classifier are required. Attackers may have limited or perfect knowledge about the training dataset, the features, and the classification algorithm [32][34][35]. In evasion attacks, attackers can only modify malicious samples while preserving their intrusive functionality. For instance, spam emails have to evade the defences but also remain readable by humans.

This behaviour has been formalized in terms of application-dependent constraints [34][36]; in particular, in terms of bounds on the distance in feature space between an initial sample and the malicious one created through the original sample manipulation. Bounding the distance between an initial sample and its malicious counterpart yields a *sparse attack* where the cost depends on the number of the modified features. For instance, in case of text inputs where a feature represents the occurrences of a given term, the objective is to change as few words as possible. Instead, *dense attacks* are those where the cost of modifying features is proportional to the Euclidean distance between the original and the modified sample. For example, when considering images, attackers prefer making small changes to many or even all pixels, rather than significantly modifying only few of them. As a consequence, the final effect is less visible to the human's eyes since the image is only slightly blurred. Generally speaking, the procedure for obfuscating malicious data to evade detection is an optimization problem.

In [33] the authors do not distinguish between Evasion and Oracle attacks. Therefore, we have reanalysed their contributions in two aspects: 1) removing references to Oracle attacks, and 2) reorganising them according to the machine learning algorithm attacked.

Below, the evasion attacks suffered by the most popular machine learning techniques are presented.

Naive Bayes (NB)

Naïve Bayes classifiers are a family of probabilistic classifiers that apply Bayes' theorem and strongly assume the independence of features. In this area, the Dalvi et al.'s approach [38] was that adversaries deploy optimal feature-changing strategies against spam classifiers following a two-player game strategy. The objective is to cause the classifier to misclassify malicious samples by modifying them during training. This is a white-box evasion attack since the adversary updates the approach according to the countermeasures that improves the classifier.

Huang et al. [39] focused on the spam and network anomaly detection domains. In their approach, adversaries work against classification and clustering mechanisms by launching black-box to white-box attacks. Moreover, their spam filtering attacks can be considered as poisoning attacks.

Naveiro et al. [40] proposed the Adversarial Classification Risk Analysis (ACRA). In this paper, the attackers evade the defences of a spam classifier by adapting the emails to the classifier's responses through a decision-making process. Therefore, they presented a grey-box evasion attack.

Linear Classifiers (LC)

Linear classifiers make the classification decisions based on a linear combination of the input characteristics. Unfortunately, linear classifiers can be attacked in several ways. For instance, Lowd and Meek [41] defined an Adversarial Classifier Reverse Engineering (ACRE) model where membership queries are sent to a classifier with the aim of misclassifying spam messages as benign. While Demontis et al. [42] focused on several linear classifiers applied to handwritten digit classification, spam filtering and PDF malware detection, where an adversary can launch grey-box to white-box attacks by modifying specific malicious samples.

Support Vector Machine (SVM)

SVMs are used for regression analysis and classification. Under this topic, Zhou et al. [43] provided grey-box and white-box attack models for spam email and credit card fraud detection systems. The restrained attack model proved the existence of a trade-off between disguising malicious and retaining their malicious utility.

Biggio et al. [34] studied evasion attacks on a realistic application for PDF malware detection. They proved that SVMs and neural networks algorithms can be evaded with high probability even if a copy of the classifier is learnt from a small surrogate dataset. The same team [35] assumed that adversaries launch white-box attacks against classifiers (Support Vector Machine - Logistic Regression) with an arms race game. Such evasion attacks were performed over spam email detection, biometric authentication and network intrusion detection applications. In addition, Biggio et al. [44] provided one and a half class classifier in order to achieve a good trade-off between classification accuracy and security. According to their attack model, an adversary can launch grey-box to white-box evasion attacks.

Bhagoji et al. [45] researched on image recognition and human activity recognition to use linear transformation as data defence against evasion attacks. They assumed that attackers could launch white-box attacks over a Support Vector Machine / Deep Learning classifier.

Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbour (K-NN)

Hayes and Danezis [46] researched on launching black-box attacks with adversarial examples against RF, SVM, and K-NN models hosted in an Amazon instance.

Neural networks (NN)

Artificial Neural Networks are bio-inspired algorithms that have been widely applied to image classification. In this area, Carlini and Wagner [37] studied three types of threat methods based on perfect (to perform white-box attacks), limited (to address grey-box attacks) and zero (in case of black-box attacks) knowledge of the adversaries on the AI system. Then, depending on their knowledge, a malicious agent is able to carry out the attack at training or testing phase.

Deep Learning (DL)

Deep learning are artificial neural networks that learn from large amounts of data. These techniques have been widely researched in the literature. For instance, Goodfellow et al. [47] proposed the “adversarial nets” based on a minimax two-player game to create a white-box evasion attack.

Nguyen et al. [48] proved that an adversary launches black-box to grey-box attacks by using evolutionary algorithms. Where an attacker can confuse a DNN classifier by retraining it using fake images.

Papernot et al. [49] highlighted the importance of adversarial sample transferability in case of image recognition made by DNNs. While remaining correctly classified by a human observer, black-box attacks use adversarial samples to mislead a model and any other model under its influence.

Evtimov et al. [50] researched on the Robust Physical Perturbation attack to DNNs used for real road sign recognition. In this attack model, the adversary can launch white-box evasion attacks by using physical perturbations to force their misclassification.

Papernot et al. [51] focused on evasion attacks to Recurrent Neural Networks (RNNs) by crafting adversarial sequences. In this approach, the model’s architecture, computational graph and hyperparameters are learnt at training phase while grey-box attacks are carried out in the test phase.

Radford et al. [52] created the deep convolutional generative adversarial networks (DCGANs) where evasion and poisoning attacks can be carried out following grey-box and white-box approaches. Image classifications were their area of interest.

Demetrio et al. [53] proposed a black-box evasion attack against the MalConv (a CNN classifier) [54]. There, the integrated gradients technique was applied to detect malware.

Gonzalez et al. [23], studied an algorithm based on back-gradient optimization, where an adversary launches grey-box to white-box attacks with either poisoning or evasion attacks to classification algorithms.

Madry et al. [55] worked on the robustness of neural networks under poisoning and evasion attacks. They tested both attacks against classification networks and defences to counter them, and proved that networks that are robust against Projected Gradient Descent (PGD) adversarial techniques are also robust against a wide range of white-box and black-box attacks. Their work used MNIST and CIFAR10 datasets.

In the work of Schottle et al. [56], the attacker launches grey-box attacks by using PGD to create adversarial examples and optimize the misclassification of benign samples.

And Dong et al. [57] proposed both indiscriminate and targeted black-box attacks against deep neural networks. They crafted adversarial examples based on momentum iterative gradient techniques to carry out evasion attacks that increase the false positive rate of robust classification models. In this work, white-box and black-box attacks were carried out over the ImageNet database.

Reinforcement Learning (RL)

In reinforcement learning software agents perform their tasks accordingly to their context and trying to maximize the cumulative reward. Uther et al. [58] evaluated RL algorithms with a framework consisting of a two-player hexagonal grid soccer. In this model, the attackers launch grey-box attacks against the classifier.

2.4.4 Oracle Attacks

In Oracle Attacks the adversary's aim is to infer hyperparameters or characteristics of the target model or data. To this aim, the adversary uses the Application Programming Interface (API) of the target model to make predictions and analyse model outputs [59]. Thanks to the transferability principle, the input-output pairs obtained often serve then to train a substitute model that replicates the original model behaviour, which in turn, allows to generate adversarial examples (Evasion Attacks).

Oracle Attacks include *Extraction Attacks* (extraction of model parameters, structure or class probabilities), Inversion *Attacks* (inference of characteristics that allow to reconstruct training data), and *Membership Inference Attacks* (which learn whether particular data points belong or not to the same distribution as the training dataset).

Extraction Attacks

According to [60][61], a model extraction attack tries to copy a machine learning model through its APIs, without any prior knowledge. Not only the confidentiality of the model is compromised, but the aim can also be to use the new similar model for further attacks [62]. In this case, attackers have limited access regarding victim's model knowledge (in terms of architecture and hyperparameters) and training dataset. In addition, attackers may be blocked under high frequency of query submissions.

Regarding the typical workflow, attackers firstly use the target model to make predictions on input data. Then, input-output pairs and different approaches are applied to extract confidential data such as parameters [61], hyperparameters [63], architectures [64], decision boundaries [62][65], and functionality [66][67] of the model.

In essence, there are three approaches to extract models:

- *Equation Solving*. As its name suggests, given enough input samples to a classification model, these attacks try to recover the parameters of its class probability function [61]. Equation solving is only suitable for small scale models.
- *Training Metamodel*. By querying a classification model, an attacker trains a metamodel that maps input features to their corresponding class. Then, such metamodel can be used to

predict model features from outputs of specific queries. The metamodel can help infer more internal information of the model [64].

- *Training Substitute Model*. Under this attack, a substitute model mimics the behaviour of a target model. Since attackers with enough pairs of input data and their corresponding outputs are able to train a substitute model, substitute model is useful under complex models.

Regarding extraction of hyperparameters, Tramèr et al. [61] were able to steal the hyperparameters of logistic regression, MLP and decision tree models by the application of equation solving techniques. The attack was performed against the BigML and Amazon Machine Learning platforms. On the other hand, Wang et al. [63] estimated the hyperparameters of a model, knowing the machine learning technique and the training data. The goal was achieved by performing many queries to a model and then extracting the corresponding underlain linear equations.

Concerning extraction of architectures, model attributes (such as architecture, operation time and training data size) were stolen by a trained metamodel in [64]. To do so, the authors queried the model via APIs.

Regarding decision boundaries between two classes, attack black models were performed by, firstly, stealing decision boundaries and, then, generating transferable adversarial samples [62][65][49]. Papernot et al. produced synthetic samples with Jacobian based Dataset Augmentation (JbDA) method [62]. Such synthetic samples fell in the nearest boundary between the current class and all the rest of the classes. Later on, this method was extended and named as Jb-topk by Juuti et al. [65]. They produced transferable targeted adversarial samples because of the samples moved to the k nearest boundaries. On the other hand, Papernot et al. [49] proved that is was unnecessary to know in advance the model architecture since a more complex model can extract a simpler model.

Finally, a method for crafting adversarial samples with a substitute DNN was presented in [69]. The substitute DNN model used synthetic queries made directly to a classifier and then, the corresponding answers allowed to craft the adversarial samples.

Inversion Attacks

A model inversion attack reveals training data (or at least some data properties) or confidence coefficients by making predictions with a model. These attacks can be executed following a black-box or white-box approach. In case of white-box attacks, the attackers can easily obtain a substitute model that behaves similarly to the original model. However, adversaries can always make queries with particular inputs and get the corresponding class probabilities. Neural networks are prone to this type of attacks [70].

Property Inference Attacks (PIA) speculate whether a specific statistical property is present in a training dataset. The approach consists of training shadow models with training datasets with or without a certain property.

In [71][72], shadow models are built to provide training data for a meta-classifier. On the one hand, Ateniese et al. [71] trained a meta-classifier that inferred whether a dataset contained a certain property based on the original model features. As this method did not succeed with DNNs, Ganju et al. [72] developed a similar meta-classifier able to extract features of DNNs. On the other hand, Melis et al. [73] trained and updated continuously a binary classifier able to infer dataset properties.

Fredrikson et al. [16] researched on model inversion attacks that use machine learning APIs to infer sensitive features in decision tree models and facial recognition services. They proved that attackers could recover face images from models based on neural network such as softmax regression, multilayer perceptron and stacked denoising autoencoder. However, in [75], the inputs extracted from the training dataset are an average representation of the inputs belonging to the same class.

There are several strategies to fight against model inversion attacks. The main countermeasure is to expose less information to attackers. Therefore, the access should be limited to a black-box access, and the model's output should be also limited. A good option is to report only rounded confidence values like Fredrikson et al. [16] proposed.

Membership-Inference Attacks (MIA)

In membership-inference attacks, an adversary has black-box access to a model, as well as specific training samples. The aim is to learn whether a sample belongs to the training dataset. The adversary conducts such attacks by comparing the predictions on samples belonging to the training dataset against other data samples. Truex et al. [75] presented a systematic formulation of MIA.

There are several ways to conduct a membership-inference attack. Attackers can use heuristic methods to determine the membership of a record by querying the target model and calculating prediction probabilities [76][77][16][78][79]. But this approach is not generalizable since in order to reliably learn probability vectors or binary results some preconditions and auxiliary information is needed. Another approach is to train an attack model where the training data is obtained by paring input queries and their related responses [80][71]. In the inference phase, attackers first predict the class of a sample with the target model and, then, they get the membership of this sample with the attack model. On the other hand, some works have studied shadow models to provide training data for the attack model [81][76] due to the limitation of queries and model features.

Lowd and Meek [41] proposed an Adversarial Classifier Reverse Engineering (ACRE) model. In ACRE, the adversary launched grey-box evasion attacks against a spam classifier by sending data membership queries to it. The goal of the attacker is to determine whether a specific instance is classified as malicious.

Pyrgelis et al. [80] used a distinguishability game process to train a classifier (attack model) to determine whether samples are in the target dataset. They implemented MIA for aggregating location data.

Only knowing the probability vector of outputs from a target model, Salem et al. [76] used a statistical measurement method to compare whether the maximum classification probability exceeds a certain value.

Long et al. [77] proposed a generalized MIA method. They trained some reference models that mimicked the target model in order to choose vulnerable data before Softmax. They compute the probability that the data belongs to the training dataset by comparing the outputs of the target model and the reference models. No attack model is needed in this approach.

Shokri et al. [81] researched on how classification ML models leak information about individual samples of the training dataset. They trained some shadow models that mimicked the target's behaviour, aiming to verify whether some data samples belonged to the training dataset.

Hayes et al. [83] were able to determine whether a dataset belonged to the target training dataset according to the probability vector given by a classifier, based on the idea that the higher the probability the more likely is it belongs to the target training set. The classifier was constructed by the target model in case of white-box approach. However, in case of black-box attacks, they created a classifier through GAN with data gathered from querying the target model. The discriminator model of the GAN framework can be leveraged similarly to the way in which the Shokri et al.'s shadow models [81] do. Knowing the training dataset, the attacker can identify samples that have been likely trained on taking in mind the confidence of the discriminator model.

An effective defence against model inversion attacks is generalization by reporting rounded prediction results [81]. On the other hand, the best defence to resist membership inference attacks is differential privacy (DP) [84].

2.5 Threat modelling and threat analysis in AI systems

2.5.1 Key concepts

The following concepts have been defined to ease the understanding of the threat model. The terms are mostly inherited from the ENISA Glossary [85] but other sources of security and risks standards have also been used as indicated.

Table 1: Key Concepts in threat modelling

Term	Description
Agent	Either: (i) *"intent and method targeted at the intentional exploitation of a vulnerability"*; or (ii) *"a situation and method that may accidentally trigger a vulnerability"*. (NIST SP 800-18 Rev. 1) (NIST SP 800-30)
Asset	*"Anything that has value to the organization, its business operations and their continuity, including Information resources that support the organization's mission*." (ISO/IEC PDTR 13335-1)
Threat	*"Any circumstance or event with the potential to adversely impact an asset through unauthorized access, destruction, disclosure, modification of data, and/or denial of service*." (ENISA)
Weakness / Vulnerability	*"The existence of a weakness, design, or implementation error that can lead to an unexpected, undesirable event compromising the security of the computer system, network, application, or protocol involved*." (ITSEC)
Impact	*"The result of an unwanted incident."* (ISO/IEC PDTR 13335-1)
Risk	*"The potential that a given threat will exploit vulnerabilities of an asset or group of assets and thereby cause harm to the organization*." (ISO/IEC PDTR 13335-1)
Safeguards / Countermeasures	*"Practices, procedures or mechanisms that reduce risk.* *The term 'safeguard' is normally considered to be synonymous with the term 'control'. "* (ISO/IEC PDTR 13335-1)
Stakeholder	*"Any individual, group or organization that can affect, be affected by, or perceive itself to be affected by, a risk."* (ISO/IEC Guide 73)

2.5.2 Threat modelling and threat analysis

Threat Modelling is an engineering discipline that supports the identification, rationalisation and reasoning about threats, attacks, vulnerabilities, and countermeasures that could affect an application or system. A threat model is an abstraction of the system under analysis that identifies different kinds of hazards, potential attacks, or weaknesses of the system. Threat models are usually built in machine-readable format and processed by computer programs as part of the system protection decision process. Threat models are often considered as a representation of the software components or devices in a system, the data flows between them and the trust boundaries in the system.

With threat-modelling potential vulnerabilities in the system design can be discovered algorithmically by analysing the system's security properties and identifying potential threats to system assets. The threat modelling can be performed before a product or service has been implemented; this helps ensure that a product or service is as secure as possible by design. The threat modelling activity can support cybersecurity and resilience decision making processes in various ways, being one of the most important the Threat Analysis.

In the Threat modelling report by the Department of Homeland security [86] it is proposed that system threat analysis starts with the analysis and specification of its scope, architecture, and technology components, together with the threat agents and vectors (e.g. attack paths and techniques) to be

examined in the analysis. The study of security perimeters, system interfaces, and data flows enables to evaluate the attack surface. Generally, all threat analysis approaches involve the identification of assets, system boundary mapping and decomposition, threat identification and vulnerability identification [87][88].

In [86] the process of threat modelling consists in choosing a cyber threat modelling framework and then populating it with particular values of system and attack attributes for each of the perspectives included in the models. Once the framework is populated, it shall support the analysis of threat scenarios to aid in devising possible countermeasures or controls, at both system and organisational levels. The populated models can also be used for threat information sharing, internally and with third parties, in order to collaborate in threat intelligence and potential responses. For doing so, different types of threat modelling could be adopted. The work of [89] proposes to differentiate them based on three main characteristics:

1. The logical entity abstracted in the model (data, software, system, etc.),
2. The phase of the system lifecycle for which the security modelling is made, and
3. The objective of the threat modelling.

According to [86], threat modelling is a component of risk management that can be approached from the next three angles:

- *Attack(er)- and Threat-Centric Modelling*: In this approach, the threat modelling starts with identifying potential threat sources and adversaries, and characterizing them, including defining their goals, motivation, means to gain the objective, behaviour, etc. [90].

  Attacker-centric approaches use attack trees to visualize the various pathways by which invaders can compromise the security of the asset [91][92]. Another notable formal approach is attack nets proposed by [93] that aim at improving the expressiveness of attack trees.

- *Software-centric and System-Centric Threat Modelling***:** Software-centric models are built in the software design and development phases to try to cut down the number of vulnerabilities, while system-centric threat models abstract operation systems security [89]. Software-centric modelling focuses on software flaws, vulnerabilities, misuses, etc. Data flow diagrams are a type of model that often supports the analysis of system, data, and boundaries, so as to be able to identify threats over particular components and trust boundaries.

  A particular type of system threat modelling is the *Data-centric system threat modelling* which focus is the protection of particular types of data rather than devices, operating systems or applications. In this analysis authorized locations to store and process data within the system are identified together with data flows between authorized locations and users [89].

- *Asset- and Impact-Centric Modelling***:** An asset-oriented approach tries to identify first the organizational assets that could be potential targets or be impacted by threats and where they fit in the system or organisation processes [94][89]. Then, the characterization of those intentional or unintentional threats follows systems. All the devised threat scenarios over the assets need to be analysed, studying primary assets one by one and considering any other assets that may be used as a channel to harm them [95].

  Analysts using asset-centric approaches mainly use summative ranking of low, medium and high to estimate the level of risk.

These three approaches are illustrated in Figure 19. It should be highlighted that the level of details in threat modelling depends on the depth and scope of the analysis required. Each of the three modelling types focuses on a different aspect of threats and assumes information of the other aspects or concepts to be able to establish the primary aspect scope and characteristics.

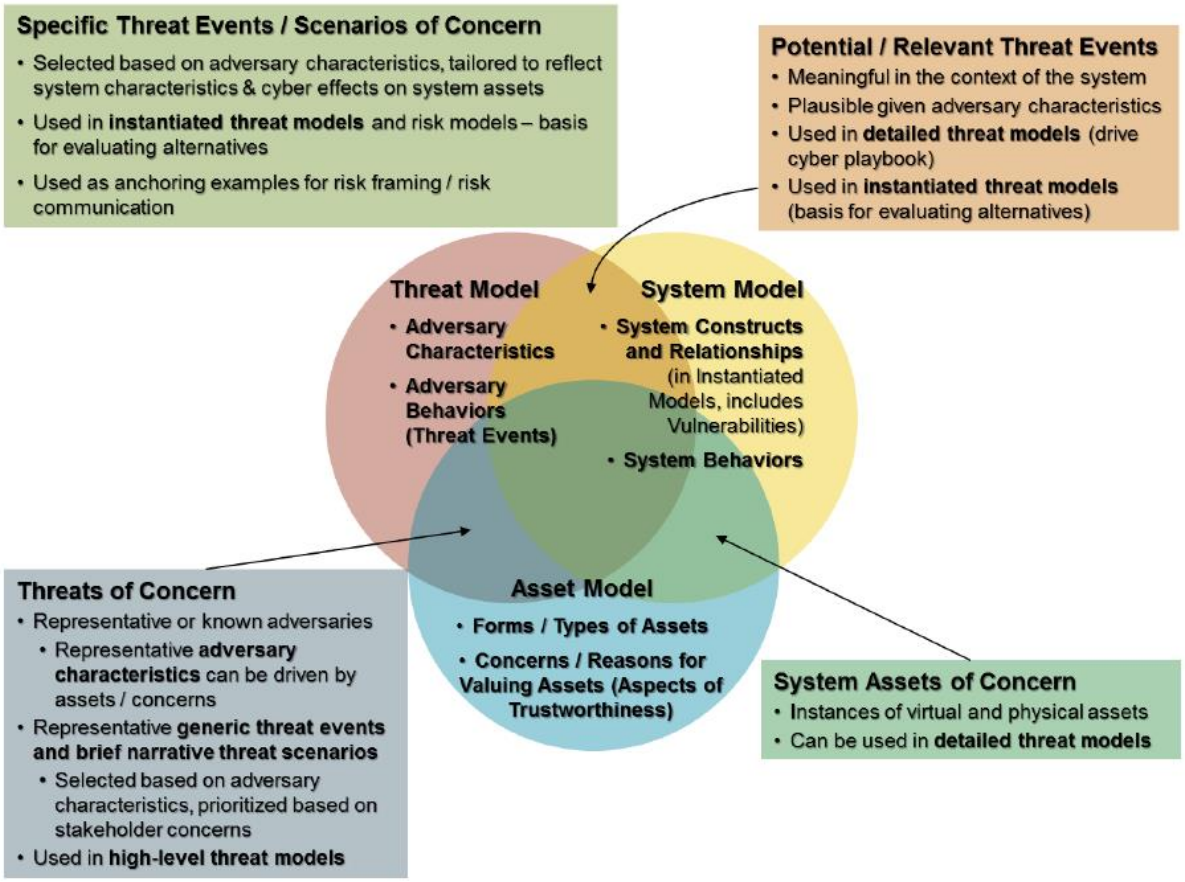

Figure 19: Threat modelling approaches [86]

Different techniques exist to implement the threat models following any of the three approaches above, and threat modelling formalisms range from text or spreadsheet-oriented to graph-oriented. In the last years, attack-oriented threat modelling is gaining adepts, and graph-based techniques such as Direct Acyclic Graphs (DAG) - including Bayesian networks and (extended) Attack Trees – are most commonly used. The main reason for their popularity is likely their capability to support threat modelling for large scale systems and to enable the automatization of the security analysis on top of them. It is not in the scope of this deliverable the review of all threat modelling techniques currently available. Please refer to the comprehensive survey on DAG-based security modelling approaches reported by Kordy et al. in [96].

The SAFAIR threat model presented in Chapter 3 offers support to threat modelling as part of security-by-design activities in AI systems. The objective of the SAFAIR threat model proposed is not to abstract a particular AI system and its potential threats, since it is not intended to support the analysis of a real system, but to capture the different concepts and elements that may be considered when analysing threats to AI systems. Therefore, the SAFAIR threat model collects information on potential attacks and weaknesses of AI systems in general, which can be used to support threat analysis activity of the AI system development when trying to identify the potential issues and attacks that a particular system may suffer.

2.6 Conclusions

Multiple relevant taxonomies addressing big data and AI systems and threat concepts can be found in the state of the art. Their analysis indicates that none of them covers all aspects around threat modelling in AI. For example, AI system taxonomies based on the application of machine learning models seem not to be appropriate for our purposes, since they propose separations of machine

learning models that are not always compatible to map threat and attacks to them. This is for example the case of Golstein [8] taxonomy in Section 2.2.2.1 which separates Regression type models and Neural Network models, while actually, Neural Networks can be considered a succession of different Regressions and some attacks such as inversion attacks [14] can be implemented in both types of models. Thus, other taxonomies that follow the paradigm of classifying the systems according to the learning method seem preferably. As a consequence, the AI Threat model in SAFAIR will be built as a combination of different taxonomies that allow for best mapping of AI threats to AI systems.

The ENISA threat taxonomy [2] is not specifically dedicated to AI but to big data systems in general. This taxonomy is one of the pioneers and its major strength resides in that it embraces multiple types and aspects of threats with a wider perspective, and not only addressing attacks but also unintentional flaws or other issues around the system such as compliance. Therefore, the threat types and threat agent classification in SAFAIR threat model were inherited from ENISA taxonomy which generalises to any type of big data system.

The NIST IR 8269 AML is currently under development and is well-known and referenced in literature. This report focuses on adversarial machine learning attacks and does not include asset classification. The report describing the comments to the NIST IR 8269 AML provided by the Carnellie Mellon University's Software Engineering Institute (SEI) [5] clarifies and details the process of AI system training, deployment and operation. From this report we borrowed the AI system asset classification since it is richer and more specific to ML systems than ENISA's in [2].

The ETSI SAI threat taxonomy is also expected to be a very relevant taxonomy of reference, but its development has been initiated in 2020 and no public report is available yet. SAFAIR partner Tecnalia is contributing to this effort and working on trying to align SAFAIR results with it.

With regards to the protections or safeguards against threats in AI which are part of the threat modelling practice, please note that the study of the countering of AI threats falls under deliverable D7.2 of SAFAIR and no classification of protections is provided in SAFAIR threat model, since the definition of AI security mechanisms in SAFAIR and their evaluation is still under work.

Chapter 3 SAFAIR Threat model for AI systems and supporting tool

3.1 Introduction

In this section, we describe the SAFAIR support to the collection and formalisation of information related to threats against AI systems. SAFAIR proposes the use of a comprehensive AI threat model developed for this purpose that enables to capture in a structured way all the knowledge about the different aspects of potential threats associated to AI systems.

The SAFAIR threat model for AI systems is the result of a thorough stocktaking of available initiatives, guidelines and taxonomies dealing with the difficult subject of organising and clarifying the landscape of AI threats. Some of these works embrace both aspects of the relationships between cybersecurity and AI, that is, i) the cybersecurity of AI systems and ii) the use of AI as a means to protect systems from security and privacy threats. However, the approach adopted in the SAFAIR threat model is limited to the first one of these perspectives and we only consider the threats that particularly affect AI systems and the artefacts used or created along the AI system creation and operation activities.

In the following sections, we describe first the methodology and principles that guided the creation of the SAFAIR AI Threat model (Section 3.2), and then we provide the description of the model itself (Section 3.3). Following, in Section 3.4, the technical details of the Knowledge Base created are presented. This tool adopts the Threat model and offers the initial content of the body of knowledge of AI threats developed in SAFAIR.

Finally, this chapter ends (Section 3.5) with the illustration of how an organisation could benefit from the use of the SAFAIR AI Threat model through the use of the supporting Knowledge Base tool as part of the AI system life-cycle process.

3.2 Design principles and methodology

The first step for designing the AI threat modelling method in SAFAIR was the thorough analysis of existing taxonomies of AI systems and their components, together with the stocktaking of taxonomies for threats against these systems. Then, existing threat modelling and analysis techniques were studied from the angle of whether they could be applicable to the case of AI threats and under which considerations. A summary of this information was reported in Chapter 2.

Then a comprehensive study of publications about incidents and attacks against AI systems was performed by combining inputs from different public sources such as surveys, mainly [14][15][17][18] and multiple references provided therein, and other relevant examples of attacks from the literature.

As a result of the analysis of all this information, the design of the SAFAIR AI Threat Model was carried out trying to adhere as much as possible to the following principles:

- The attributes selected to capture the AI threat information should enable structuring it, so as the model can support organisations in a pragmatic way and serve as the cornerstone for a reference body of knowledge.
- The values for AI threat attributes in the model should reflect standard classifications as much as possible.
- The initial body of knowledge content should be as complete as possible with currently available information.
- The body of knowledge should be extensible in the future with information of additional attack techniques and attack examples that may appear in the real world or in the literature as plausible attacks.

- The body of knowledge is initially focused on nefarious activities and abuse behaviour from adversaries, though the model architecture should be complete enough to embrace other types of threats such as incidents caused by unintentional software or hardware flaws, issues or imperfections in organisational procedures, and legal threats such as non-compliance aspects.
- As the knowledge about the threats that needs to be captured in the model needs to serve organisations in improving the cybersecurity of AI systems, the model shall include information related to how to prevent or protect the AI system against the threats instances identified. Therefore, the countermeasure or defence information shall also be part of the threat information.

3.3 AI Threat model in SAFAIR

The following Figure 20 illustrates the AI Threat domain model developed within SAFAIR, which links the main concepts around threats against AI systems. The main goal of the model is to aid in the identification and analysis of potential threats against AI systems.

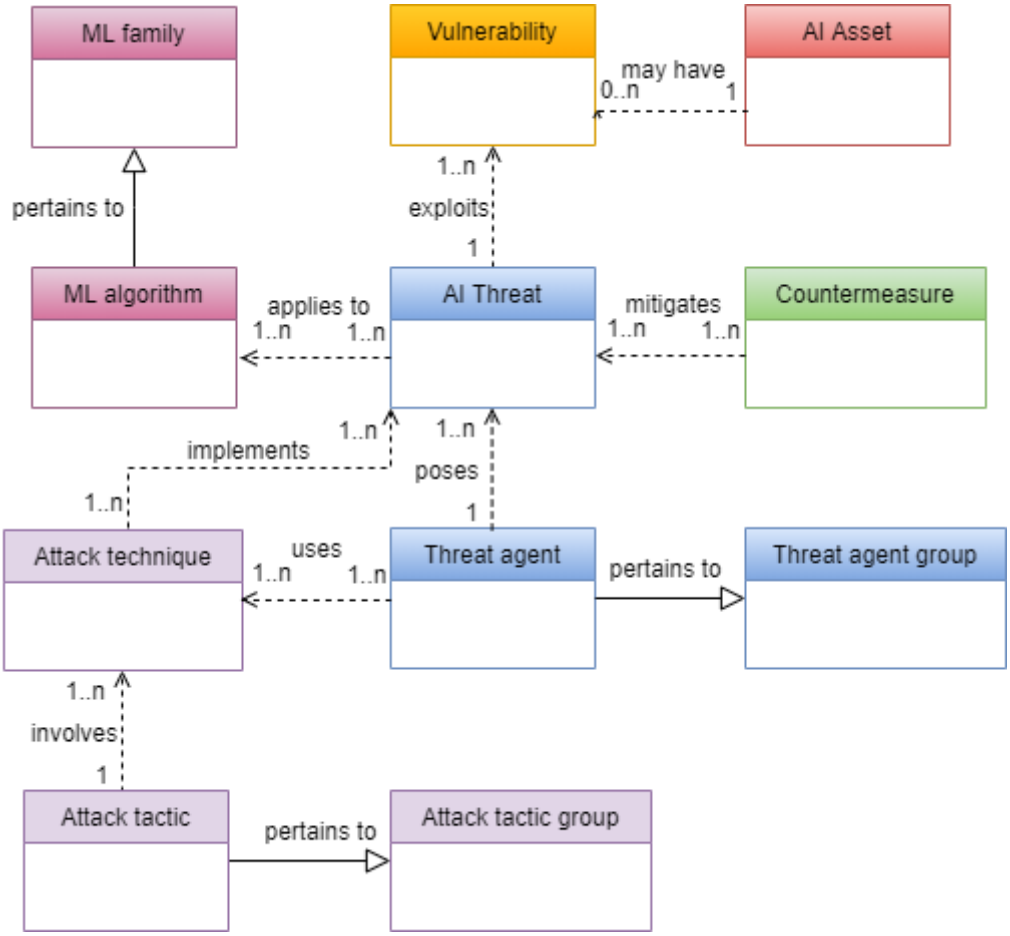

Figure 20: SAFAIR AI Threat model

In the following subsections we describe in detail these concepts captured in the SAFAIR AI Threat model. Please note that even if initially the focus of the SAFAIR AI Threat model is to capture knowledge of ML-based system threats, we will keep the more generic term “AI algorithm”, “AI threat” and “AI asset” to name the concepts. The main reason for this is that it is expected that in the future the AI Threat Knowledge base is extended with information about threats against other intelligent systems that mimic human behaviour beyond the ML-based systems, such as Natural Language Processing (NLP), fuzzy logic, etc.

3.3.1 Threat group or type

For the identification of the Threat groups, we have considered the Threat group taxonomy in the *ENISA Big Data Threat Landscape and Good Practice Guide, Jan 2016* [2].

The following threat groups have been included in the SAFAIR AI Threat model, since they represent a high-level classification of the types of threats in any ICT system, and therefore they will allow us to easily allocate the AI threats researched so far and most likely any new incoming attack technique that may get documented in the literature in the future:

- *Nefarious Activity / Abuse*.
- *Unintentional damage / loss of information or IT assets*.
- *Eavesdropping / Interception/ Hijacking*.
- *Legal*.
- *Organisational*.

The body of knowledge in SAFAIR is initially focused on technical threats covering *nefarious* activities and *abuse behaviour* from adversaries, since all the attack techniques surveyed in the literature and the ones that SAFAIR will develop are intentional attacks rather than accidental or unintentional losses.

Furthermore, it is expected that GDPR compliance issues fit within *Legal* threats. Other organisational issues on engineering processes, policies and value chain around the AI system will fit in the *Organisational* type of threats.

3.3.2 Target AI asset taxonomy

For the identification of the Target AI assets, we have considered the Software Engineering Institute's Comments to NIST IR 8269 AML [5] and complemented with other concepts that are used in the literature when explaining potential attacks to AI systems such as *raw data*.

The target AI assets are classified in the following categories in the SAFAIR AI Threat model:

- *Raw data*: This is the data collected from sensors and data sources in data capturing phase before it is started to be processed.
- *Training data*: This refers to the data used to train the model that is resulting from the pre-processing of raw data.
- *Test data – Training – Model Training*: This is the test data used in the model Training phase, named "Test Data 1" by SEI in [5] (see Section 2.2.1.4).
- *Test data – Training – Model Deployment*: This is the test data used in the model Deployment phase, named "Test Data 2" by SEI in [5] (see Section 2.2.1.4).
- *Trained model*: This refers to the function and hyperparameters or for short parameters of the ML model and the configuration of the model once it is trained.
- *Operational data*: This is the data for which the deployed model generates the inference result. These data are collected in operation in the data capturing phase. The concept herein refers to data coming from sensors and other input data sources of the system in operation, regardless it is input raw data or already pre-processed input data.
- *Benchmark data*: This is the dataset that is used to validate that the system is functioning in operation as intended.
- *Model testing tool*: This refers to tools that are used in testing phase to evaluate whether the Trained model is working well.

- *Runtime Model monitoring tool***:** This refers to tools that are used in operation phase to continuously monitor that the model is behaving as expected.
- *Inference result*: The outcome of the application of the machine learning algorithm to the Operational Data may also be subject to attacks.

3.3.3 AI algorithm taxonomy

As explained before, the focus of SAFAIR in the analysis of AI threats is basically on threats targeting Machine Learning systems. In this section we describe the Machine Learning types classified in SAFAIR by extending the state-of-the-art taxonomies related to learning types (see Section 2.2). This taxonomy will be the one used in the proposed threat modelling methods by SAFAIR (see Section 3.3.4).

For the identification of this taxonomy, we have considered the CSA Big Data Taxonomy [4] that has also been extended with some algorithms from the literature where attacks were identified (see examples in Section 2.4). The following categories of Machine Learning Types, ML Algorithm Families and ML Algorithms have been included in the SAFAIR AI Threat model:

- *Unsupervised Learning*: There are two main ML Algorithm Families herein:
  - *Clustering:*
    - *K-means Clustering*.
    - *Spectral Clustering*.
    - *Hierarchical Clustering*.
    - *Expectation-Maximization (EM)*.
    - *Gaussian mixtures*.
  - *Dimensionality Reduction:*
    - *Principal Component Analysis (PCA)*.
    - *Linear Discriminant Analysis (LDA)*.
    - *T-Distributed Stochastic Neighbour Embedding (T-SNE).*
- *Supervised Learning* split in two ML Algorithm Families called Regression and Classification.
  - *Regression*:
    - *Linear regression*.
    - *MARS*.
    - *Logistic regression*.
  - *Classification***:**
    - *Neural Networks (NN)*.
    - *Bayesian Network.*
    - *Support Vector Machine (SVM)*.
    - *Maximum Entropy*.
    - *Decision Trees*.
    - *Conditional Random Fields (CRF)*.
    - *Random Forests*.
- *Semi-Supervised Learning* gathers the main four types:
  - *Generative Models.*

- *Graph-Based Models.*
- *Self-training.*
- *Multi-View Models.*

- *Reinforcement Learning*, subdivided in two ML Algorithm Families as follows.
  - *Markov*:
    - *Iterative Value*.
    - *Iterative Policy*.
    - *Q-learning*.
    - *SARSA algorithm*.
  - *Evolution*:
    - *Learning Classifiers*.
    - *Stochastic Gradient*.
    - *Genetic Algorithm.*
    - *Deep reinforcement learning (DRL).*

For a complete description of the algorithms, we refer the reader to the corresponding literature on threats against these algorithms that were classified in SAFAIR, particularly to Table 3 in Section 3.4.2 below.

3.3.4 AI attack technique taxonomy

For the identification of the AI attack technique taxonomy, we have considered two major works Auernhammer et al. [14] and the NIST IR 8269 AML [15] and other adversarial machine learning attacks surveys such as [17] and [18]. Some discrepancies were found between these studies with respect to how they classify and name attacks techniques. Particularly, in some cases there is not a clear distinction between *poisoning* and *evasion* attacks, as the ultimate goal of many poisoning attacks is to evade the proper inference of the machine learning algorithm in operation.

Aiming at trying to be as complete as possible and differentiating the techniques as much as possible, the following categories of attack techniques have been included in the SAFAIR AI Threat model:

- *Adversarial label flips*
- *Ateniese et al.*
- *Backdoor poisoning*
- *Basic Iterative Method*
- *Carlini and Wanger*
- *Deblurring*
- *Dictionary*
- *Enchanting*
- *Equation-Solving*
- *Fast Gradient Sign Method (FGSM)*
- *Feature Collision*
- *Feature Deletion*

- *Generative poisoning*
- *Hyperparameter Stealing*
- *Lowd-Meek*
- *Model Inversion from confidence values*
- *Obfuscation*
- *Path-Finding*
- *Projected Gradient Descent*
- *Random label noise*
- *Robbery of Model IPR*
- *Side-channel attack*
- *Strategically-timed attack*
- *The worst-case label noise*
- *Training Data Extraction*
- *Transferable clean-label*
- *Trojan Trigger*
- *Watermarks*

Please note that the description of each technique is provided in the corresponding references mapped in Table 3 in Section 3.4.2.

3.3.5 AI attack tactic taxonomy

For the identification of the AI attack tactic taxonomy, we have taken into account the NIST IR 8269 AML [5]. The following 4 main categories of Attack Tactic Groups have been identified in the SAFAIR Threat model where main Attack Tactics have been mapped:

- *Data Access*: In these attacks, some or all of the data points in the training dataset are accessed by the adversaries who can use them to create a replica or substitute model, which will be used to prepare attacks in the model Testing (Inference) phase.
- *Poisoning*: In Poisoning Attacks, the data or model are altered indirectly or directly.
    - *Indirect poisoning*: In this case, adversaries poison the data before pre-processing since they do not have direct access to the pre-processed data.
    - *Direct poisoning - Data Injection*: In this case, adversarial data points are injected in the original training data, which changes the data distribution, but still the labels or features of the original training data remain.
    - *Direct poisoning - Data Manipulation - Label Manipulation*: It involves adversarial modification of output labels of the original training data.
    - *Direct poisoning - Data Manipulation - Input Manipulation*: It involves adversarial modification of input data of the original training data.
    - *Direct poisoning - Logic Corruption*: Logic Corruption is achieved when the adversary tampers with the algorithm and thus modifies the learning process of the model and thus the model itself.
- *Evasion*: In Evasion Attacks, the adversary solves a constrained optimisation problem to find a perturbation in the input (Operational Data) that causes the desired misclassification.

  - *Gradient-based - Single Step*: It typically involves gradient-based search algorithms.
  - *Gradient-based – Iterative*: It is based on iterative gradient-based search algorithms.
  - *Gradient-free*: It is not based on gradient-based search algorithms, but typically requires access to model's prediction confidence values.
- *Oracle*: In this case, an adversary uses the system Application Programming Interface (API) to enter inputs for the model and to observe the model's predictions so as they can learn the oracle of the model.
  - *Oracle – Extraction*: The adversaries are able to extract the parameters or structure of the model from model's prediction outputs and estimations of probabilities of the classes.
  - *Oracle – Inversion*: In this case, the adversary tries to infer or reconstruct the training data, which may include private information of individuals, posing privacy threats.
  - *Oracle - Membership Inference*: The adversaries query the target model to identify data points belonging to the same distribution as the training dataset, that is, learning the model's confidence on entered data points which may or may not be part of the training dataset.

As it can be seen above, in *poisoning* attacks that manipulate the data, we have differentiated between tampering with labels in Training data or the Training data itself. In *oracle* attacks we have also made a distinction between reverse engineering of the model itself (*Extraction*), the inference of the training data (*Inversion*) and the inference of whether data points pertain to the Training Data (*Membership Inference*). Please refer to Section 2.4 for complete description of Attack Tactic Groups and Attack Tactics.

3.3.6 Threat agents and their knowledge

For the classification of the threat agent types, we have taken into account the Agent taxonomy in *ENISA Big Data Threat Landscape and Good Practice Guide, Jan 2016* [2]. The following categories of threat agents have been included in the SAFAIR AI Threat model which will keep the definition by [2]:

- *Corporations* or organisations.
- *Cyber criminals*.
- *Cyber terrorists.*
- *Script kiddies* representing unskilled individuals.
- *Online social hackers* who are individuals with political or social motivations.
- *Employees* who refer to the staff or contractors of an organisation.
- *Nation states*.

This ENISA classification of threat agents is valid for multiple types of ICT systems including AI systems. Therefore, it can be easily used for the techniques and attack types studied in SAFAIR from the literature, where in most cases the article authors refer to "adversaries", "attackers" or "opponent" as performers of the adversarial machine learning techniques, and they do not actually give details about them, so it is difficult to establish which of the roles above they fit better.

For the identification of the threat agents' knowledge, we have taken into account the NIST IR 8269 AML [5] rather than ENISA taxonomy [2], since [5] is closer to threats against adversarial systems. Starting from the NIST taxonomy that distinguishes between black-box, grey-box and white-box knowledge of the adversaries, the following categories of threat agents' knowledge have been created in the SAFAIR AI Threat model, which detail more the specific knowledge about the AI system that the adversary may have:

- *Black-box – Samples:* The adversary has no knowledge about the model under attack but has insights of Training samples used in model training such as probabilities of sample labels.
- *Black-box – Oracle:* The adversary has no knowledge about the model under attack but usually has query access to the model.
- *Grey-box - Model Architecture:* The adversary has partial knowledge about the model under attack and knows the model type and structure only.
- *Grey-box - Parameters Values:* The adversary has partial knowledge about the model under attack and knows the parameters of the model when obtaining results
- *Grey-box - Training Method (Loss Function):* The adversary has partial knowledge about the model under attack and knows the model Training method.
- *Grey-box - Training Data:* The adversary has partial knowledge about the model under attack and knows the Training Data set used to train the model.
- *White-box or Perfect knowledge:* The adversary has full knowledge of the model, including its architecture, the hyperparameter values, the model Training method, and in some cases the Training Data set too.

It is interesting to note that the degree of knowledge about the AI system together with the specific system attributes that need to be known by the adversaries to carry out the attacks is a source of root cause analysis and understanding of what type of countermeasures could be adopted to prevent the attacks and issues in AI. Furthermore, the knowledge necessary to perpetrate the attack may also give hints on the transferability property of the attack, and whether it could be effective or not against a wider target.

3.4 AI Threat Knowledge Base

This section describes the Knowledge Base of AI Threats created in SAFAIR to support the collection, structuring and reuse of knowledge about threats against AI systems.

3.4.1 Knowledge Base creation methodology

The SAFAIR Knowledge Base on AI threats has been crafted upon the structure and design of the SAFAIR AI Threat model described in previous section 3.3. The Knowledge Base has been implemented as a MySQL repository with tables and relationships that conform to the taxonomy and concepts of the Threat model defined.

The Knowledge Base has been initially populated with the information and analysis results of the stocktaking of existing literature and works on AI threats against AI, as collected in Chapter 2. The objective is to extend the body of knowledge in the future by adding understanding and knowledge gained from the work on SAFAIR solutions for AI, particularly:

- Task *T7.2 Design defensive and reactive security mechanisms* is expected to enhance the knowledge on the understanding of how particular attacks work and which protections could be implemented in the AI system to counter them.
- Task *T7.3 Enhance explainability of AI* will define means to overcome issues and threats derived from the lack of explainability of the system.
- Task *T7.4 Design measures for fairness in AI* will deliver measures to counter threats related to potential bias in AI systems.

All these results will be validated in task T7.5 “Testing and validation” through demonstrators of SAFAIR solutions. The validation will serve to confirm the effectiveness of the protections and solutions proposed and refine them so as the body of knowledge is accurate.

3.4.2 Initial Knowledge Base contents

The AI threat knowledge initially captured is focused on malicious activities and abuse behaviour of adversaries, i.e. attacks against AI systems, rather than unintentional weaknesses or incidents. This initial core, extracted from [14][15][17][18], is intended to be extended in the future with information on incidents or unintentional threats as well as legal threats such as non-compliance aspects.

With regards to attacks information, we have collected those attacks that have been documented in the literature or in experiments. It is important to note that, when it is indicated that an attack applies to a certain ML family or method it is because at least one reference was found that proves it. However, it could be the case that in the future the threat utility against another ML method is documented. Therefore, this statement does not preclude the attack from being applicable to other ML methods that were not identified yet.

Table 2 shows the extract of the main threat attributes of the initial contents of the Knowledge base with respect to only attacks information.

Table 3 shows the mapping of AI attacks studied in SAFAIR to most relevant ML techniques as defined in Section 3.3.4, and all the references are provided.

Table 2: Initial Knowledge Base contents – AI Attack Techniques

Attack Technique	Target Asset	ML Phase	Attack Tactic Group	Attack Tactic	ML Algorithm family
Adversarial label flips attack	Training data	Training	Poisoning	Direct poisoning - Data Manipulation - Label Manipulation	Classification
Ateniese et al.	Training Data	Testing	Oracle	Oracle - Inversion	Classification
Backdoor poisoning	Training Data	Training	Poisoning	Direct poisoning - Data Manipulation - Input Manipulation	Classification
Basic Iterative Method	Operational Data	Testing	Evasion	Gradient-based - Iterative	Classification
Carlini and Wanger	Operational Data	Testing	Evasion	Gradient-based - Iterative	Classification
Deblurring	Operational Data	Testing	Oracle	Oracle - Inversion	Classification
Dictionary attack	Training data	Training	Poisoning	Direct poisoning - Data Injection	Classification
Enchanting	Operational Data	Testing	Evasion	Gradient-based - Iterative	Markov
Equation-Solving	Trained Model	Testing	Oracle	Oracle - Extraction	Classification
Fast Gradient Sign Method (FGSM)	Operational Data	Testing	Evasion	Gradient-based - Single Step	Classification
Feature Collision	Training Data	Training	Poisoning	Direct poisoning - Data Manipulation - Label Manipulation	Classification
Feature Deletion	Operational Data	Testing	Evasion	Gradient-free	Classification
Generative poisoning	Training Data	Training	Poisoning	Direct poisoning - Data Injection	Classification
Hyperparameter Stealing	Trained Model	Testing	Oracle	Oracle - Extraction	Classification
Lowd-Meek	Trained Model	Testing	Oracle	Oracle - Membership Inference	Classification

Attack Technique	Target Asset	ML Phase	Attack Tactic Group	Attack Tactic	ML Algorithm family
Model Inversion from confidence values	Trained Model	Testing	Oracle	Oracle - Inversion	Classification
Obfuscation	Operational Data	Testing	Evasion	Gradient-free	Clustering
Path-Finding	Trained Model	Testing	Oracle	Oracle - Extraction	Classification
Projected Gradient Descent	Prediction	Testing	Evasion	Gradient-based - Iterative	Classification
Random label noise attack	Training Data	Training	Poisoning	Direct poisoning - Data Manipulation - Label Manipulation	Classification
Robbery of Model IPR	Trained Model IPR	Testing	Oracle	Oracle - Extraction	Classification
Side-channel Attack	Trained Model Engine	Testing	Oracle	Oracle - Extraction	Classification
Strategically-timed Attack	Operational Data	Testing	Evasion	Gradient-free	
The worst-case label noise attack	Training Data	Training	Poisoning	Direct poisoning - Data Manipulation - Label Manipulation	Classification
Training Data Extraction	Training Data	Training	Data Access	Data Access	Dimensionality reduction, Classification
Transferable clean-label	Training Data	Training	Poisoning	Direct poisoning - Data Manipulation - Label Manipulation	Classification
Trojan Trigger	Training Data	Training	Poisoning	Direct poisoning - Data Manipulation - Input Manipulation	Classification
Watermark	Operational Data	Testing	Evasion	Direct poisoning - Data Manipulation - Input Manipulation	Classification

Table 3: Mapping between Attack techniques and ML algorithms

Machine Learning & Threats			Adversarial label flips attack	Ateniese et al.	Backdoor poisoning	Basic Iterative Method	Carlini and Wanger	Deblurring	Dictionary attack	Enchanting	Equation Solving	Fast Gradient Sign Method	Feature Collision	Feature Deletion	Generative poisoning	Hyperparameter Stealing	Lowd Meek	Model Inversion from	Obfuscation	Path Finding	Projected Gradient Descent	Random label noise attack	Robbery of Model IPR	Side-channel Attack	Strategically-timed Attack	The worst-case label noise	Training Data Extraction	Transferable clean-label	Trojan Trigger	Watermark
Unsupervised	Clustering	K-means Clustering		[71]																									
		Spectral Clustering																											
		Hierarchical Clustering																[100] [127]											
		Expectation-Maximization (EM)																											
		Gaussian mixtures																											
	Dimensionality Reduction	PCA																											
		LDA																								[31]			
		t-SNE																											
Supervised	Regression	Linear Regression									[120] [61]						[41] [61]	[116]											
		MARS									[120] [61]																			
		Logistic Regression		[71]							[120] [61]	[109]		[108]		[63]	[41] [82]				[34] [55] [82] [106] [118]	[82] [102]								

Machine Learning & Threats			Adversarial label flips attack	Ateniese et al.	Backdoor poisoning	Basic Iterative Method	Carlini and Wanger	Deblurring	Dictionary attack	Enchanting	Equation Solving	Fast Gradient Sign Method	Feature Collision	Feature Deletion	Generative poisoning	Hyperparameter Stealing	Lowd Meek	Model Inversion from	Obfuscation	Path Finding	Projected Gradient Descent	Random label noise attack	Robbery of Model IPR	Side-channel Attack	Strategically-timed Attack	The worst-case label noise	Training Data Extraction	Transferable clean-label	Trojan Trigger	Watermark
	Classification	NN		[71]	[27]	[99] [118]	[119] [118] [82] [121] [103] [104]	[117]		[122] [113]	[120] [61]	[109] [118] [125] [114] [104] [113]	[22]	[108]	[26]	[63]		[16] [82] [116]			[34] [82]		[107]	[115]	[122]		[132]	[124]	[82] [123]	[105]
		Bayesian Network		[71]					[21]					[108]			[41] [82]				[34] [82]						[31]			
		SVM	[68] [102] [111] [112]	[71]							[120]			[108]		[63]	[41] [61] [82]				[34] [82]	[102]	[34] [82]			[112]				
		Maximum Entropy		[71]										[108]			[41] [82]				[34] [82]									
		Decision trees		[71]										[108]				[16] [82] [116]		[61]	[34] [82]									
		CRF		[71]										[108]							[34] [82]									
		Random Forests		[71]										[108]				[16] [82] [116]		[61]	[34] [82]									
Reinforcement	Markov	Iterative Value																							[122]					
		Iterative Policy																							[122]					
		Q-learning								[122] [113]															[122]					

Machine Learning & Threats		Adversarial label flips attack	Ateniese et al.	Backdoor poisoning	Basic Iterative Method	Carlini and Wanger	Deblurring	Dictionary attack	Enchanting	Equation Solving	Fast Gradient Sign Method	Feature Collision	Feature Deletion	Generative poisoning	Hyperparameter Stealing	Lowd Meek	Model Inversion from	Obfuscation	Path Finding	Projected Gradient Descent	Random label noise attack	Robbery of Model IPR	Side-channel Attack	Strategically-timed Attack	The worst-case label noise	Training Data Extraction	Transferable clean-label	Trojan Trigger	Watermark
	SARSA																							[122]					
Evolution	Learning Classifiers																							[122]					
	Stochastic Gradient																							[122]					
	Genetic Algorithm																							[122]					
	DRL																												

3.4.3 Implementation of the Knowledge Base tool

Figure 21 shows the schema of the SAFAIR AI Threat Knowledge Base tool, which has been designed as a MySQL database.

The Knowledge Base implements the SAFAIR threat domain model depicted in Figure 20, and details each of the concepts implemented and their relationships.

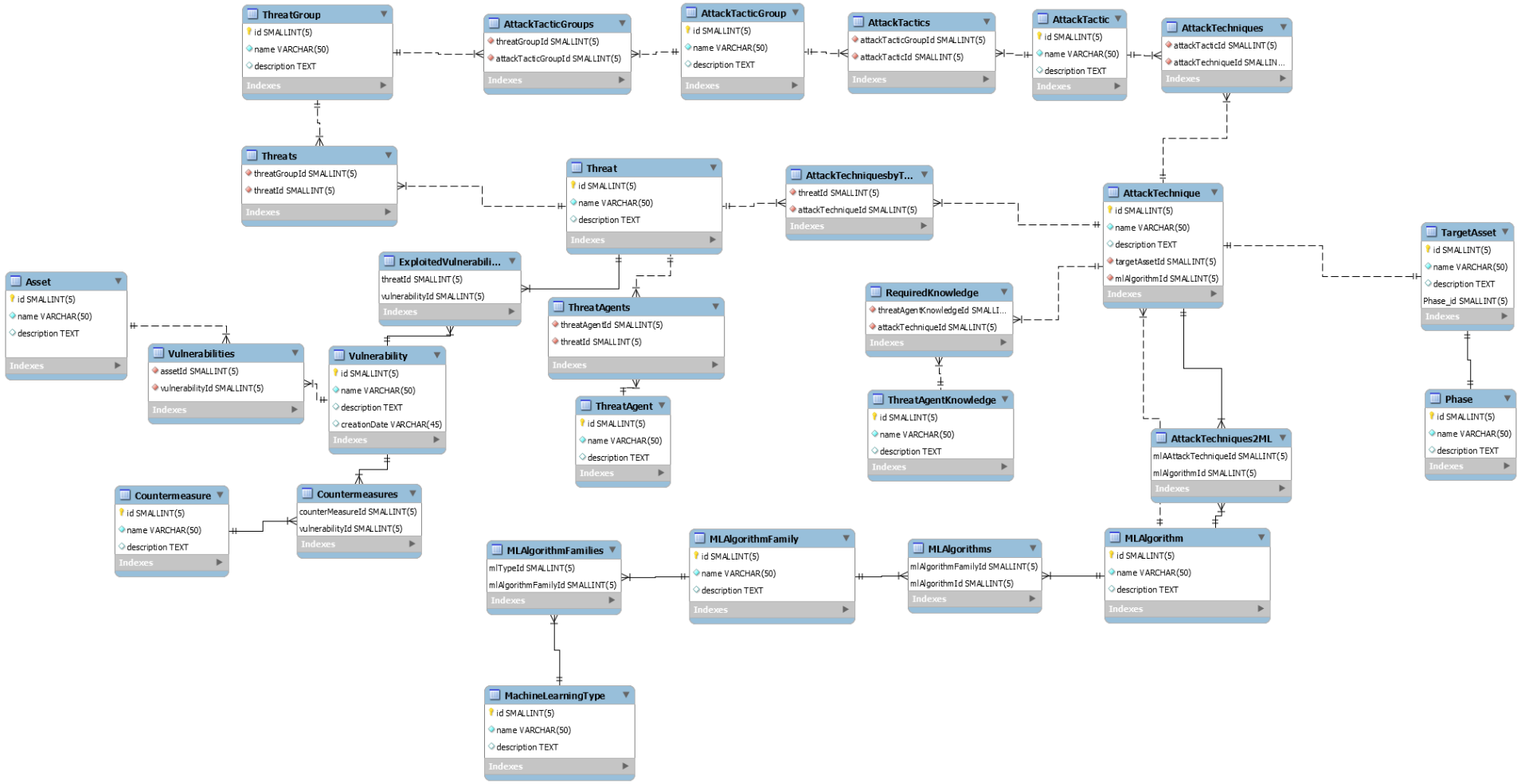

Figure 21: Schema of the SAFAIR AI Threat Knowledge Base tool

3.5 AI Threat Analysis methodology in SAFAIR

This section explains how the AI Threat model defined in SAFAIR and the Knowledge Base created on top of it could be used to support the AI Threat Analysis activity of the AI system design and creation process.

The SAFAIR Knowledge Base on AI threats is structured following the AI Threat model described in previous sections and aims at supporting organisations in enhancing their cyber threat intelligence with regards to artificial intelligence-based systems.

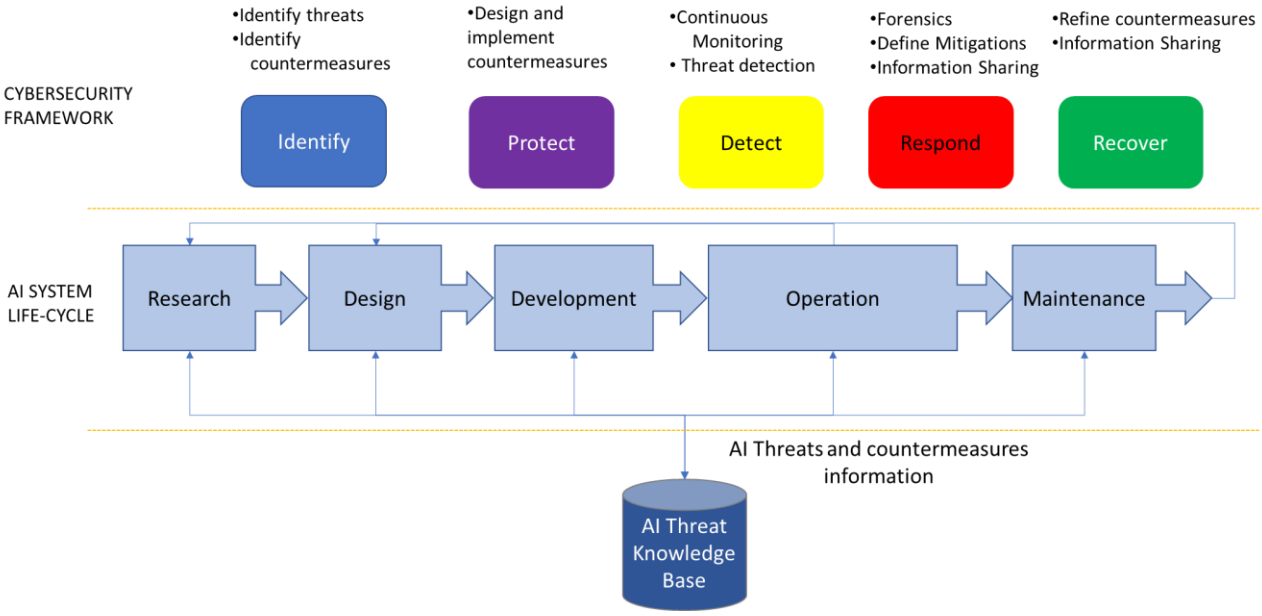

Figure 22: SAFAIR AI Threat Knowledge Base tool in use

As shown in Figure 22, the present deliverable D7.1 together with the AI Threat repository can be used in the life-cycle of AI systems as a means to aid in the following activities:

1) **Research of secure and fair AI systems:** The AI threat Knowledge Base can be useful in the research activities to better understand the state of the art in attacks and defences of AI, as well as in the conceptualisation of the AI-based system when investigating which flaws and threats it may have. The repository collects information from multiple state-of-the-art surveys that researchers and interested stakeholders in cybersecurity of AI systems can benefit from when studying the trustworthiness of the system. The gathered knowledge has been structured so as it is easy to identify references and sources of information about the threats and the countermeasures that have proved to be effective to minimise their effects.

2) **Design of secure and fair AI systems**: Since the AI threat model includes information related to countermeasures and how to prevent or protect the AI system against the threat instances identified, it will serve organisations in improving trustworthiness and other capabilities of AI systems. Particularly, the Knowledge Base can be used as a tool where different types of queries can be made to extract the information therein to help in the following activities:

- **Security-by-design**: an extensive collection of AI attack types such as *Adversarial Machine Learning* attacks have been identified together with the countermeasures applicable to specific types of machine learning algorithms. Therefore, this will allow AI system developers understanding better the techniques used by the adversaries and offer them clues on how to protect from them.
- **Privacy-by-design** and **Privacy-by-default**: the D7.1 includes a thorough study about possible privacy threats in form of challenges and issues that should be considering when building AI systems so as a privacy-respectful processing and storage is ensured.

- **Fairness-by-design**: some of the attack techniques collected in D7.1 and the repository refer to adversarial machine learning targeting to alter the result of the AI system when in operation. This could be done through different means such as tampering with Training data (*poisoning*) or directly with Operational Data (*evasion*). In any case, understanding how to prevent these types of attacks can aid to ensure fairness.
- **Explainability-by-design**: as the knowledge on how to achieve transparency and explainability of AI systems in their inference steps and results increases, the Knowledge base will be enriched with this information.

3) **Development of secure and fair AI systems**: The Knowledge Base will help developers in, first, understanding which attack tactics adversaries could use against the particular AI system under study, and also in identifying potential safeguards against them. As part of system development, in the testing phase the utility and efficiency of the implemented countermeasures could be checked and information of the results could be added to the Knowledge base to evidence and compare these results with the results at system operation.

4) **Operation of secure and fair AI systems**: System operators would act as consumers and providers of information in the Knowledge Base. The information of the Knowledge Base shall be treated as a live body of knowledge that needs to be kept up to date to serve the purpose of deciding effective security and fairness strategy for the AI system under study.

- **Benefiting from Knowledge Base**: During system operation, administrators and operators could consume the knowledge of potential attacks and implement means against those attacks targeting operation phase such as *evasion* attacks. Countermeasure results in testing phase could also serve for them to check whether desired protection levels are being achieved or not and deviations from designed efficiency identified.
- **Enriching Knowledge Base**: The knowledge on whether specific countermeasures implemented in the system are actually performing well could be fed into the repository and enrich it with data from real usage upon the system under study. Refinements on how to improve countermeasures efficiency could also be made. Furthermore, in case of actual incidents and attacks are detected, this information shall be added in the repository.

5) **Maintenance of secure and fair AI systems**: At this phase, the AI threat repository will intervene in supporting the following two tasks:

- **Forensics**: When detected attacks pertain to the group that already have a countermeasure identified, investigations on why the countermeasure did not work properly shall be carried out. And when the attack registered is a new attack not yet identified in the Knowledge Base, it shall be carefully studied and the root cause effects analysed so as to try to identify safeguards against it that need to be tested and registered in the Knowledge Base too.
- **Information sharing of AI systems**: The Knowledge Base could be extended by or integrated with other information sharing tools such as MISP [97] so as the knowledge on suffered attacks and incidents serves the cyber intelligence of other organisations. This way, all participants in the community interested in AI system protection could benefit from common understanding about real attack techniques as well as about means and practices to protect the AI systems from them.

Chapter 4 GDPR Compliance of AI systems

4.1 Introduction

One of the principal goals of the General Data Protection Regulation[1] (hereinafter "GDPR") is to strengthen the protection of the right to the protection of the personal data and privacy[2]. In the same time, the use of automated processing of personal data is increasing. As explained by the Article 29 Working Party (replaced by the European Data Protection Board, hereafter "EDPB"[3]): "*The widespread availability of personal data (…), and the ability to find correlations and create links, can allow aspects of an individual's personality or behaviour, interests and habits to be determined, analysed and predicted*"[4].

Therefore, it is imperative that AI systems are compliant with the Regulation. In particular, the European legislator has strengthened the guarantees that must govern the processing of personal data by a principle of data integrity and confidentiality (Article 5 of the GDPR). In addition, Article 25 GDPR enshrines the data protection obligation by design and by default. On the one hand, the controller of an AI system must, both at the moment of setting up the means of processing and at the moment of the processing itself, apply adequate technical and organisational measures aimed at effectively applying data protection standards and incorporating the required safeguards in the processing. This must intervene as soon as possible in the process of "creation" of any data processing[5] as the principle of privacy by design aims at avoiding gap between the technical work and the data protection rules. On the other hand, the controller must use appropriate technical and organisational measures to guarantee that, by default, only personal data that are necessary for each specific purpose of the processing are processed.

In view of the above, the objective of this chapter is twofold. Firstly, in order to assist in identifying challenges, requirements and potential issues, we present the applicable legal framework for making AI systems compliant with the principles governing data protection set out in the GDPR. Secondly, particular attention is paid to the law requirements regarding privacy threats and data breaches. Indeed, one major challenge at the heart of the Regulation in the adoption of AI system is its security. A data controller could not legally use an artificial intelligence system without the assurance that the integrity and confidentiality of the information would be respected, which lead to increase trust in AI system. Moreover, the regulation now requires notification of data breaches to the supervisory authority or even to the data subjects.

The structure followed in the chapter is as follows. Section provides 4.2 some definitions useful to understand the discussion. Section 4.3 is devoted to the main principles of personal data protection. Section 4.4 examines the concept of data breach. Section 4.5 analyses the division of roles and liabilities between the data controller and the processor. Sections 4.6 analyses the concept of privacy by design and by default and its implications for artificial intelligence systems. Section 4.7 is devoted to the security of personal data. Section 4.8 outlines the obligations in case of a data breach. Finally,

---

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC ("GDPR").

[2] See Communication from the Commission to the European Parliament and the Council, Stronger protection, new opportunities- Commission guidance on the direct application of the General Data Protection Regulation as of 25 May 2018, 24.01.2018, COM(2018) 43 final.

[3] https://edpb.europa.eu/edpb_en

[4] Art. 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 03.10.2017 (revised and adopted on 06.02.2018), WP251 rev. 01, p. 5.

[5] *"When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations"* (recital 78)

Section 4.9 offers the major conclusions of the analysis. Please note that, for the sake of readability and understandability, all the references in this section, including clarifications and references to specific articles, are provided as footnotes.

4.2 Preliminary definitions

In this section some definitions of key terms from GDPR that will aid the reader in the rest of the section are provided.

Data controller: *"the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data"*[6].

Data processor: *"a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller"*[7].

Personal data: *"any information relating to an identified or identifiable natural person"*[8].

Identifiable natural person: *"who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"*[9].

Processing of personal data: *"any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction"*[10].

Pseudonymisation: *"processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person"*[11].

4.3 Key principles to have a lawful processing

The purpose of this section is to present the general principles ensuring the lawfulness of the processing of personal data (see Figure 23). The use of AI systems needs to be accountable and correct according to EU regulations, in particular regarding the GDPR. Moreover, due to the amount of personal data processed in AI mechanisms, a proper implementation of these principles is already a first step for the controller to avoid threats as much as possible. The different steps for the management of personal data in AI can be resumed in the following figure.

[6] Article 4.7 of the GDPR.
[7] Article 4.8 of the GDPR.
[8] Article 4.1 of the GDPR.
[9] Article 4.1 of the GDPR.
[10] Article 4.2 of the GDPR.
[11] Article 4.5 of the GDPR.

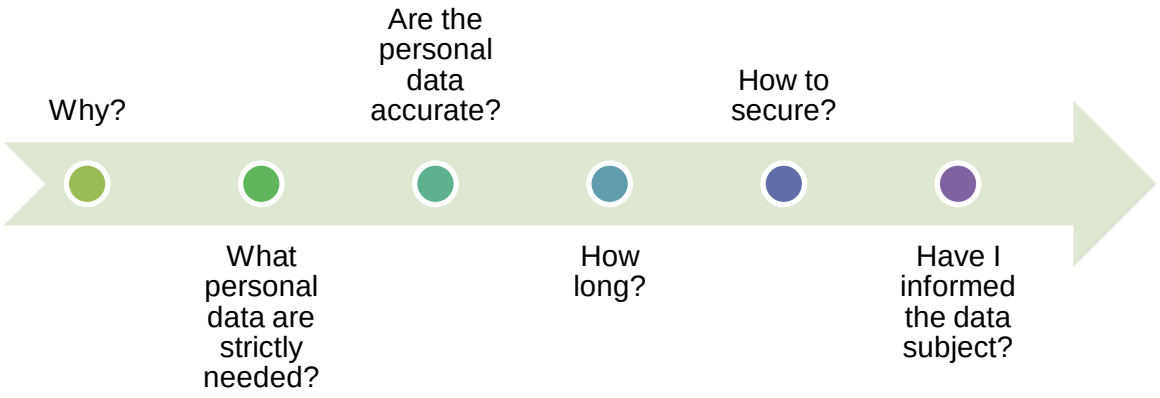

Figure 23: List of key questions for a lawful processing of personal data

4.3.1 Why[12]?

First step. The data controller must determine beforehand the reason why he/she wants to collect and process personal data. It means that all personal data must be processed only for "*specified, explicit and legitimate purposes*"[13]. This implies the purpose limitation principle.

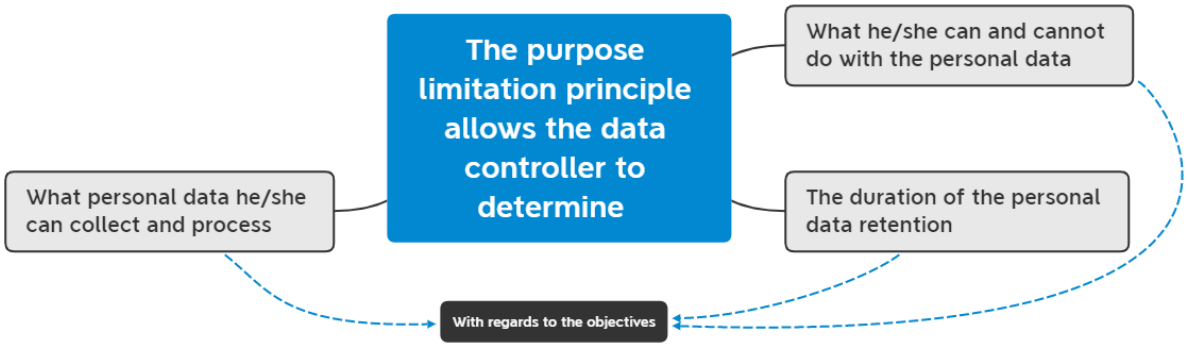

Figure 24: Explanation of the purpose limitation principle

In particular, this principle rejects imprecise formulation such as "Promote safety and security" or "Provide, improve and develop services"[14].

For scientific research, the data controller could obtain the consent from the data subject by providing more general information. Taking into account the recognised ethical standards, the accuracy of the information given should be improved as the research progresses[15].

[12] Article 5.1, b) of the GDPR: "*Personal data shall be: (b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation').*"

[13] Article 5.1, b) of the GDPR.

[14] C. DE TERWANGNE, « Les principes relatifs au traitement des données à caractère personnel et à sa licéité », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier.

[15] Article 29 Working Party, Guidelines on consent under Regulation 2016/679, 28.11.2017 (Revised and adopted on 10.04.2018), WP 259 rev. 01.; Recital 33 of the GDPR.

Secondly, further processing for scientific or historical research purposes or for statistical purposes shall be considered compatible[16]. However, conditions to be respected are provided for in Article 89[17]. The European legislator specifies the notion of scientific research[18] and the notion of statistical purposes[19].

4.3.2 What personal data are strictly needed[20]?

Second step. The controller must determine which data is strictly needed to accomplish the purpose beforehand defined. This is the principle of minimisation. The first question that need to be asked to respect the minimisation requirement is to give the priority to anonymous or at least pseudonymised data. Figure 25 demonstrates the distinction between anonymous data and pseudonymised data.

Figure 25: Distinction between anonymisation and pseudonymisation for personal data protection

Attention must be paid on the impact of the evolution of the technology on the anonymity of the data. Indeed, a data might be anonymous today and not anymore in 6 months. The consequence of such a loss of anonymity is that the data become personal and fall in the scope of GDPR with its duties and sanctions. In consequence, anonymity must be handled with care.

[16] Indeed, article 5.1 b) of the GDPR states that: "*if further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes*"

[17] Article 89 of the GDPR

[18] Recital 159 of the GDPR: "*The notion of scientific research means "the processing of personal data for scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research. In addition, it should take into account the Union's objective under Article 179(1) TFEU of achieving a European Research Area. Scientific research purposes should also include studies conducted in the public interest in the area of public health*".

[19] Recital 162 of the GDPR: "*The notion of statistical purposes means: "any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person*".

[20] Article 5.1, c) of the GDPR:" *Article 5 - Principles relating to processing of personal data 1. Personal data shall be: c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimization')*."

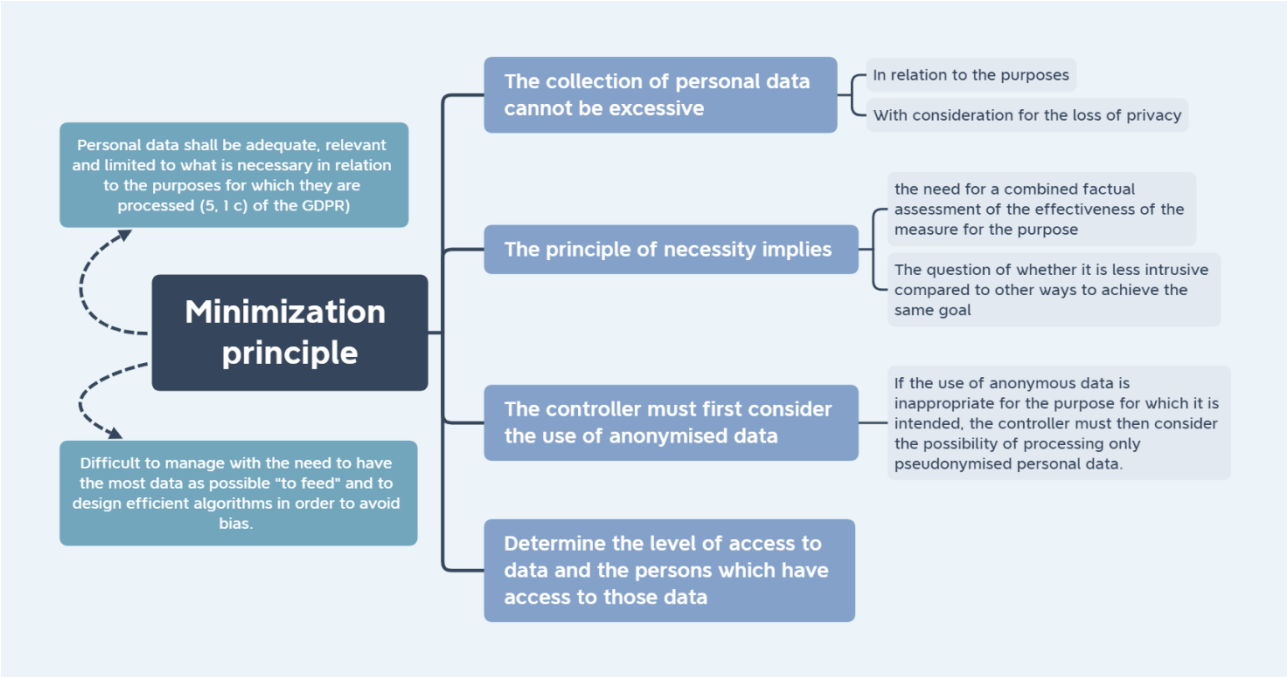

Figure 26: Minimization principle

The principle of minimization involves both organizational measures and technical measures to be taken (see Figure 26). In terms of examples of organizational measures for AI, we could think about listing the data they really need or informing AI researchers about data protection aspects. In terms of technical measures, they mainly concern learning data sets. It would be advisable to start with a small volume of learning data and then check the accuracy of the model when it is populated with new data. One could also think of creating algorithms that would gradually erase the data using automatic forgetting mechanisms[21].

4.3.3 Are the personal data accurate[22]?

Third step. The data controller must put in place reasonable measures to keep the information up to date and accurate[23].

The obligation to keep the data up to date will be judged with more or less severity depending on the context of the processing[24]. For example, in the scenario where an AI system processed several personal data in order to determine if the data subject fulfil the conditions to receive a loan, the accuracy of the personal data on which the decision is based is crucial to avoid any prejudice to the citizen.

[21] Council of Europe, Intelligence artificielle et protection des données. Available at https://rm.coe.int/2018-lignes-directrices-sur-l-intelligence-artificielle-et-la-protecti/168098e1b8.
[22] Article 5, 1, d) of the GDPR: "*Article 5 - Principles relating to processing of personal data 1. Personal data shall be: (d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy')*".
[23] Article 5.1, d) of the GDPR.
[24] C. DE TERWANGNE, « Les principes relatifs au traitement des données à caractère personnel et à sa licéité », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier.

4.3.4 How long[25]?

Fourth step. For the management of personal data, the data controller must determine a retention period for the personal data. This is the storage limitation principle. The data retention period must be aligned with the purpose followed by the data controller[26]. For each type of data collected and in consideration of the relevant purposes, it is necessary to determine if the personal data needs to be stored or whether it can be deleted or anonymized.

The controller must carry out this operation of deletion or anonymization of the data spontaneously, and not at the request of the data subjects[27]. Recital 39 of the GDPR recommends that time limits be set by the controller at the outset for the erasure of personal data or for periodic verification, to ensure that the storage does not exceed what is necessary. By applying the principle of privacy by design and by default, we can establish a technical mechanism whereby conservation of the data automatically ends as soon as the time required to achieve the stated purpose has passed[28].

4.3.5 How to secure[29]?

Fifth step. The data controller and the data processor must ensure an adequate level of protection. One major element for the security policy is to define clearly who can have access to personal data[30]. The GDPR does not impose any specific measures. Security of personal data is an integral part of the effectiveness of the right to privacy.[31]. In the following lines, the security principle will be developed in detail.

[25] Article 5.1, e) of the GDPR: "*Article 5 - Principles relating to processing of personal data 1.Personal data shall be: e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation')*".

[26] See Article 5.1 e) of the GDPR.

[27] C. DE TERWANGNE, « Les principes relatifs au traitement des données à caractère personnel et à sa licéité », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, p. 113.

[28] C. DE TERWANGNE, « Les principes relatifs au traitement des données à caractère personnel et à sa licéité », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, p. 114.

[29] Article 5.1, f) of the GDPR: "*Principles relating to processing of personal data 1. Personal data shall be: (f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality').*"

[30] Article 5.1, f) of the GDPR.

[31] E.C.H.R., 17 July 2008, *I. v. Finlande*, n° 20511/03.

4.3.6 Have I informed the data subject[32]?

Sixth step. Personal data must be processed lawfully[33], fairly and transparently[34]. Transparency implies that some information is provided spontaneously by the controller to the persons concerned by the processing of their personal data[35] as shown in Figure 27.

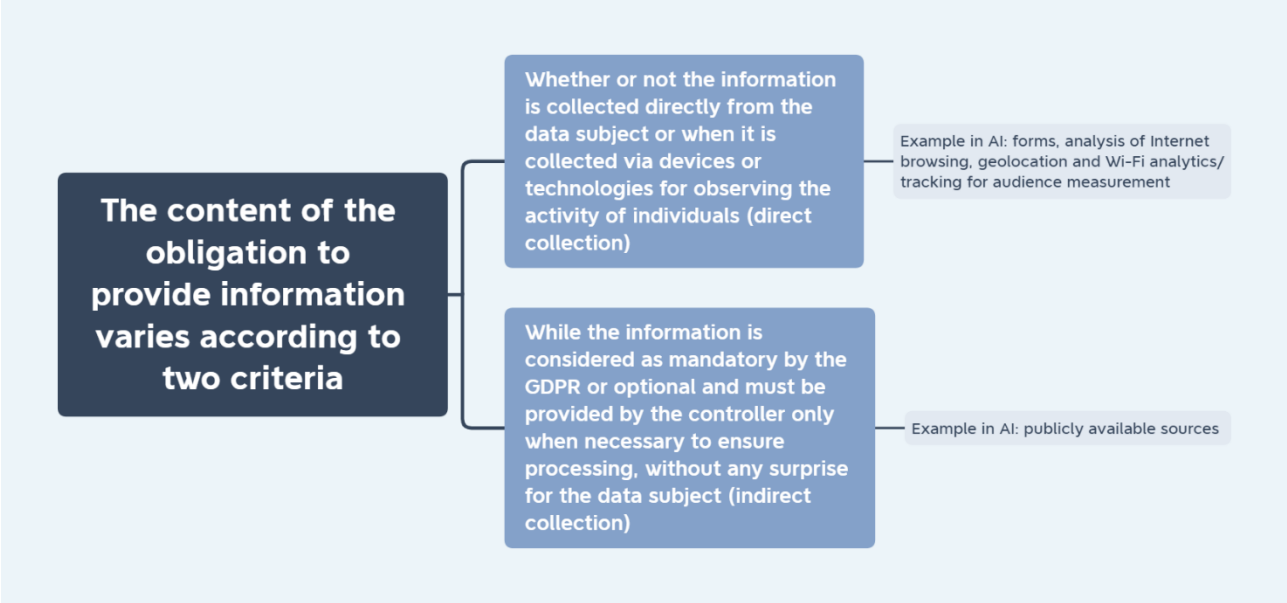

Figure 27: Transparency principle and direct or indirect collection of personal data

Furthermore, information shall be provided in a "*concise, transparent, intelligible and easily accessible*" way[36], for example with a QR code on the sensors or a flash-code to explain the type of sensors and the information's captured and the purposes of the data collections[37].

The following table shows the information to be given for both direct and indirect collection.

[32] Article 5.1, a) of the GDPR: "*Principles relating to processing of personal data. 1. Personal data shall be: (a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency').*"
[33] According to Recital 40 of the GDPR, "*In order for processing to be lawful, personal data should be processed on the basis of the consent of the data subject concerned or some other legitimate basis, laid down by law, either in this Regulation or in other Union or Member State law as referred to in this Regulation, including the necessity for compliance with the legal obligation to which the controller is subject or the necessity for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract*".
[34]. See article 12 of the GDPR
[35] C. DE TERWANGNE, « Les principes relatifs au traitement des données à caractère personnel et à sa licéité », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, p. 91.
[36]. See article 12 of the GDPR.
[37] Art. 29 Working Party, Opinion 8/2014 on the on Recent Developments on the Internet of Things, 16.09.2014, WP 223, p. 18.

Table 4: Information to give to the data subject

INFORMATION TO GIVE TO THE DATA SUBJECT	DIRECT COLLECTION	INDIRECT COLLECTION
Identity and contact of the controller	YES	YES
DPO contact	YES	YES
Purposes of the collection	YES	YES
Legal basis	YES	YES
Legitimate interests (if the legal basis)	YES	YES
Compulsory or optional nature of the data collection	YES	YES
Recipients or categories of recipients of the data	YES	YES
Duration of data storage	YES	YES
Rights of data subjects	YES	YES
Intention of sub-processing	YES	YES
Intention of transfer	YES	YES
Existence of automated decision making	YES	YES
Right to complain	YES	YES
Categories of data collected	NO	YES
Origin of the data	NO	YES

4.4 Notion of data breaches

According to the GDPR, a personal data breach means "*a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed*"[38].

The security of personal data is an integral part of the fundamental rights of the right to privacy and the right to the protection of personal data [39].

[38] Article 4.12 of the GDPR.
[39] ECHR, 17 July 2008, *I v. Finland*, req. n° 20511/03.

As it can be seen in Figure 28, there are several possibilities of data breach[40].

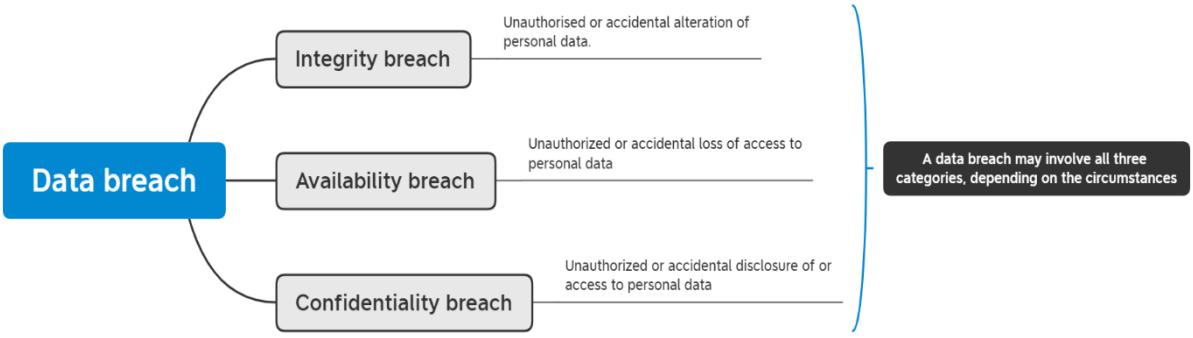

Figure 28: Data breach situations

The Article 29 Working Party gives, as examples of data breach, the loss of the decryption key and the unauthorized reversal of pseudonymisation[41].

4.5 Sharing of roles between the data controller and the data processor

4.5.1 Responsibility

We recall that:

- the data controller is the person who determines the purposes (Why?) and the means (How?) of the processing[42].
- the data processor is the one who actually processes the personal data on behalf of the data controller[43]. We often observe that the processor intervenes at the level of the means determined by the controller.

As an example, in the case of the creation of a medical robotic system, it might be difficult to identify a responsibility regime in an abstract manner. It will be necessary to analyse, on the basis of the factual circumstances, who has the authority to determine the main characteristics of a treatment, according where the AI system is running (e.g. Home environment or workplace), the utilization and the complexity of the robot/AI system[44].

While the liability for compensation was originally limited to the data controller[45] under the Directive 95/46/CE, this is no more the situation set by the new regulation. There are two situations in which the subcontractor may be liable. Firstly, if there is any infringement to the GDPR. Secondly, if the data processor fails to comply with the lawful orders of the data controller. [46].

[40]Art. 29 Working Party, Guidelines on Personal data breach notification under Regulation 2016/679, 03.10.2017 (Revised and adopted on 06.02.2018), WP 250, p. 7-8.

[41] Art. 29 Working Party, Guidelines on Personal data breach notification under Regulation 2016/679, 03.10.2017 (Revised and adopted on 06.02.2018), WP 250.

[42] Article 4.7 of the GDPR.

[43] Article 4.8 of the GDPR.

[44] A. DELFORGE and L. GERARD, "Notre vie privée est-elle réellement mise en danger par les robots ? Etude des risques et analyse des solutions apportées par le GDPR », *in L'intelligence artificielle et le droit*, H. JACQUEMIN and A. DE STREEL (coord.), Bruxelles, Larcier, p.p. 143 et s.

[45.] Article 23 of the Directive 95/46 of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with record to the processing of personal data and on the free movement of such data, L 281/31 (hereinafter "The Directive 96/46/EC").

[46.] Articles 82.2 and 82.3 of the GDPR.

The GDPR also addresses the situation where there is more than one controller or processor implicated in the same processing operation. Each controller or processor will be held accountable for the entirety of the damage. The party that has covered the full amount of compensation has the right to reclaim from the other controllers or processors involved in the same treatment the part of the indemnity corresponding to their share of the responsibility for the damage[47].

As regards security, the controller and the processor he has appointed are both jointly responsible for the security of personal data. Indeed, the controller may only choose a processor who ensures an appropriate level of security for the personal data to be processed[48]. In this respect, the regulation provides that the processing of personal data by a processor on the controller's behalf must be governed by a contract. This should contain the information shown in the Figure 29.

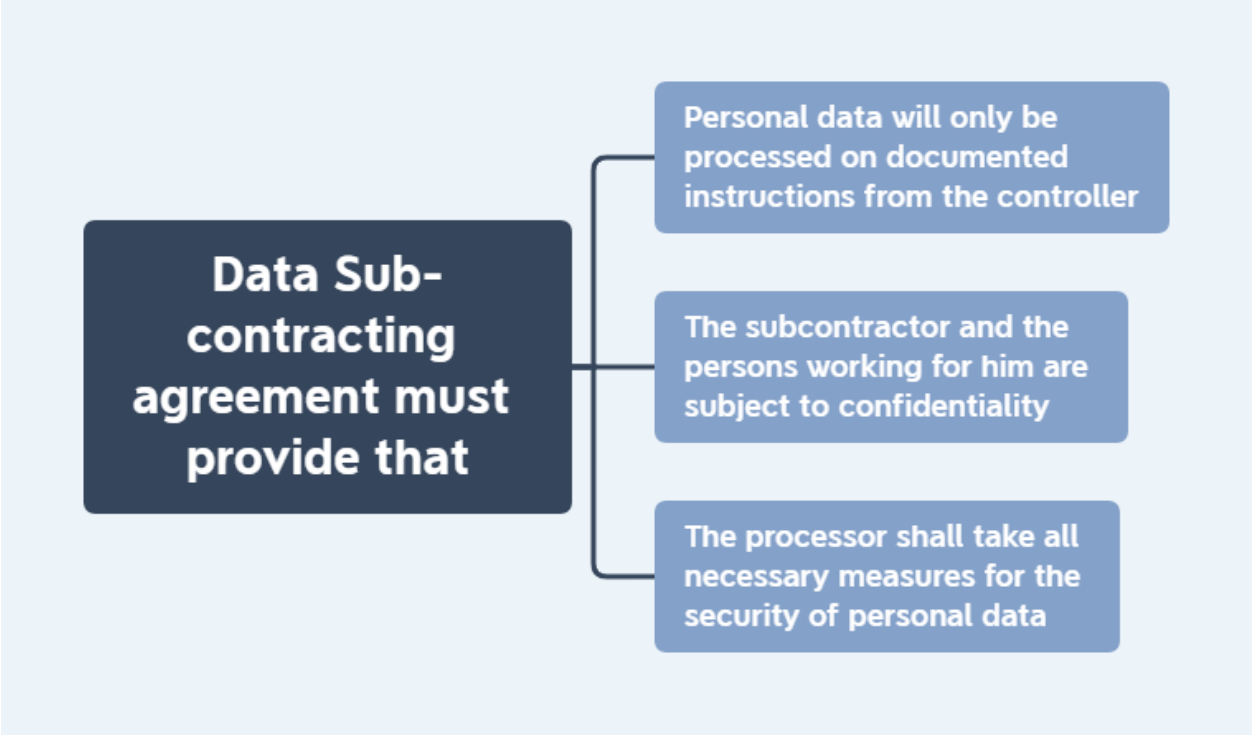

Figure 29: Contractual relationship between the data controller and the data processor

As a core principle of the GDPR, the principle of accountability is of utmost importance[49]. The data controller is therefore held responsible for the security of personal data and must be able to demonstrate compliance with the security requirements laid down by the regulations. Therefore, it is responsible for regularly checking the compliance and efficiency of the technological and organisational safeguards implemented by the data processor to secure the personal data. In this respect, the contract between the data controller and the data processor must provide for audits and inspections to be carried out by the data controller or another auditor appointed by the controller[50].

[47] Articles 82.4 and 82.5 of the GDPR.
[48] Article 28.1 of the GDPR
[49] Article 5.2 of the GDPR.
[50] Article 28.3 h) of the GDPR.

4.5.2 Obligations to the data controller

As it can be seen in Figure 30, the data controller receives the following obligations in case of a data breach[51]:

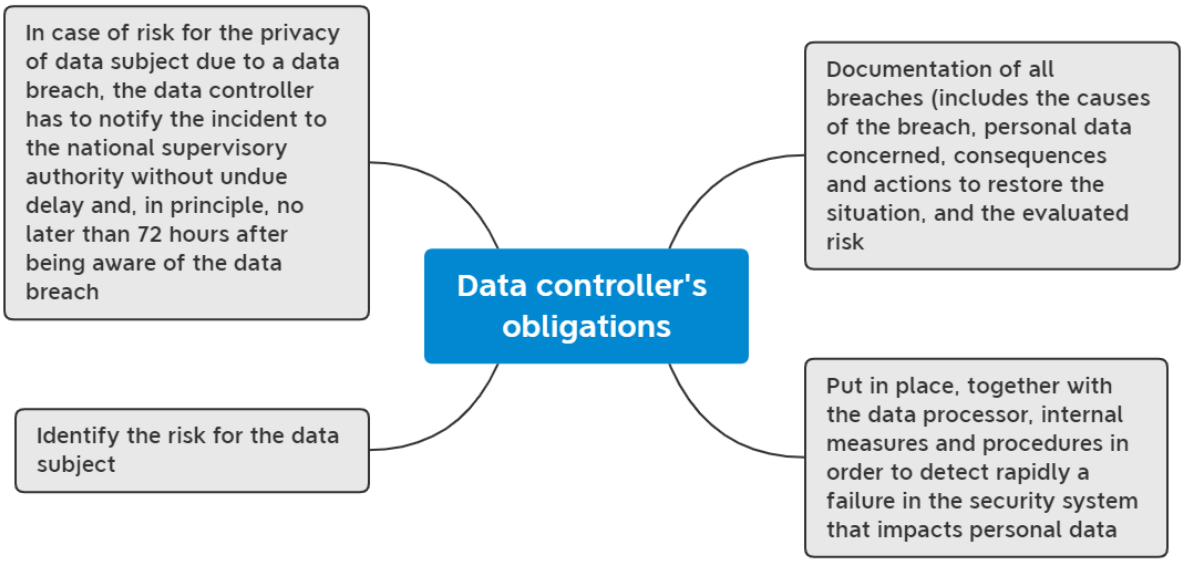

Figure 30: Obligations for the data controller

The Article 29 Working Party recommends the appointment of an internal person that will be responsible for addressing security incidents and assessing the concrete risk for the data subjects[52].

4.5.3 Obligation of data processor

The data processor has to help the data controller to comply with the obligations imposed by the GDPR. Article 33.2 of the GDPR states that: "*The processor shall notify the controller without undue delay after becoming aware of a personal data breach.*"

It is important to keep in mind that as the data controller has to choose and ensure the compliance with the GDPR and the level of security offered by the data processor, he/she retains overall responsibility in case of data breach.

As it can be seen in Figure 31, the data processor receives the following obligations to address a data breach:

[51] Art. 29 Working Party, Guidelines on Personal data breach notification under Regulation 2016/679, 03.10.2017 (Revised and adopted on 06.02.2018), WP 210 and article 33 of the GDPR.

[52] Art. 29 Working Party, Guidelines on Personal data breach notification under Regulation 2016/679, 03.10.2017 (Revised and adopted on 06.02.2018), WP 250, p. 12.

Figure 31: Obligations for the data processor

4.6 Privacy by design and by default[53]

The privacy by design[54] is a formalized obligation introduced by the GDPR to strengthen these data protection principles.

This requirement obliges the data controller to verify that the AI system implemented complies with the core personal data protection principles. Personal data protection obligations must be considered from the outset when designing an artificial intelligence tool. The technology and its algorithmic processing must therefore be developed to respect the legal framework surrounding the use of personal data. [55].

This responsibility rests on the shoulders of the data controller. The legal framework does not provide for products manufacturers, services providers and applications producers for such requirement, although there is a non-binding provision encouraging them[56].

The notion of data protection by design is reflected in the recent modernization of Convention 108 for the Protection of Individuals with regard to Automatic Processing of Personal Data[57]. It is interesting to indicate that this is the first international instrument concerning data protection. The Convention stresses the importance of developing technology that directly respects the protection of personal data. This technology must be accompanied by good data protection practices[58].

While its advancement to the status of legal rules at the European level is new, the Court of Justice of the European Union insisted in the past on the importance of the implementation of measures in order to respect the proportionality of the interference into the privacy of individuals caused by the

[53] Article 25.1 of the GDPR: "*Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects*."

[54] Article 25.1 of the GDPR.

[55] C. DE TERWANGNE K. ROSIER and B. LOSDYCK, « Lignes de force du nouveau Règlement relatif à la protection des données à caractère presonnel », *Journal de droit européen. 2016, pp. 32-33.*

[56] Recital 78 of the GDPR.

[57] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, signed in Strasbourg the 28 January 1981, ETS No.108 (Convention 108+, hereafter). and Additional Protocol to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, regarding supervisory authorities and transborder data flows, Strasbourg, 08/11/2001, ETS No.181.This convention is open as other countries than the members of the CoE can join it.

[58] Council of Europe, Explanatory Report to the Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, 10.X.2018, Strasbourg, p. 15. Available at: https://rm.coe.int/16808ac91a

use of their personal data[59]. The Court has recently recalled that the processing of personal data constitutes in itself an interference with the fundamental right to privacy, irrespective of whether the information concerned is sensitive or not and whether the individuals concerned have suffered any inconvenience. [60].

In the *Google Spain* case, the Court had the opportunity to call for the deployment of the full effectiveness of the data protection requirements by setting up guarantees by the data controller[61]. In the *Digital Rights* decision[62], the Court also called for the development of *specific* and *adapted* rules[63] tailored to the large amount of data required to be stored, the sensitive nature of the data and the risk of unlawful access, in order to ensure their full integrity and confidentiality[64].

In addition to the legal obligation, respect of the privacy principle by the controller from the time of conception makes it possible to compensate for the lack of knowledge of potential users of artificial intelligence services[65]. However, it should be noted that products' manufacturer that are not responsible for data processing are not bound by the rule of privacy by design (for example, when using a drone, it is not the manufacturer of the product that will be liable for the processing if no personal data has been processed by the manufacturer but by the user)[66]. Nonetheless, the GDPR encourages product manufacturers, service providers and application producers to take data protection regulations into account when they develop and design their products or services[67]. This provision is only contained in a recital and not in a mandatory rule. This recital is, however, important to guarantee that all those involved in the design and development of an artificial intelligence system are concerned about security.

The GDPR provides practical examples of measures that can be implemented, as shown in Figure 32. These include minimizing the processing of personal data, giving priority to pseudonymization and ensuring transparency of processing with regard to data subjects[68].

[59] Case C-131/12, *Google Spain v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González*, 13 May 2014, ECLI: EU: C: 2014:317; Joined Cases C-293/12 and C-594/12, pt. 72, *Digital Rights Ireland Ltd and Seitlinger and Others*, 8 April 2014, ECLI:EU:C:2014:238, pts. 38, 46, 61; Bygrave, Lee A., Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements (June 20, 2017). Oslo Law Review, Volume 4, No. 2, 2017, pp.109-110. Available at SSRN: https://ssrn.com/abstract=3035164.

[60] Joined Cases C-293/12 and C-594/12, *Digital Rights Ireland Ltd and Seitlinger and Others*, 8 April 2014, ECLI: EU: C: 2014:238, pt. 33.

[61] Case C-131/12, *Google Spain v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González*, 13 May 2014, ECLI: EU: C: 2014:317, pt. 38.

[62] Let us indicate that this decision was about the validity of the Directive 2006/24/EC on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks ; Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks, *O.J.*, L 105.

[63] Joined Cases C-293/12 and C-594/12, *Digital Rights Ireland Ltd and Seitlinger and Others*, 8 April 2014, ECLI: EU: C: 2014:238, pt. 66.

[64] Joined Cases C-293/12 and C-594/12, *Digital Rights Ireland Ltd and Seitlinger and Others*, 8 April 2014, ECLI: EU: C: 2014:238, pt. 67. See also Bygrave, Lee A., Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements (June 20, 2017). Oslo Law Review, Volume 4, No. 2, 2017, p.111. Available at SSRN: https://ssrn.com/abstract=3035164.

[65] A. DELFORGE and L. GERARD, "Notre vie privée est-elle réellement mise en danger par les robots ? Etude des risques et analyse des solutions apportées par le GDPR », *in L'intelligence artificielle et le droit*, H. JACQUEMIN and A. DE STREEL (coord.), Bruxelles, Larcier, p. 169.

[66] A. DELFORGE and L. GERARD, "Notre vie privée est-elle réellement mise en danger par les robots ? Etude des risques et analyse des solutions apportées par le GDPR », *in L'intelligence artificielle et le droit*, H. JACQUEMIN and A. DE STREEL (coord.), Bruxelles, Larcier, p. 178.

[67] Recital 78 of the GDPR.

[68] Recital 78 of the GDPR.

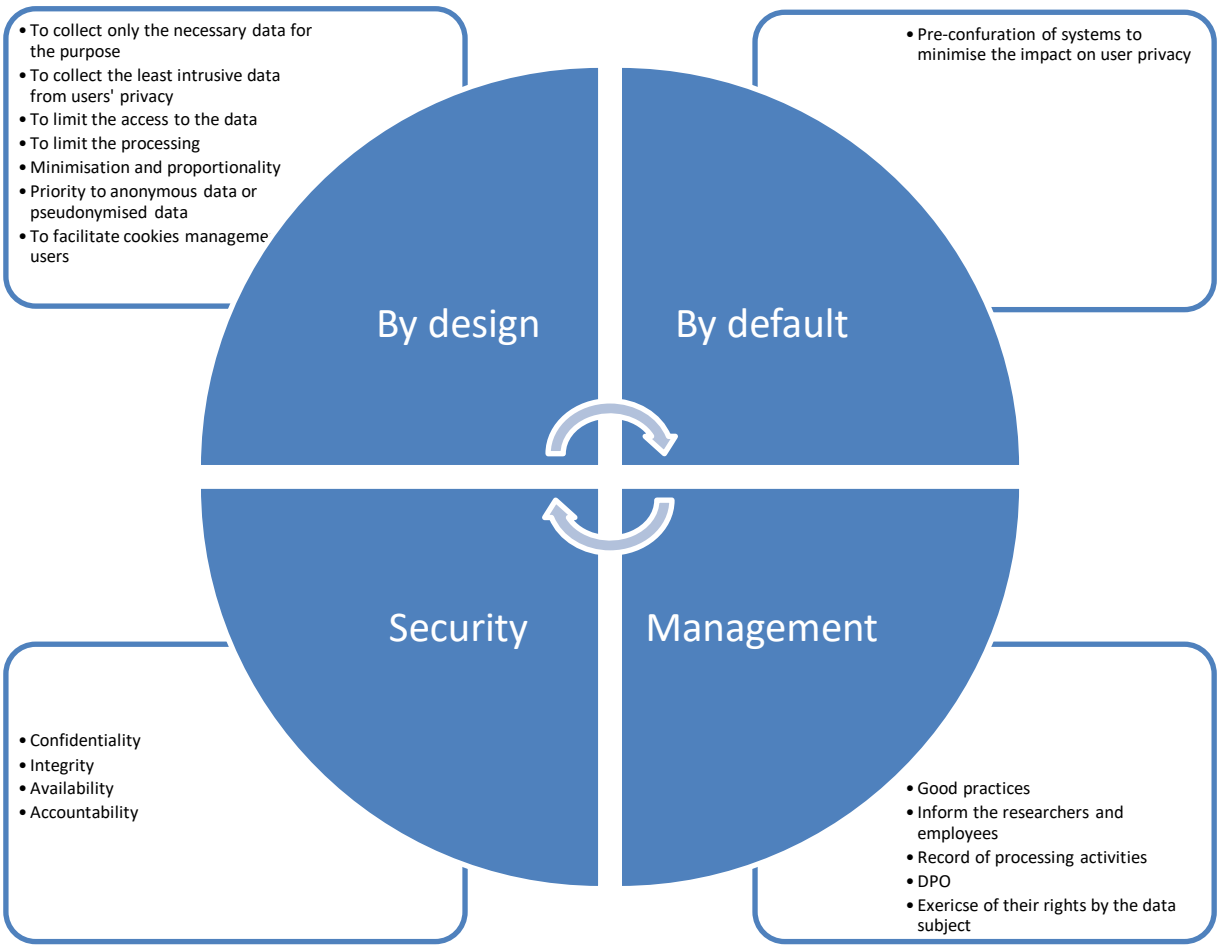

Figure 32: Privacy by design and management of personal data

A parallel path is taken by ENISA in order to facilitate the configuration of the technology developed[69].

It is important to underline that the controller is responsible for the configuration of his device or service and must be able to prove, in case of an audit by the competent data protection authority, compliance with the requirements of the regulation.

In addition, ENISA emphasises the crucial role of transparency. Indeed, the persons concerned must be able to be informed both of the initial configuration put in place and of the possibilities of modifying the pre-settings in full knowledge of the consequences[70].

The following figure shows the overview on choices in the design process regarding functionality or behaviour of an IT system.

[69] ENISA, Recommendations on shaping technology according to GDPR provisions. Available at: https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions-part-2
[70] ENISA, Recommendations on shaping technology according to GDPR provisions, p. 19. Available at: https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions-part-2 .

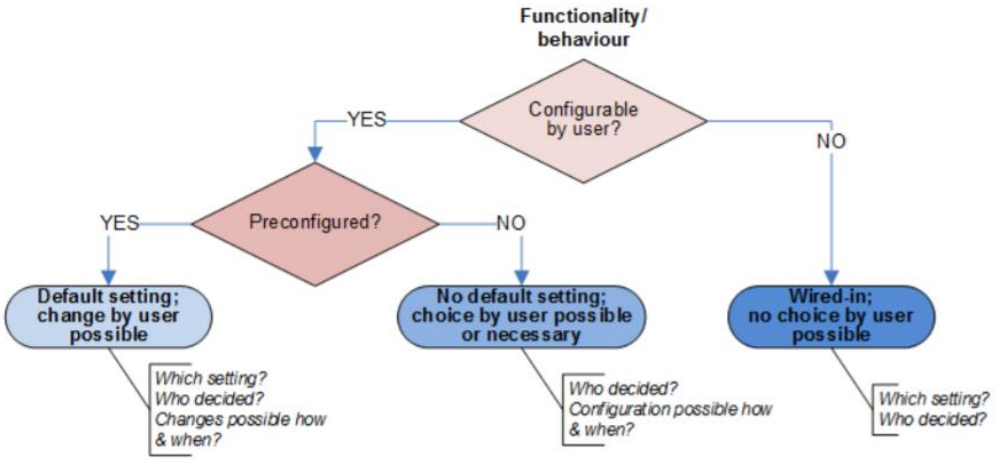

Figure 33: Overview on choices in the design process regarding functionality or behaviour of an IT system [134]

4.7 Security of the personal data[71]

Article 32 of the GDPR defines what surrounds the security of personal data and the technical and organisational elements to be put in place to ensure a level of security appropriate to the risk. The Figure 34 shows the timeline to process.

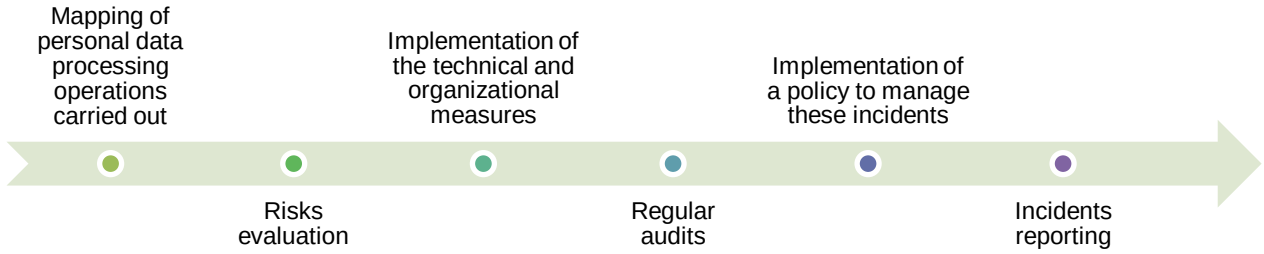

Figure 34: Timeline for Article 32 of the GDPR on the security of personal data

4.7.1 Risks evaluation

The GDPR calls for appropriate security measures and thus adopts a risk-based approach. Therefore, depending on the nature and amount of personal data and the processing carried out, the controller must determine the risks, the likelihood of those risks occurring and the severity of the risks to individuals in a data protection impact assessment, as shown in the Figure 35 [72]. These are not only risks related to privacy and personal data protection, but also risks related to freedom of expression, freedom of thought, freedom of movement, discrimination, etc. For example, theft, discrimination, financial loss, unauthorised removal of pseudonymisation, identity theft, etc[73].

[71] Article 32 of the GDPR.
[72] See Article 32 of the GDPR.
[73] F. DUMORTIER, « La sécurité des traitements de données, les analyses d'impact et les violations de données », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, p. 188.

In order to examine the probability of a risk occurring, ENISA has put in place a dedicated methodology[74]. In Table 5, we can see that its methodology is very close to the risk-based approach enshrined in the legislation:

Table 5: GDPR Risk-based approach and the ENISA methodology

GDPR: risk-based approach	ENISA's methodology[75]
• The nature of the personal data • The volume of the personal data • The processing operations	Step 1: "Definition of the processing operations and its context"
Evaluation of the risk	Step 4: "Evaluation of risk"
Evaluation of the probability that a risk occurs	Step 3: "Definition of possible threats and evaluation of their likelihood"
The seriousness of the risks for data subject	Step 2: "Understanding and evaluating impact"

Article 35 of the GDPR provides for a new obligation attributed to the data controller. Firstly, the data controller has to verify if the nature, the volume and the processing of personal data may lead to a substantial risk to the protection of the data subjects' rights and freedoms. Secondly, if there is a high risk, an assessment of the impact needs to be achieved[76].

This obligation is of utmost importance for AI activities. Article 35.7 establishes a list of minimum requirements[77].

Figure 35: Data protection impact assessment

If after the elaboration of a DPIA and after taking security measures a residual risk persists, the data controller must address to the national supervisory authority[78].

[74] ENISA,Handbook on Security of Personal Data Processing, December 2017. Available at: https://www.enisa.europa.eu/publications/handbook-on-security-of-personal-data-processing.
[75] ENISA,Handbook on Security of Personal Data Processing, December 2017. Available at: https://www.enisa.europa.eu/publications/handbook-on-security-of-personal-data-processing.
[76] C. DE TERWANGNE, K. ROSIER and B. LOSDYCK, « Le règlement européen relatif à la protection des données à caractère personnel : quelles nouveautés », *J.D.E*, 2017, p. 309.
[77] Art. 29 Working Party, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, 04.10.2017, WP 248, p. 17.
[78] Article 36.1 of the GDPR ; V. VERBRUGGEN, "Mise en œuvre du nouveau Règlement général sur la protection des données: coup de projecteur sur certaines nouvelles obligations à charge des responsables de traitement et des sous-traitants", *Orientations*, 2017, p. 15.

For more information, the Article 29 Working Party adopted two annexes on existing EU DPIA frameworks and criteria for an acceptable DPIA[79].

4.7.2 Implementation of technical and organisation measures

Even before the GDPR, the Article 29 Working Party already insisted on proactive risk management[80]. We could say that the core of the regulation on the protection of personal data is based on both the technical and organisational measures that Article 32 requires the controller to implement. While these measures are designed to secure personal data, they also require the controller to ask himself the right questions in order to comply with the main principles around the processing of personal data set out in Article 5. Therefore, these technical and organisational measures are also intended to enable the controller to comply with the GDPR. However, the GDPR is pragmatic as the choice of technical measures could be influenced by considering the state of the art, the cost of implementation and the nature, scope, context and purposes of processing[81]. We remind that the regulation is technically neutral.

To fulfil the requirements from the article 32 of the GDPR, the organisational and technical measures must be designed to respect four considerations[82] that can be seen in Figure 36.

Figure 36: Components of the security concept for the protection of personal data

Accountability is defined as " *the property that ensures that an entity's actions are tracked and attributed to that single entity.*"[83]

The data controller must notably verify that:

- The rules for accessing and processing personal data are clearly defined in an internal policy[84].
- The authenticity of the personal data and that their manipulation has not had for effect to modify the content of the information[85].

[79] Available at: http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083.

[80] See Art. 29 Working Party, Opinion 03/2014 on "Personal Data Breach Notification, 25.03.2014, WP213 and Art. 29 Working Party, Statement on the role of a risk-based approach in data protection legal frameworks", 30.05.2014, WP 218.

[81] Article 25.1 of the GDPR.

[82] F. DUMORTIER, "La sécurité des traitements de données, les analyses d'impact et les violations de données", *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier.

[83] Autorité de protection des données, «La sécurité des données à caractère personnel Available at: https://www.autoriteprotectiondonnees.be/publications/note-relative-a-la-securite-des-donnees-a-caractere-personnel.pdf

[84] Recital 39 of the GDPR.

[85] See Art. 29 Working Party, Opinion 05/2012 on Cloud Computing, 01.07.2012, WP 196, p. 15.

- The risk has been correctly determined and the seriousness for the data subjects in the case of a temporary interruption of service. Furthermore, the sustainability of the supports is also important[86].
- The procedure to be able to effectively determine and trace who has had access to personal data and when is established[87].

The GDPR gives some examples of security measures[88]. Other technical means could be, for example, the establishment of internal privacy filters to limit the access, installation of firewalls and intrusion detection systems. These are only examples and by no means constitute an exhaustive list of security measures. We could also suggest the adoption of standards as ISO[89].

As it can be seen in Figure 37, in order to comply with the four security objectives listed above, the data controller must implement various measures[90].

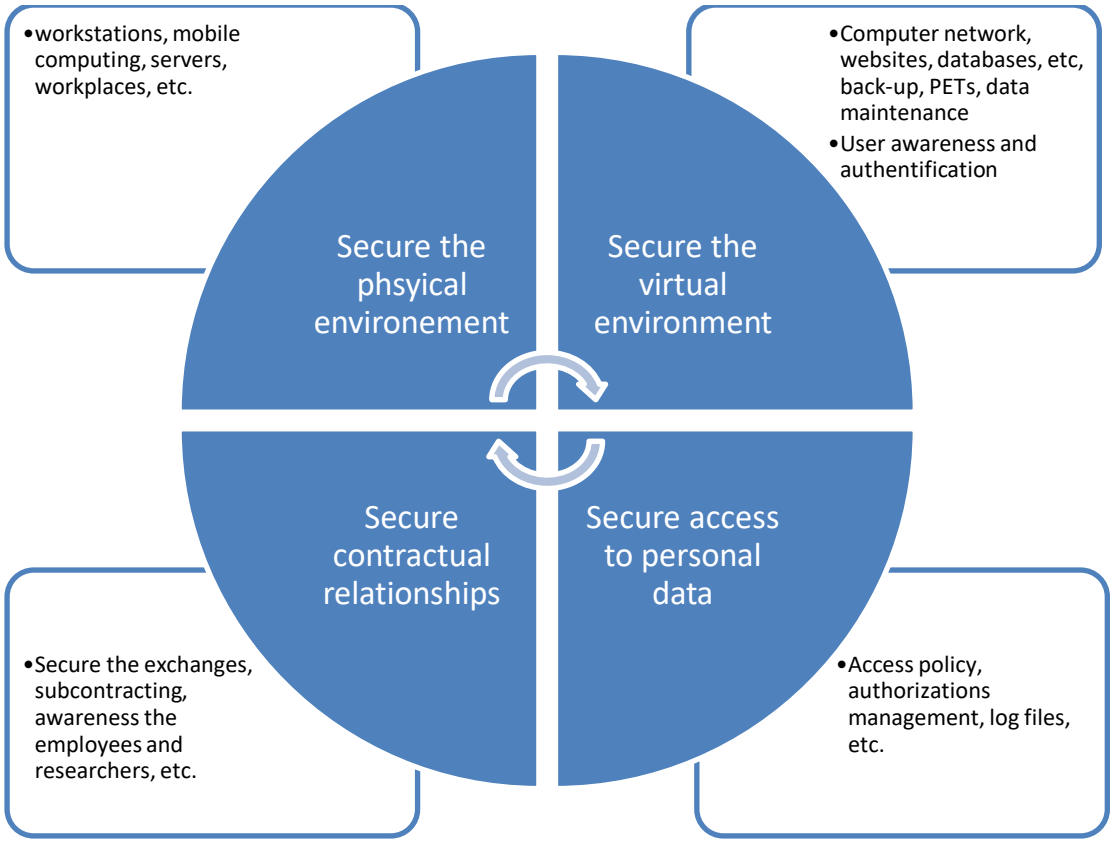

Figure 37: Measures to ensure security of personal data

[86] F. DUMORTIER, « La sécurité des traitements de données, les analyses d'impact et les violations de données », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, p. 154.
[87] F. DUMORTIER, « La sécurité des traitements de données, les analyses d'impact et les violations de données », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier, p. 157.
[88] See Article 32 of the GDPR.
[89] C. DE TERWANGNE, J-M.VAN GYSEGHEM, « Analyse détaillée de la loi de protection des données et de son arrêté royal d'exécution » , in *Vie privée et données à caractère personnel,* Bruxelles, Politeia, 2013, p. 122.
[90] This list is the result of the combination of ENISA, Handbook on Security of Personal Data Processing and CNIL, La sécurité des données personnelles. On this matter, see also F. DUMORTIER, « La sécurité des traitements de données, les analyses d'impact et les violations de données », *in Le règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. DE TERWANGNE et K. ROSIER (coord.), Brussels, Larcier.

4.8 Data breaches notification

4.8.1 Obligation of notification

One of the core principle for the protection of personal data is the guarantee of the 'integrity and confidentiality'[91] of the information.

In order to enhance compliance, the GDPR puts in place a system of personal data breach notification for both data controller and data processor[92].

The data controller and data processor have to implement technological and organizational measures to identify a data breach without undue delay[93].

As it is seen in Table 6, there are two possible notifications[94].

Table 6: Obligation to notify security breaches

Concept	Notification to the supervisory authority	Notification to the data subject
Description of the data breach	Yes	Yes
Contact details about the DPO	Yes	Yes
Impact of the data breach	Yes	Yes
Mitigation measures	Yes	Yes
Clear and plain language	No specific obligation	Yes
For any data breach	Yes	Only in case of high risk for the data subject

[91] Article 5.1 f) of the GDPR.

[92] Article 33 and Recital 81 of the GDPR.

[93] Recital 87 states that: "*It should be ascertained whether all appropriate technological protection and organisational measures have been implemented to establish immediately whether a personal data breach has taken place and to inform promptly the supervisory authority and the data subject. The fact that the notification was made without undue delay should be established taking into account in particular the nature and gravity of the personal data breach and its consequences and adverse effects for the data subject. Such notification may result in an intervention of the supervisory authority in accordance with its tasks and powers laid down in this Regulation*".

[94] See Article 33 and Rectal 85 of the GDPR for the notification to the supervisory authority; Article 34 and Recital 84 for the notification to the data subject; Art. 29 Working Party, Guidelines on Personal data breach notification under Regulation 2016/679, 03.10.2017 (Revised and adopted on 06.02.2018), WP 250.

The data controller must inform the data subject about the security breach only when this has led to significant risks to his or her rights. In order to determine concretely what a high risk is, the Article 29 Working Party has identified several criteria[95] that can be seen in Figure 38.

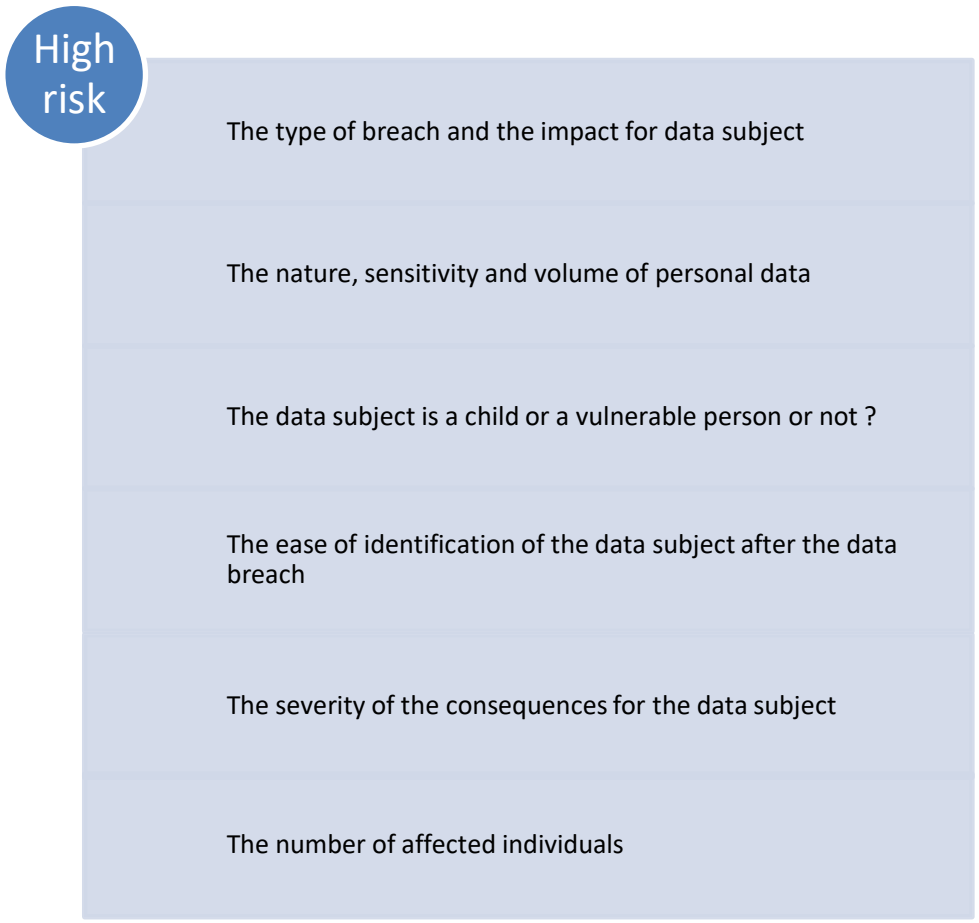

Figure 38: Criteria to evaluate the risk of a data breach for the data subject

4.9 Conclusions

The General Data Protection Regulation imposes several obligations on the data controller of an artificial intelligence system. Due to the amount of personal data processed in AI mechanisms, a proper implementation of these principles is already a first way for the controller to avoid threats as much as possible. Before any processing, the data controller must answer the following questions:

- Why would I like to collect and process personal data?
- What personal data are strictly needed?
- Is the personal data accurate?
- How long should I keep the personal data?
- How to secure the system and the personal data?
- Have I informed in a proper manner the data subject?

[95] Art. 29 Working Party, Guidelines on Personal data breach notification under Regulation 2016/679, 03.10.2017 (Revised and adopted on 06.02.2018), WP 250, pp. 22 and seq.

All these issues must be addressed prior to any collection or use of personal data to consider the privacy by design and by default requirements.

In particular, the fifth question is of the greatest importance in AI system. Indeed, the security of personal data is an integral part of the fundamental rights of the right to privacy and the right to the protection of personal data[96].

The GDPR imposes a duty to take appropriate security precautions and therefore adopts a risk-based approach. In order to ensure an appropriate level of security regarding the risk of the processing and the nature of the personal data, the following management should be put in place:

- Mapping of the personal data processing operations.
- Evaluation of the risks of the processing for the data subjects.
- Elaboration of a data protection impact assessment if the nature, the volume and the processing of personal data may entail a high risk for the protection of the rights and freedoms of the data subject.
- Implementation of the appropriate technical and organisation measures regarding the risks evaluated.
- Organisation of regular audits.
- Implementation of a policy to manage these incidents.

While the responsibility for compensation was limited to the controller under Directive 95/46/EC, this is no longer the case under the new Regulation. Indeed, according to Article 82 of the GDPR, the data processor is accountable for damages caused by processing that does not comply with the obligations specifically imposed on the data processor by the GDPR. He will also be liable if he has acted outside or contrary to the legitimate instructions of the controller. However, the data processor will be exempt from liability if it proves that it is not responsible for the event causing the damage. The GDPR also addresses the situation where there is more than one controller or processor involved in the same processing operation. Each controller or processor will be declared responsible for the full amount of the damage in order to ensure that the data subject is fully and effectively compensated. The entity that has paid the total compensation has the right to reclaim from the other controllers or processors engaged in the same treatment the part of the compensation corresponding to their share of responsibility[97].

[96] ECHR, 17 July 2008, *I v. Finland*, req. n° 20511/03.

[97] Articles 82.4 et 82.5 of the GDPR.

Chapter 5 Summary and Conclusion

This report analyses the AI threats landscape from the point of view of securing AI systems from threats (incidents and attacks) particular to the nature of machine learning and deep learning systems. The use of AI for either attacking or protecting systems in general is not addressed.

The report focuses on Machine Learning (ML) systems leaving aside other types of Artificial Intelligence systems, since this is the work scope of SAFAIR Program. Therefore, AI term in the document refers to ML field.

The document first reviews the main taxonomies of AI systems since the threat analysis starts from understanding the architecture system parts (assets to protect) and how they work and could be exploited as attack vectors. The report documents existing threat classifications, and published examples of threats against different AI systems are also studied, with a focus on adversarial machine learning attack techniques and methods.

As a result of these studies, SAFAIR has designed an AI Threat model that integrates into a single domain model multiple concepts and aspects related to threats against AI systems. This AI threat domain model has been used as the core of the common understanding between SAFAIR work tasks and has served to structure the body of knowledge around AI threats. A MySQL-based Knowledge Base has also been implemented and populated with first contents from the survey, which are expected to be enriched and extended in the future with results from all other tasks in SAFAIR, particularly understanding on attack techniques and possible protections to counter them.

The lawful processing and GDPR compliance by AI systems are challenges that deserve special attention. As part of SAFAIR objective to support trustworthy, fair and GDPR compliant AI systems, in Chapter 4 we gave a practical summary of the key concepts involved in lawful and compliant AI. The section describes also the main threats faced by AI system providers and operators as data processors, and how they impact the design and operation of the AI systems.

Chapter 6 List of Abbreviations

Abbreviation	Translation
AI	Artificial Intelligence
AML	Adversarial Machine Learning
API	Application Programming Interface
APT	Advanced Persistent Threat
BDE	Big Data Ecosystem
CNN	Convolutional Neural Network
CSA	Cloud Security Alliance
DNN	Deep Neural Network
DPIA	Data protection impact assessment
DPO	Data Protection Officer
DRL	Delayed Reinforcement Learning
EDPB	European Data Protection Board
ENISA	European Network and Information Security Agency
ETSI	European Telecommunications Standards Institute
EU	European Union
ICT	Information and Communication Technology
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
IEC	International Electrotechnical Commission
IPR	Intellectual Property Rights
ISG	Industry Specification Group
ISO	International Organization for Standardization
IT	Information Technology
ITSEC	Information Technology Security Evaluation Criteria
MARS	Multivariate Adaptive Regression Splines
MIA	Membership-Inference Attacks
ML	Machine Learning
MLP	Multilayer Perceptron
NBDRA	NIST Big Data Reference Architecture
NIST	National Institute of Standards and Technology
PDTR	Proposed Draft Technical Report
PIA	Property Inference Attacks
QR	Quick Response
RL	Reinforcement Learning

Abbreviation	Translation
SAI	Securing Artificial Intelligence
SARSA	State–Action–Reward–State–Action
SEI	Software Engineering Institute
SOTA	State Of The Art
SVM	Support Vector Machine
TFEU	Treaty on the Functioning of the European Union

Chapter 7 Bibliography

[1] EC High-Level Expert Group on Artificial Intelligence. Available at: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[2] E Damiani, C. A. Ardagna, F. Zavatarelli, E. Rekleitis (ed.) and L. Marinos, "Big Data Threat Landscape and Good Practice Guide", 2016. Available at: https://www.enisa.europa.eu/publications/bigdata-threat-landscape/at_download/fullReport

[3] NIST NBD-WG. 2017. NIST Big Data Reference Architecture, 2017. Available at: https://bigdatawg.nist.gov/_uploadfiles/M0639_v1_9796711131.docx

[4] Big Data Working Group Big Data Taxonomy, September 2014. Available at: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf

[5] Software Engineering Institute's Comments on NIST IR 8269: A Taxonomy and Terminology of Adversarial Machine Learning. Available at: https://resources.sei.cmu.edu/asset_files/WhitePaper/2020_019_001_637336.pdf

[6] Demchenko, Y., De Laat, C., & Membrey, P. (2014, May). Defining architecture components of the Big Data Ecosystem. In 2014 International Conference on Collaboration Technologies and Systems (CTS) (pp. 104-112). IEEE.

[7] Gartner Glossary, Big Data definition. Available at: http://www.gartner.com/it-glossary/big-data/

[8] Golstein, B. (October 2018). A Brief Taxonomy of AI. Available at: https://www.sharper.ai/taxonomy-ai/

[9] Wang, S., Chaovalitwongse, W., & Babuška, R. (2012). "Machine learning algorithms in bipedal robot control". IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 42(5), 728–743. https://doi.org/10.1109/TSMCC.2012.2186565

[10] How, H. (2016). "The master algorithm: how the quest for the ultimate learning machine will remake our world". In Choice Reviews Online (Vol. 53). https://doi.org/10.5860/choice.194685

[11] Peng, W., Wang, H., Bailey, J., Tseng, V., Ho, T., Zhou, Z. & Chen, A. (2014). "Trends and Applications in Knowledge Discovery and Data Mining". PAKDD 2014 International Workshops, Taiwan, May 13-16, 2014.

[12] Types of machine learning algorithms (2015). Available at: https://en.proft.me/2015/12/24/types-machine-learning-algorithms/

[13] Mindmeister taxonomy of Machine Learning Algorithms Grouped by Similarity. Available at: https://www.mindmeister.com/es/675097196/machine-learning-algorithms-grouped-by-similarity

[14] Auernhammer, K., Kolagari, R. T., & Zoppelt, M. (2019). "Attacks on machine learning: Lurking danger for accountability". CEUR Workshop Proceedings, 2301.

[15] NIST IR 8269 Draft A Taxonomy and Terminology of Adversarial Machine Learning report: https://csrc.nist.gov/publications/detail/nistir/8269/draft

[16] M. Fredrikson, S. Jha, T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures". In proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (2015).

[17] Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. (2018). "A survey on security threats and defensive techniques of machine learning: A data driven view". IEEE access, 6, 12103-12117.

[18] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, "Adversarial attacks and defences: A survey", CoRR abs/1810.00069 (2018). arXiv:1810.00069. URL http://arxiv.org/abs/1810.00069

[19] X. Liao, L. Ding, Y. Wang, “Secure machine learning, a brief overview”, in: 2011 Fifth International Conference on Secure Software Integration and Reliability Improvement - Companion, 2011, pp. 26–29 (June 2011). doi:10.1109/SSIRI-C.2011.15.

[20] B. Biggio, B. Nelson, P. Laskov, “Poisoning Attacks against Support Vector Machines”, arXiv e-prints (2012) arXiv:1206.6389 (Jun 2012). arXiv:1206.6389.

[21] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, K. Xia, “Exploiting machine learning to subvert your spam filter”, in: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, LEET’08, USENIX Association, Berkeley, CA, USA, 2008, pp. 7:1–7:9 (2008). URL http://dl.acm.org/citation.cfm?id=1387709.1387716

[22] A. Shafahi, W. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, T. Goldstein, “Poison frogs! targeted clean-label poisoning attacks on neural networks” (04 2018).

[23] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, F. Roli, “Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization”, arXiv e-prints (2017) arXiv:1708.08689 (Aug 2017). arXiv:1708.08689.

[24] S. Mei, X. Zhu, “Using machine teaching to identify optimal training-set attacks on machine learners”, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI’15, AAAI Press, 2015, pp. 2871– 6 2877 (2015). URL http://dl.acm.org/citation.cfm?id=2886521.2886721

[25] S. Szyller, B. G. Atli, S. Marchal, N. Asokan, “Dawn: Dynamic adversarial watermarking of neural networks”, arXiv preprint arXiv:1906.00830 (2019).

[26] C. Yang, Q. Wu, H. Li, Y. Chen, “Generative Poisoning Attack Method Against Neural Networks”, arXiv e-prints (2017) arXiv:1703.01340 (Mar 2017). arXiv:1703.01340.

[27] X. Chen, C. Liu, B. Li, K. Lu, D. Song, “Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning”, arXiv e-prints (2017) arXiv:1712.05526 (Dec 2017). arXiv:1712.05526.

[28] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, F. Roli, “Is feature selection secure against training data poisoning”?, CoRR abs/1804.07933 (2018). arXiv:1804.07933. URL http://arxiv.org/abs/1804.07933

[29] B. Biggio, G. Fumera, F. Roli, “Pattern recognition systems under attack: Design issues and research challenges”, International Journal of Pattern Recognition and Artificial Intelligence 28 (07) (2014) 1460002 (2014). doi:10.1142/S0218001414600027.

[30] O. Suciu, R. Marginean, Y. Kaya, H. D. III, T. Dumitras, “When does machine learning FAIL? generalized transferability for evasion and poisoning attacks”, in: 27th USENIX Security Symposium (USENIX Security 18), USENIX Association, Baltimore, MD, 2018, pp. 1299–1316 (Aug. 2018). URL: https://www.usenix.org/conference/usenixsecurity18/presentation/suciu

[31] N. Papernot, P. McDaniel, A. Sinha, M. P. Wellman, “Sok: Security and privacy in machine learning”, in: 2018 IEEE European Symposium on Security and Privacy (EuroS&P), 2018, pp. 399–414 (April 2018). doi:10.1109/EuroSP.2018.00035

[32] Paolo Russu, Ambra Demontis, Battista Biggio, Giorgio Fumera, and Fabio Roli. “Secure Kernel Machines against Evasion Attacks”. In: Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security - ALSec ’16 (2016), pp. 59–69. DOI: 10.1145/2996758.2996771.

[33] Pitropakis, Nikolaos, et al. "A taxonomy and survey of attacks against machine learning." Computer Science Review 34 (2019): 100199.

[34] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndi_c, P. Laskov, G. Giacinto, and F. Roli. “Evasion attacks against machine learning at test time”. In H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezn y editors, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Part III, volume 8190 of Lecture Notes in Computer Science, pages 387{402. Springer Berlin Heidelberg, 2013.

[35] B. Biggio, G. Fumera, and F. Roli. “Security evaluation of pattern classifiers under attack”. IEEE Transactions on Knowledge and Data Engineering, 26(4):984-996, April 2014.

[36] F. Wang, W. Liu, and S. Chawla. “On sparse feature attacks in adversarial learning”. In IEEE Int'l Conf. on Data Mining (ICDM), pages 1013{1018. IEEE, 2014.

[37] N. Carlini, D. Wagner, “Adversarial examples are not easily detected: Bypassing ten detection methods”, arXiv preprint arXiv:1705.07263 (2017).

[38] N. Dalvi, P. Domingos, S. Sanghai, D. Verma, et al., “Adversarial classification”, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 99–108.

[39] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, J. Tygar, “Adversarial machine learning”, in: Proceedings of the 4th ACM workshop on Security and artificial intelligence, ACM, 2011, pp. 43–58.

[40] R. Naveiro, A. Redondo, D. R. Insua, F. Ruggeri, “Adversarial classification: An adversarial risk analysis approach”, arXiv preprint arXiv:1802.07513(2018).

[41] D. Lowd, C. Meek, “Adversarial learning”, in: 11th ACM SIGKDD international conference on Knowledge discovery in data mining, ACM, 2005, pp.641–647.

[42] A. Demontis, P. Russu, B. Biggio, G. Fumera, F. Roli, “On security and sparsity of linear classifiers for adversarial settings”, in: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, 2016, pp. 322–332.

[43] Y. Zhou, M. Kantarcioglu, B. Thuraisingham, B. Xi, “Adversarial support vector machine learning”, in: 18th ACM SIGKDD International conference on Knowledge discovery and data mining, ACM, 2012, pp. 1059–1067.

[44] B. Biggio, I. Corona, Z.-M. He, P. P. Chan, G. Giacinto, D. S. Yeung, F. Roli, “One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time”, in: International Workshop on Multiple Classifier Systems, Springer, 2015, pp. 168–180.

[45] A. N. Bhagoji, D. Cullina, C. Sitawarin, P. Mittal, “Enhancing robustness of machine learning systems via data transformations”, in: Information Sciences and Systems (CISS), 2018 52nd Annual Conference on, IEEE, 2018, pp. 1–5.

[46] J. Hayes, G. Danezis, “Machine learning as an adversarial service: Learning black-box adversarial examples”, arXiv preprint arXiv:1708.05207 (2017).

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, “Generative adversarial nets, in: Advances in neural information processing systems”, 2014, pp. 2672–2680.

[48] A. Nguyen, J. Yosinski, J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images”, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436

[49] N. Papernot, P. McDaniel, I. Goodfellow, “Transferability in machine learning: from phenomena to black-box attacks using adversarial samples”, arXiv preprint arXiv:1605.07277 (2016).

[50] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, D. Song, “Robust physical-world attacks on machine learning models”, arXiv preprint arXiv:1707.08945 (2017).

[51] N. Papernot, P. McDaniel, A. Swami, R. Harang, “Crafting adversarial input sequences for recurrent neural networks”, in: Military Communications Conference, MILCOM 2016-2016 IEEE, IEEE, 2016, pp. 49–54.

[52] A. Radford, L. Metz, S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks”, arXiv preprint arXiv:1511.06434 (2015).

[53] L. Demetrio, B. Biggio, G. Lagorio, F. Roli, A. Armando, “Explaining vulnerabilities of deep learning to adversarial malware binaries”, arXiv preprint arXiv:1901.03583 (2019)

[54] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, C. K. Nicholas, “Malware detection by eating a whole exe”, in: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[55] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, “Towards deep learning models resistant to adversarial attacks”, arXiv preprint arXiv:1706.06083 (2017).

[56] P. Schöttle, A. Schlögl, C. Pasquini, R. Böhme, “Detecting adversarial examples-a lesson from multimedia forensics”, arXiv preprint arXiv:1803.03613 (2018).

[57] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, “Boosting adversarial attacks with momentum”, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9185–9193.

[58] W. Uther, M. Veloso, “Adversarial reinforcement learning”, Technical Report, Tech. rep., Carnegie Mellon University. Unpublished, 1997.

[59] Tabassi, Elham, et al. "A Taxonomy and Terminology of Adversarial Machine Learning." (2019).

[60] He, Yingzhe, et al. "Towards Privacy and Security of Deep Learning Systems: A Survey." arXiv preprint arXiv:1911.12562 (2019).

[61] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. “Stealing machine learning models via prediction apis”. In 25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016., pages 601–618, 2016.

[62] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. “Practical black-box attacks against machine learning”. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS, Abu Dhabi, United Arab Emirates, April 2-6, 2017, pages 506–519.

[63] B. Wang and N. Z. Gong. “Stealing hyperparameters in machine learning”. In IEEE Symposium on Security and Privacy (SP), San Francisco, California, USA, 21-23 May 2018, pages 36–52, 2018.

[64] S. J. Oh, M. Augustin, M. Fritz, and B. Schiele. “Towards reverse engineering black-box neural networks”. In International Conference on Learning Representations, 2018.

[65] M. Juuti, S. Szyller, A. Dmitrenko, S. Marchal, and N. Asokan. “PRADA: protecting against DNN model stealing attacks”. CoRR, abs/1805.02628, 2018.

[66] T. Orekondy, B. Schiele, and M. Fritz. “Knockoff nets: Stealing functionality of black-box models”. June 2019.

[67] J. R. C. da Silva, R. F. Berriel, C. Badue, A. F. de Souza, and T. Oliveira-Santos. “Copycat CNN: stealing knowledge by persuading confession with random non-labelled data”. In International Joint Conference on Neural Networks, IJCNN, Rio de Janeiro, Brazil, pages 1–8, July 8-13, 2018.

[68] Han Xiao, Huang Xiao, and Claudia Eckert. “Adversarial label flips attack on support vector machines”. In: Frontiers in Artificial Intelligence and Applications 242.4 (2012), pp. 870–875. ISSN: 09226389. DOI: 10.3233/978-

[69] J. Hamm, Y. Cao, and M. Belkin. “Learning privately from multiparty data”. In Proceedings of the 33nd International Conference on Machine Learning, ICML, New York City, NY, USA, pages 555– 563, June 19-24, 2016.

[70] C. Song, T. Ristenpart, and V. Shmatikov. “Machine learning models that remember too much”. In Proceedings of ACM SIGSAC Conference on Computer and Communications Security, CCS

[71] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici. “Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers”. IJSN, 10(3):137–150, 2015.

[72] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov. “Property inference attacks on fully connected neural networks using permutation invariant representations”. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, CCS, Toronto, ON, Canada, pages 619–633, October 15-19, 2018.

[73] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov. “Inference attacks against collaborative learning”. CoRR, abs/1805.04049, 2018

[74] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in Computer vision–ECCV 2014. Springer, 2014, pp. 818–833.

[75] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei. “Towards demystifying membership inference attacks”. CoRR, abs/1807.09173, 2018.

[76] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes. “Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models”. CoRR, abs/1806.01246, 2018.

[77] Y. Long, V. Bindschaedler, L.Wang, D. Bu, X.Wang, H. Tang, C. A. Gunter, and K. Chen. “Understanding membership inferences on well-generalized learning models”. CoRR, abs/1802.04889, 2018.

[78] B. Hitaj, G. Ateniese, and F. P´erez-Cruz. “Deep models under the GAN: information leakage from collaborative deep learning”. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, CCS, Dallas, TX, USA, pages 603–618, October 30-November 03, 2017.

[79] K. S. Liu, B. Li, and J. Gao. “Generative model: Membership attack, generalization and diversity”. CoRR, abs/1805.09898, 2018.

[80] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro. “Knock knock, who’s there? membership inference on aggregate location data”. 2017.

[81] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. “Membership inference attacks against machine learning models”. In 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017, pages 3–18, 2017.

[82] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” Pattern Recognition, vol. 84, pp. 317–331, 2018.

[83] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro. “LOGAN: evaluating privacy leakage of generative models using generative adversarial networks”. CoRR, abs/1705.07663, 2017.

[84] Dwork, Cynthia. "Differential Privacy: A Survey of Results." TAMC. Vol. 4978. 2008.

[85] ENISA Glossary. Available at: https://www.enisa.europa.eu/topics/threat-risk-management/risk-management/current-risk/risk-management-inventory/glossary

[86] Department of Homeland Security. “Cyber Threat Modelling: Survey, Assessment, and Representative Framework”. April 7, 2018. Authors: Deborah J. Bodeau, Catherine D. McCollum, David B. Fox

[87] Mockel, C. & Abdallah, A. (2010). “Threat modeling approaches and tools for securing architectural designs of an e-banking application”. Sixth International Conference on Information Assurance and Security (IAS), 149-154, doi: 10.1109/ISIAS.2010.5604049.

[88] Shostack, A. (2008). “Experiences threat modelling at Microsoft”. Modelling Security Workshop. Dept. of Computing, Lancaster University, UK.

[89] Souppaya, M. & Scarfone, K. 2016. “Guide to Data-Centric System Threat Modelling” (NIST Special Publication 800-154). Gaithersburg: National Institute of Standards and Technology

[90] NIST Special Publication 800-30, “Guide for Conducting Risk Assessments”. 2012. Gaithersburg: National Institute of Standards and Technology

[91] Schneier, B. (1999). “Modelling security threats: Attack Trees”. Dr. Dobb’s Journal, 1–9

[92] Sjouke, M. & Oostdijk, M. (2006). “Foundations of attack trees”. 8th Annual International Conference on Information Security and Cryptology (ICISC'05), 186-198.

[93] McDermott, J. (2001). “Attack net penetration testing”. Proceedings of the 2000 workshop on New Security Paradigms, 15-21.

[94] Potteiger, B. Martins, G., and Koutsoukos, X. 2016. “Software and Attack Centric Integrated Threat Modeling for Quantitative Risk Assessment,” Hot Topics in Science of Security Symposium (HotSoS ’16), April 19-21, 2016.

[95] Shostack, A. 2014. “Threat Modelling: Designing for Security”. Indianapolis: John Wiley & Sons, Inc.

[96] Kordy, B., Piètre-Cambacédès, L., & Schweitzer. P., (2014). “DAG-based attack and defense modelling: Don’t miss the forest for the attack trees”. Computer science review, 13, 1-38.

[97] Wagner, C., Dulaunoy, A., Wagener, G., & Iklody, A. (2016, October). “Misp: The design and implementation of a collaborative threat intelligence sharing platform”. In Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security (pp. 49-56).

[98] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, “Towards Deep Learning Models Resistant to Adversarial Attacks”, 2019, https://arxiv.org/abs/1706.06083.

[99] Alexey Kurakin, Ian Goodfellow, Samy Bengio, “Adversarial examples in the physical world”, In ICLR Workshop, 2017. https://arxiv.org/abs/1607.02533.

[100] Battista Biggio, Ignazio Pillai, Samuel Rota Bulo, Davide ` Ariu, Marcello Pelillo, and Fabio Roli. “Is data clustering in adversarial settings secure?”, (2013).

[101] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., ... & Roli, F. (2013, September). “Evasion attacks against machine learning at test time”. In Joint European conference on machine learning and knowledge discovery in databases (pp. 387-402). Springer, Berlin, Heidelberg.

[102] Biggio, B., Nelson, B., & Laskov, P. (2011, November). “Support vector machines under adversarial label noise”. In Asian conference on machine learning (pp. 97-112).

[103] Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text”. 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.

[104] Chen, H., Zhang, H., Chen, P. Y., Yi, J., & Hsieh, C. J. (2017). “Attacking visual language grounding with adversarial examples: A case study on neural image captioning”. arXiv preprint arXiv:1712.02051.

[105] Chen, L., & Xu, W. (2020). “Attacking Optical Character Recognition (OCR) Systems with Adversarial Watermarks”. arXiv preprint arXiv:2002.03095.

[106] Chen, Yizheng, et al. "Practical attacks against graph-based clustering”. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017.

[107] D. Hitaj, Luigi V. Mancini, “Have you stolen my model? Evasion attacks against deep neural network watermarking techniques”, Sapienza University of Rome (2018).

[108] Globerson, A., & Roweis, S. (2006). “Nightmare at test time: robust learning by feature deletion”. Proceedings of the 23rd International Conference on Machine Learning - ICML ’06, 353–360. https://doi.org/10.1145/1143844.1143889.

[109] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". arXiv preprint arXiv:1412.6572 (2014).

[110] Grosse, Kathrin, et al. "Adversarial examples for malware detection". European Symposium on Research in Computer Security. Springer, Cham, 2017.

[111] Han Xiao, Huang Xiao, and Claudia Eckert. "Adversarial label flips attack on support vector machines". In: Frontiers in Artificial Intelligence and Applications 242.4 (2012), pp. 870–875. ISSN: 09226389. DOI: 10.3233/978-1-61499-098-7-870.

[112] Huang Xiao, Battista Biggio, Blaine Nelson, Han Xiao, Claudia Eckert, and Fabio Roli. "Support vector machines under adversarial label contamination". In: Neurocomputing 160 (2015), pp. 53–62. ISSN: 18728286. DOI: 10 .1016/j.neucom.2014.08.081.

[113] Huang, Sandy, et al. "Adversarial attacks on neural network policies". arXiv preprint arXiv:1702.02284 (2017).

[114] Kreuk, F., Adi, Y., Cisse, M., & Keshet, J. (2018, April). "Fooling End-To-End Speaker Verification With Adversarial Examples". In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1962-1966). IEEE.

[115] Lingxiao Wei, Yannan Liu, Bo Luo, Yu Li, and Qiang Xu. "I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators". In: (2018). arXiv: 1803.05847.

[116] Matt Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing". In: Proceedings of the 23rd USENIX Security Symposium (2014), pp. 17–32.

[117] McPherson, Richard, Reza Shokri, and Vitaly Shmatikov. "Defeating image obfuscation with deep learning". arXiv preprint arXiv:1609.00408 (2016).

[118] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey", in IEEE Access, vol. 6, pp. 14410-14430, 2018. doi: 10.1109/ACCESS.2018.2807385.

[119] N. Carlini and D. Wagner. "Towards evaluating the robustness of neural networks". IEEE Symposium on Security and Privacy, 2017.

[120] Tam N. Nguyen. "Attacking Machine Learning models as part of a cyber kill chain", North Carolina State University (2017).

[121] Yakura, Hiromu, and Jun Sakuma. "Robust Audio Adversarial Example for a Physical Attack". arXiv preprint arXiv:1810.11793 (2018).

[122] Yen Chen Lin, Zhang Wei Hong, Yuan Hong Liao, Meng Li Shih, Ming Yu Liu, and Min Sun. "Tactics of adversarial attack on deep reinforcement learning agents", (2017).

[123] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Authors Yingqi Liu, Weihang Wang, and Xiangyu Zhang. "Trojaning Attack on Neural Networks", Purdue University (2017).

[124] Zhu, Chen, W. Ronny Huang, Ali Shafahi, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. "Transferable clean-label poisoning attacks on deep neural nets". arXiv preprint arXiv:1905.05897 (2019).

[125] Zügner, Daniel, Amir Akbarnejad, and Stephan Günnemann. "Adversarial attacks on neural networks for graph data". Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018.

[126] Battista Biggio, Samuel Rota Bul`o, Ignazio Pillai, Michele Mura, Eyasu Zemene Mequanint, Marcello Pelillo, and Fabio Roli. “Poisoning Complete-Linkage Hierarchical Clustering”. In: ed. by Ana Fred, Terry M. Caelli, Robert P. W. Duin, Aur´elio C. Campilho, and Dick de Ridder. Vol. 3138. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, Aug. 2004. ISBN: 978-3-540-22570-6. DOI: 10. 1007 / b98738. arXiv: 9780201398298

[127] Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, and Fabio Roli. “Poisoning behavioral malware clustering”. In: Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop - AISec ’14. New York, USA: ACM Press, Nov. 2014, pp. 27–36. ISBN: 9781450331531. DOI: 10.1145/2666652.2666666.

[128] Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K. Jha. “Systematic poisoning attacks on and defenses for machine learning in healthcare”. In: IEEE Journal of Biomedical and Health Informatics 19.6 (2015), pp. 1893–1905. ISSN: 21682194. DOI: 10 . 1109/JBHI.2014.2344095.

[129] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning”. In: 2018 IEEE Symposium on Security and Privacy (SP). IEEE, May 2018, pp. 19–35. ISBN: 978-1-5386-4353-2. DOI: 10.1109/SP.2018.00057. arXiv: 1804.00308.

[130] Chang Liu, Bo Li, Yevgeniy Vorobeychik, and Alina Oprea. “Robust Linear Regression Against Training Data Poisoning”. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security - AISec ’17 (2017), pp. 91–102. DOI: 10.1145/3128572.3140447.

[131] Alex Beatson, Zhaoran Wang, and Han Liu. “Blind Attacks on Machine Learners”. In: 30th Conference on Neural Information Processing Systems (NIPS 2016) (2016), pp. 2397–2405. ISSN: 10495258

[132] Nicholas Carlini, Chang Liu, Jernej Kos, U´ lfar Erlingsson, and Dawn Song. “The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets”. In: (Feb. 2018). arXiv: 1802.08232.

[133] ENISA Threat taxonomy. Available at: https://commons.wikimedia.org/wiki/File:Threat_Taxonomy_Mind_Map.png

[134] ENISA recommendations on shaping technology according to GDPR provisions, p10. Available at: https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions-part-2

[135] Choraś, Michał, Marek Pawlicki, Damian Puchalski, and Rafał Kozik. "Machine Learning, the results are not the only thing that matters! What about security, explainability and fairness?." In International Conference on Computational Science, pp. 615-628. Springer, Cham, 2020.