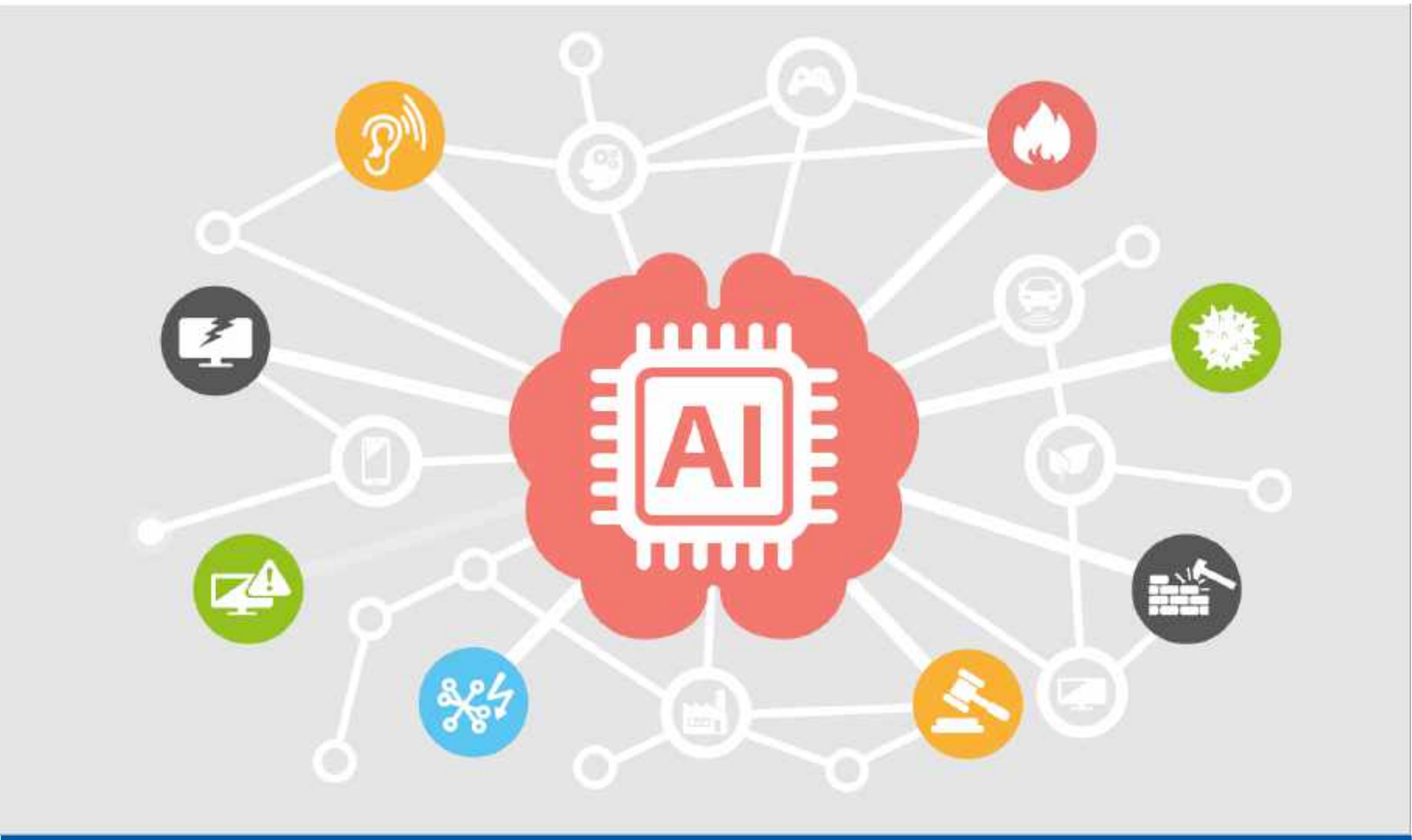
EUROPEAN UNION AGENCY
FOR CYBERSECURITY



AI CYBERSECURITY CHALLENGES

Threat Landscape for Artificial Intelligence

DECEMBER 2020

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies, and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure, and, ultimately, to keep Europe's society and citizens digitally secure. For more information, visit www.enisa.europa.eu.

CONTACT

For contacting the authors please use AIWG@enisa.europa.eu
For media enquiries about this paper, please use press@enisa.europa.eu

EDITORS

Apostolos Malatras, Georgia Dede – European Union Agency for Cybersecurity

ACKNOWLEDGEMENTS

We would like to thank the Members and Observers of ENISA ad hoc Working Group on Artificial Intelligence[1]:

Members:

- Caroline Baylon, AXA
- Christian Berghoff, Federal Office for Information Security Germany (BSI)
- Stephan Brunessaux, Airbus
- Luis Burdalo, S2 Grupo
- Giuseppe Dacquisto, Italian Data Protection Authority
- Ernesto Damiani, Universita degli Studi di Milano
- Sven Herpig, Stiftung Neue Verantwortung
- Caroline Louveaux, Mastercard
- Jochen Mistiaen, DigitalEurope
- Duy Cu Nguyen, Post Luxembourg
- Nineta Polemi, University of Piraeus
- Isabel Praca, Instituto Superior de Engenharia do Porto (ISEP)
- George Sharkov, Ministry of Defense Bulgaria and European Software Institute CEE
- Vincent Slieker, The National Cyber Security Center of the Netherlands
- Ewelina Szczekocka, Orange Polska SA

Observers:

- EC DG CONNECT
- EC DG JRC
- ETSI
- EU-LISA
- European Defence Agency
- Europol/EC3

[1] https://www.enisa.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence/adhoc_wg_calls

We would also like to thank EC DG JUSTICE, members of ENISA Advisory Group (AG) and ENISA National Liaison Officers (NLO) network for their valuable insights and comments.

LEGAL NOTICE

COPYRIGHT NOTICE

ISBN 978-92-9204-462-6 - DOI 10.2824/238222

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Artificial Intelligence (AI) is influencing people's everyday lives and playing a key role in digital transformation through its automated decision-making capabilities. The benefits of this emerging technology are significant, but so are the concerns. The EU Agency for Cybersecurity warns that AI may open new avenues in manipulation and attack methods, as well as new privacy and data protection challenges.

This report presents the Agency's active mapping of the AI cybersecurity ecosystem and its Threat Landscape, realised with the support of the Ad-Hoc Working Group on Artificial Intelligence Cybersecurity. The main highlights of the report include the following:

- Definition of the scope of AI in the context of cybersecurity following a lifecycle approach. Taking into account the different stages of the AI lifecycle from requirements analysis to deployment, the ecosystem of AI systems and applications is delineated.
- Identification of assets of the AI ecosystem as a fundamental step in pinpointing what needs to be protected and what could possibly go wrong in terms of security of the AI ecosystem.
- Mapping of the AI threat landscape by means of a detailed taxonomy. This serves as a baseline for the identification of potential vulnerabilities and eventually attack scenarios for specific use cases and thus serve in forthcoming sectorial risk assessments and listing of proportionate security controls.
- Classification of threats for the different assets and in the context of the diverse AI lifecycle stages, also listing relevant threat actors. The impact of threats to different security properties is also highlighted.

The ENISA AI Threat Landscape not only lays the foundation for upcoming cybersecurity policy initiatives and technical guidelines, but also stresses relevant challenges. One area of particular significance is that of the supply chain related to AI and accordingly it is important to highlight the need for an EU ecosystem for secure and trustworthy AI, including all elements of the AI supply chain. The EU secure AI ecosystem should place cybersecurity and data protection at the forefront and foster relevant innovation, capacity-building, awareness raising and research and development initiatives.

1. INTRODUCTION

Artificial Intelligence (AI) has gained traction over the last years facilitating intelligent and automated decision-making across as span of deployment scenarios and application areas. We are witnessing a convergence of different technologies (e.g. Internet of Things, robotics, sensor technologies, etc.) and growing amount and variety of data as well as their novel characteristics (e.g. distributed data) to employ AI at scale. In the context of cybersecurity, AI may be seen as an emerging approach and accordingly AI techniques have been used to support and automate relevant operations, e.g. traffic filtering, automated forensic analysis, etc. Whereas undoubtedly beneficial, one should not sidestep the fact that AI and its application to for instance automated decision making—especially in safety critical deployments such as in autonomous vehicles, smart manufacturing, eHealth, etc.—may expose individuals and organizations to new, and sometimes unpredictable, risks and it may open new avenues in attack methods and techniques, as well as creating new data protection challenges.

AI is increasingly influencing people's everyday lives and playing a key role in digital transformation through its automated decision-making capabilities. The benefits of this emerging technology are significant, but so are the concerns. It is thus necessary to highlight the role of cybersecurity in establishing the reliable and deployment of trustworthy AI.

When considering security in the context of AI, one needs to be aware that AI techniques and systems making use of AI may lead to unexpected outcomes and may be tampered with to manipulate the expected outcomes. This is particularly the case when developing AI software that is often based on fully black-box models[2], or it may even be used with malicious intentions, e.g. AI as a means to augment cybercrime and facilitate attacks by malicious adversaries. Therefore, it is **essential to secure AI itself. In particular, it is important:**

- **to understand what needs to be secured (the assets that are subject to AI-specific threats and adversarial models),**
- **to understand the related data governance models (including designing, evaluating and protecting the data and the process of training AI systems),**
- **to manage threats in a multi-party ecosystem in a comprehensive way by using shared models and taxonomies,**
- **to develop specific controls to ensure that AI itself is secure.**

Accordingly, securing AI is one of the areas on which ENISA will initially focus and this **threat landscape is the first effort to set the baseline for a common understanding on relevant cybersecurity threats.**

[2] Evidently white box models are also susceptible to cyber attacks because adversaries have widely accessible information to tailor attacks.

Artificial Intelligence and cybersecurity have a multi-dimensional relationship and a series of interdependencies. The dimensions that may be identified include the following three:

1. **Cybersecurity for AI:** lack of robustness and the vulnerabilities of AI models and algorithms, e.g. adversarial model inference and manipulation, attacks against AI-powered cyber-physical systems, manipulation of data used in AI systems, exploitation of computing infrastructure used to power AI systems' functionalities, data poisoning, environment variations which cause variations in the intrinsic nature of the data[3], credible and reliable training datasets, algorithmic validation/verification (including the integrity of the software supply chain), validation of training and performance evaluation processes, credible and reliable feature identification, data protection/privacy in the context of AI systems, etc.

2. **AI to support cybersecurity:** AI used as a tool/means to create advanced cybersecurity by developing more effective security controls (e.g. active firewalls, smart antivirus, automated CTI (cyber threat intelligence) operations, AI fuzzing, smart forensics, email scanning, adaptive sandboxing, automated malware analysis, automated cyber defence, etc.) and to facilitate the efforts of the law enforcement and other public authorities to better respond to cybercrime, including the analysis of the exponential growth of Big Data in the context of investigations, as well as the criminal misuse of AI.

3. **Malicious use of AI:** malicious/adversarial use of AI to create more sophisticated types of attacks, e.g. AI powered malware, advanced social engineering, AI-powered fake social media accounts farming, AI-augmented DDoS attacks, deep generative models to create fake data, AI-supported password cracking, etc. This category includes both AI-targeted attacks (focused on subverting existing AI systems in order to alter their capabilities), as well as AI-supported attacks (those that include AI-based techniques aimed at improving the efficacy of traditional attacks).

Cybersecurity can be one of the foundations of trustworthy Artificial Intelligence solutions. It will serve as a springboard for the widespread secure deployment of AI across the EU. However, it will do so only when common understanding of the relevant threat landscape and associated challenges are mapped in a consistent manner. **This report serves the purpose of setting the ground for defining the AI Threat Landscape**. The AI Threat Landscape is vast and dynamic, since it evolves alongside the innovations observed in the AI field and the continuous integration of numerous other technologies in the AI quiver.

1.1 POLICY CONTEXT

ENISA's WP2020 Output O.1.1.3 on Building Knowledge on Artificial Intelligence Security and the **European Commission White Paper on Artificial Intelligence[4]** have brought about the need for ENISA to look into the topic of AI Cybersecurity. The focus is mostly from the perspective of securing AI, but also looking into other aspects of AI and cybersecurity as mentioned above in a holistic and coordinated manner. In particular, the mapping of the AI Threat Landscape (AI TL) using threat modelling and assessment techniques has emerged as

[3] This refers to both physical attacks on AI systems as well as robustness of AI systems against naturally occurring variations and events.
[4] See EC White Paper on Ai under consultation at: https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

an important topic, as well as the drawing of proportionate security measures and recommendations[5].

Moreover, the European Commission (EC) has highlighted the importance of AI in society and the economy in its White Paper on Artificial Intelligence, which is the frontrunner to upcoming policy initiatives on the technology. The Commission has also recognised the strategic importance of AI in its "Coordinated Plan on Artificial Intelligence"[6], which aims to harmonise and coordinate AI initiatives across the Union, including addressing its security-related aspects. Additionally, in July 2020, the newly unveiled Security Union Strategy[7] of the European Commission underlined the significance of AI, noting that it will bring both new benefits and new risks.

In June 2018, the EC established the High-Level Expert Group on Artificial Intelligence (AI HLEG)[8] with the general objective to support the implementation of the European Strategy on Artificial Intelligence[9]. The AI HLEG has been looking into not only related policy developments, but also ethical, legal and societal aspects related to AI. Accordingly, the AI HLEG has put forward Policy and investment recommendations for trustworthy Artificial Intelligence[10], as well as Ethics Guidelines for Trustworthy AI[11] and an assessment list for trustworthy AI[12] including specific recommendations on assessing trustworthiness of AI systems.

In terms of policy context and relevant developments in the EU, it is noteworthy to mention the work of the European Defence Agency (EDA) that has developed a thorough taxonomy for AI[13] in the domain of defence. EDA's taxonomy is structured along three lines: algorithms, functions carried out by algorithms and support or related areas such as ethics, hardware implementation or learning techniques. Identifying the potential impact of AI in security and the interplay between the two domains, the European Telecommunication Standards Institute (ETSI) has set up an Industry Specification Group on Securing Artificial Intelligence (ISG SAI)[14]. The objective of the ISG SAI is to create standards to preserve and improve the security of new AI technologies. Additionally, the EC's Joint Research Centre (JRC) has established the AI Watch initiative in order to serve as a knowledge service to monitor the development, uptake and impact of artificial intelligence for Europe and monitor relevant research across the vast field of AI. One of the seminal works of the AI Watch is the report on defining AI[15] that sets the basis for relevant methodological conventions, introduces common vocabulary and more importantly common understanding of the diverse terms.

1.1.1 AI security under the Data Protection Prism

The General Data Protection Regulation (GDPR) establishes, under Article 5, security as a principle when processing personal data. This is an advanced role for security and an important conceptual shift from the past, when security was a mere technical-organizational provision on top of the processing operation. Under GDPR security is a pre-requisite, and not implementing

5 In doing so, ENISA will take stock of existing initiatives and studies that are ongoing in the area of AI cybersecurity such as the results of EU projects in this area (H2020) and will avoid duplication of efforts, rather focus on provide harmonized view of ongoing works.
6 See https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence
7 See https://ec.europa.eu/info/sites/info/files/communication-eu-security-union-strategy.pdf
8 See https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
9 See http://ec.europa.eu/digital-single-market/en/artificial-intelligence
10 See https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence
11 See https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines
12 See https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
13 See https://www.eda.europa.eu/info-hub/press-centre/latest-news/2020/08/25/artificial-intelligence-joint-quest-for-future-defence-applications
14 See https://www.etsi.org/technologies/securing-artificial-intelligence
15 See https://ec.europa.eu/jrc/en/publication/ai-watch-defining-artificial-intelligence

appropriate security measures invalidates the processing and makes it unlawful. Similar to the other GDPR data protection principles, security is not an option but a necessity.

Article 32 of the GDPR calls for security measures that must scale-up according to the risk of varying likelihood and severity for the rights and freedoms of data subjects. Therefore, the "asset" to protect with security measures is the unconstrained exercise of individuals' rights, and not only the informational asset per se. Personal data must be protected in a progressive way (the higher the risks, the stricter the measures). Security is a way to reinforce individuals' rights and freedoms as a whole and enables the centrality of humans vis-a-vis machines. AI systems are logic systems and as such, they may not be fully consistent and complete, meaning that humans will never be able to predict, upfront during the design phase, all the possible contextual factors that may impair their functioning. This exposes individuals to the inherent risks of unexpected outcomes where the outputs of an AI system are not properly constrained.

Security is also a data protection by design instrument, as envisaged in art 25 of the GDPR. Taking into account a number of contextual factors (like the state of the art) data controllers must put in place appropriate technical and organisational measures[16]. These measures must be in place to implement data-protection principles in an effective manner, from minimization to data accuracy, integrating the necessary safeguards into the processing. Notably, the GDPR mentions specifically pseudonymisation as one of those effective measures. The dimension of security for data protection, in the context of AI, is very important in order to be able to introduce the necessary technical and organisational safeguards (for the protection of rights and freedoms of individuals) already at the design phase of new AI applications.

To this end, security can also be an enabler of new types of processing operations, especially related to emerging technologies, such as AI. For instance, the implementation of specific security measures, like pseudonymisation or encryption, may bring data to a new format so that it cannot be attributed to a specific data subject without the use of additional information data (like a decryption key). These options could be explored in the context of AI environment, to shape new relationships between humans and machines, in a way that individuals are not by default identifiable by machines unless they wish to do so. For instance, to revertthe effect of the implemented pseudonymisation or encryption.

Putting security among the principles of data protection, as said, means that this is a precondition for the processing. However, a misinterpreted approach purely based on the assessment of economic risks might not foster the adoption of security measures, hindering the effective implementation of this principle. This phenomenon is broadly known[17] and may lead economic actors to bargain risks with investments, accepting information security risks (sometimes very high risks) on the assumption that security incidents are unlikely, and any investment that may just determine a reduction of an expected economic loss can always be procrastinated[18]. The GDPR offers a possible escape from this deadlock, raising security to the level of principles of data protection, and promotes security as a token of accountability at the largest possible scale, including the myriads of actors of the very complex AI value chain. The turning point is the shift from security as a defensive instrument to security as a functional element of digital ecosystem. Security should not be implemented only to prevent losses, but to create value. Only if security threats do not materialize, the AI ecosystem may generate trust, attract investments, retain users and create a positive feedback to develop every time new beneficial applications.

[16] This applies both at the time of the determination of the means for processing and at the time of the processing itself.
[17] Lawrence Gordon e Martin Loeb, *The Economics of Information Security Investment*, in ACM Transactions on Information and System Security, vol. 5, n. 4, November 2002, pp. 438–457, and Lawrence Gordon e Martin Loeb, *You May Be Fighting the Wrong Security Battles*, in the Wall Street Journal, 26 September 2011
[18] Such incidents may lead to severe risks that can compromise and impact data protection and individuals' fundamental rights and freedoms.

In this respect, two strategic choices could be explored concerning security:

- reflecting on the necessary functional role that security may have in the faultless functioning of an AI system, and on how to embed security since the very early stage of design (security by design) in order to create trust in new AI applications;
- considering the positive role that security certification may have in promoting a culture of security among economic actors, especially considering that certification may relieve the stakeholders from the complexity of the managerial decision between being idle and accepting security risks, or investing for reinforcing security.

Against this background, this report explores the AI security, considering also security as a data protection principle. While this dimension forms part of the threat landscape analysis discussed later in the report, it must be noted that the report strictly focuses on AI security and does not further address in any way data protection requirements and/or aspects of GDPR compliance in AI applications.

1.2 SCOPE & OBJECTIVES

In the context of the ENISA AI Threat Landscape (AI TL), the main focus of the work will be on cybersecurity of AI since secure AI is the foundation for any further work on AI. Only when AI itself is secure can we use it in a trustworthy manner and can we further utilise it for additional cybersecurity operations. Ethics of AI remain outside the scope of this work, since this was one of the focus areas of the EC High-Level Expert Group on AI (AI HLEG)[19].

According to the AI HLEG, "as a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)." It is therefore evident that the field of AI is vast and this is the reason for the necessity to scope the work in the context of the threat landscape.

Given that the driving force in terms of technologies nowadays is that of Machine Learning (ML), the main focus of the ENISA AI TL is on these technologies. Nonetheless, the work presented strives to also consider broader aspects of AI (e.g. data, infrastructure, algorithms, platforms, etc.) that are far more generic than ML and in this respect is representative of the wider AI ecosystem.

Moreover, sectorial domain-specific applications to AI, will not be considered in the AI TL per se. Sectorial approaches to the AI TL will need to be developed in the future to assess the likelihood and impact of threats for specific applications and to identify the risks that are specific to the context of use. This report of the ENISA AI TL aims at being horizontal and transversal, agnostic to application of domain of use.

In the context of the ENISA AI Threat Landscape, the focus of the work will be on cybersecurity of AI[20]. More specifically, the objectives of this work include:

- Identification, analysis and correlation of a list/taxonomy of assets (including interdependencies between assets) and respective asset owners. The identification will be performed through selected use cases that will embody/highlight key AI features and functionalities.

[19] See https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines, April 2019
[20] Due consideration to existing and ongoing pieces of work on cybersecurity for AI such as the ones by ETSI, ISC/IEC JTC 1/SC42, etc. in an effort to avoid duplication and base the work on already established knowledge sources.

- Identification and correlation of list of threats and vulnerabilities to be mapped against the list of assets mentioned above.
- Description of a representative set of attack scenarios/failure modes pertaining to core AI lifecycle stages.

1.3 METHODOLOGY

The method adopted for this study is in line with the methodology developed by ENISA for the preparation of its annual cyber Threat Landscape. According to this methodology, the process requires an initial identification of critical assets within the architecture before performing a threat assessment, which evaluates the different levels of asset exposure. Threats play a central role in a risk assessment, especially when considering the different components of risks. The ISO 27005[21], a widely adopted risk management standard, defines that risks emerge when: "Threats abuse vulnerabilities of assets to generate harm for the organisation".

Following this methodology, we have identified assets, threats and threat actors. Threats have been derived after analysing the functional behaviour of assets and pinpointing potential failure modes that represent the manifestation of threats. The combination of assets, threats and threat actors constitute the core of the AI Threat Landscape presented in this report[22]. The work has been conducted with the invaluable support of the ENISA ad hoc Working Group on Artificial Intelligence[23] (ahWGAI), which has provided feedback, insight and validation for the content of this report.

In the course of developing this work, the introduction of a new terminology was debated. However, given the noteworthy work of several fora we opted against introducing a new terminology that would replicate existing ones. Terms used throughout the document are following standard definitions based on the work of EC AI HLEG[24], EC JRC AI Watch[25], EDA[26], NIST[27], ETSI[28], SNV[29], MITRE[30], ISO[31] (e.g. ISO/IEC CD 22989.2, WD 5259-1), etc.

1.4 TARGET AUDIENCE

The target audience of this report includes a number of different stakeholders that are concerned by cybersecurity threats to AI systems. We have divided these stakeholders into the broad categories presented here.

- Public/governmental sector (EU, EU Institutions, European Commission, Member States regulatory bodies, supervisory authorities in the field of data protection, military and intelligence agencies, law enforcement community, international organisations and national cybersecurity authorities).
- Industry (including Small and Medium Enterprises) that makes use of AI solutions and/or is engaged in cybersecurity, including operators or essential services.
- AI technical community, AI cybersecurity experts and AI experts (designers, developers, machine learning experts, data scientists, etc.) with an interest in

[21] See https://www.iso.org/standard/75281.html
[22] The the identification and analysis of assets and cyber threats are based on work conducted by ENISA and the ENISA Ad hoc Working Group on AI (ahWGAI) based on the study of specifications, white papers and literature, without attempting any interpretation/evaluation of the assumptions stated in these reports.
[23] See https://www.enisa.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence/adhoc_wg_calls
[24] See https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines
[25] See https://publications.jrc.ec.europa.eu/repository/bitstream/JRC118163/jrc118163_ai_watch._defining_artificial_intelligence_1.pdf and https://publications.jrc.ec.europa.eu/repository/bitstream/JRC113826/ai-flagship-report-online.pdf
[26] See https://www.eda.europa.eu/info-hub/press-centre/latest-news/2020/08/25/artificial-intelligence-joint-quest-for-future-defence-applications
[27] See https://csrc.nist.gov/publications/detail/nistir/8269/draft
[28] See https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp34_Artificial_Intellignce_and_future_directions_for_ETSI.pdf
[29] See https://www.stiftung-nv.de/de/publikation/securing-artificial-intelligence
[30] See https://github.com/mitre/advmlthreatmatrix
[31] See ISO/IEC JTC 1/SC 42: https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0

developing secure solutions and in integrating security and privacy by design in their solutions.
- Cybersecurity community.
- Academia and research community.
- Standardization bodies.
- Civil society and general public.

1.5 STRUCTURE OF THE REPORT

After this Introduction, the report is structured as follows:

- Chapter 2 presents a generic reference model for the lifecycle of AI systems, in order to set the foundation for asset and processes identification.
- Chapter 3 details the assets in the AI ecosystem based on the lifecycle stages defined in Chapter 2 and categorizes them in 6 groups.
- Chapter 4 introduces the threat taxonomy of AI systems, where relevant threats are presented and mapped to corresponding assets that were introduced in Chapter 3.
- Chapter 5 concludes the report by highlighting cybersecurity-related challenges to AI and proposes high-level recommendations.

2. AI LIFECYCLE

In order to properly frame the domain of AI, it is essential to follow a structured and methodical approach to understand its different facets. For this reason, we opted towards deriving a lifecycle functional view of typical AI systems. Accordingly, this Chapter is structured around the different stages of the AI lifecycle and elaborates on the involved assets (e.g. actors, processes, artefacts, hardware, etc.), as the basis for threats identification[32] that follows in Chapter 4. Particular consideration is given to data protection in the context of AI, which is a horizontal concern that permeates all stages of the AI lifecycle.

The lifecycle of an AI system includes several interdependent phases ranging from its design and development (including sub-phases such as requirement analysis, data collection, training, testing, integration), installation, deployment, operation, maintenance, and disposal. Given the complexity of AI (and in general information) systems, several models and methodologies have been defined to manage this complexity, especially during the design and development phases, such as waterfall, spiral, agile software development, rapid prototyping, and incremental[33]. The AI lifecycle defines the phases that an organization should follow to take advantage of AI techniques and in particular of Machine Learning (ML) models to derive practical business value. For the purposes of this document, ML models are used to represent a mathematical transformation of the input data into a new result, e.g. use image input data to recognize faces. Conversely, algorithms are used to update the model parameters (training) or to discover patterns and relations in newly provided data and infer the result[34].

A disclaimer needs to be made on the focus of the reference model. Given the vast range and intricacies of techniques, technologies, algorithms and models involved in AI systems, mapping their entirety in a sole AI lifecycle model is not possible. The particularities of AI systems and the many sub-fields of AI (e.g. reasoning systems, robotics, connectionist vs symbolic AI, etc.) would require the generation of targeted reference models based on the used technology. Given the current prominence of Machine Learning (ML) in the use and deployment of AI systems, we opted to gear the AI lifecycle reference model towards ML in order to on the one hand make it specific and detailed, and on the other hand address the majority of current AI systems. ML has been spearheading the explosion of AI in the last ten years regarding image- and voice-identification. Future work will ensure that the ENISA AI TL will expand to cover the other sub-fields of AI to ensure complete coverage.

Based on desktop research[35], a generic reference model of various components found in common AI systems was drafted and is depicted in Figure 1. The purpose of having a reference model is to establish a conceptual framework ensuring shared understanding of the assets composing an AI system and their significant relationships. This facilitates the assignment of owners to different assets on one hand and on the other hand provides a systematic, structured way of analysing relevant security threats. Provided that assets have been defined, threats to AI

---

[32] In this document we consider the data sources for AI have been protected and are considered to be secure. In our approach, the AI application life cycle (in short, AI Lifecycle) is considered as a generic model for the foundation of assets and threats identification, and not intended as a statement. Feedback loops presented are not exhaustive as different use cases might follow different pipelines. and omit some of the phases of the generic life cyle.. Mind maps were included as a first step towards a complete reference model.

[33] See https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

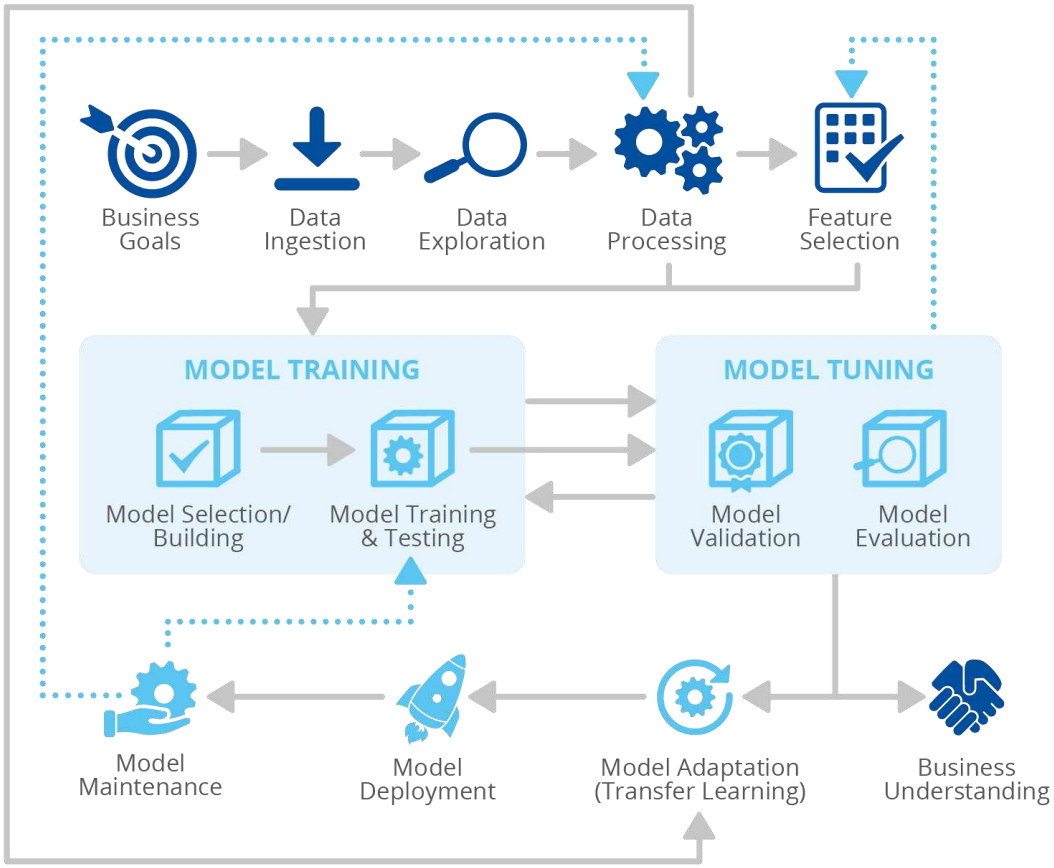
[34] The threat landscape assumes basic understanding of AI terminology and concepts. For further details and to gain deeper understanding the interested reader is referred to AI tectbook material.

[35] Including already referenced work from EC JRC, EC AI HLEG, EDA, ETSI ISG SAI, NIST, Stiftung Neue Verantwortung, Mcrosoft (https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml), Berryville Institute of Machine Learning (https://berryvilleiml.com /) and BSI (https://doi.org/10.3389/fdata.2020.00023).

systems can be mapped against these assets and following that targeted security measures to the corresponding asset owners may be delivered.

2.1 AI LIFECYCLE

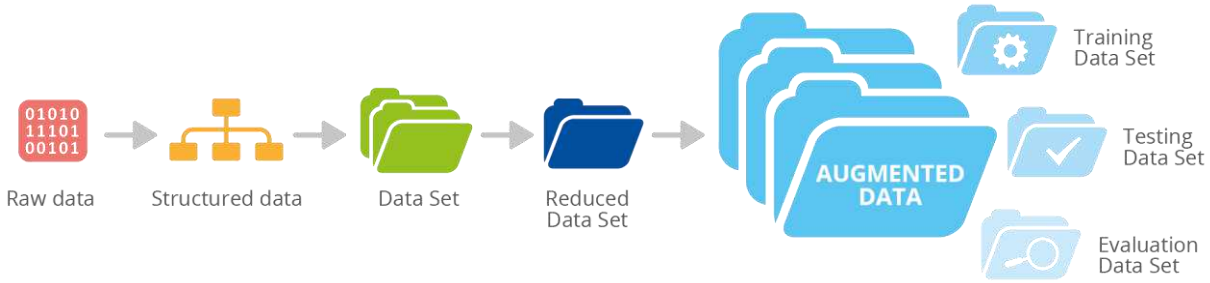
Figure 1: AI lifecycle generic reference model



Data is one of the most valuable assets in Artificial Intelligence; it is being continuously transformed along the AI Lifecycle[36]. Figure 2 illustrates data transformation along the different lifecycle stages: Data Ingestion, Data Exploration, Data Pre-processing, Feature Importance, Training, Testing and Evaluation. Data transformation along the AI Lifecycle involves several other types of assets, like the involved actors, computational resources, software, etc. and even some non-tangible assets like processes, culture and the way actors' experience and knowledge can bring potential non-intentional threats (e.g. non-intentional bias).

[36] In terms of data categories and the provenance of the data we distinguish between the following.
Self-reported data, voluntarily provided by a "trustful" operator (e.g. AIS for a ship or ADS-B for an aircraft, cooperative and governmental data).
Observed data collected by active or passive "secure" systems (e.g. IDSs, sensors, RFIDs, cameras, IoTs in general, radars), the integrity of the data depends upon a variety of parameters (resolution, range, refresh, latency, environmental conditions, size, orientation, electromagnetic characteristics).
Information registers and databases: they contain information linking data (aircrafts or ship IDs, human IDs from civil legacy systems, smart objects IDs from industries) with details on its structure, construction, appearance, history and interactions, activity, Social Media from Free and Open Internet Sources (e.g Twitter, Youtube, Facebook, WhatsApp, Media, Open DB) are also included in this category.

Figure 2: Data transformation along AI Lifecycle development stages



In what follows, we describe in detail the different stages of the AI lifecycle by giving emphasis to the different assets, processes and actors that are involved[37], as well as discussing the relevant data transformations.

2.2 AI LIFECYCLE ACTORS

There are various actors actively engaged in the context of the entire AI Lifecycle. Actors include the **AI designers / AI application designers** involved in the design and creation of AI systems. There are also the **AI developers** that develop and build the software and algorithms used in AI systems, as well as work to refine and enhance them. Their experience and capacity plays a key role in the development of secure AI systems.

AI developers and designers work closely with **data scientists**. Data scientists' work might involve helping design and develop AI models, or it can consist of using such models and analysing the results. More specifically, data scientists are involved in collecting and interpreting data, focusing on extracting knowledge and insights from that data. Other actors in the AI lifecycle are **data engineers**, whose work primarily involves extracting and collating data from different sources, then transforming, cleaning, standardizing, and storing it. Data engineers mainly focus on the design, management, and optimization of the flow of data.

Other important actors in the AI lifecycle are **data owners**[38]. Data owners own the datasets that are used to either train/validate AI systems or that these systems use to perform tasks. They are often businesses, who have their own datasets linked to their business that they provide an AI system with in order to carry out a task on their behalf. Data owners can also be **data providers / data brokers**. These are third parties that monetize data used by AI systems, either for training purposes or to perform various tasks. They might include commercial data brokers, which collect, store, and sell various types of data, in a legal manner. There are also reports of shadow data brokers that gather data about users without them being aware that their personal data is being collected, stored, and sold[39].

Other AI lifecycle actors include **model providers**, who provide models (as well as implementations of them in the form AI/ML libraries) that have already been trained and fine-tuned. Some model providers are **cloud providers**, which offer the models as a service, notably the use of AI-based computational and data analyses capabilities in the cloud. Besides model providers, other actors involve **third-party providers** who may also provide third-party software

[37] The reference model detals the typical, different phases of the AI lifecycle. A noteworthy reference needs to be made to automated machine learning solutions (offered by several vendors) that encompass the vast majority of the AI lifecycle stages to facilitate product developers. Despite numerous research and commercial initiatives for developing efficient automated machine learning mechanisms and tools, many challenges have been identified including transparency issues (black-box operation), limited reproducibility, etc.
[38] Please note that in the case of personal data the role of data owners is equivalent to that of data controllers.
[39] Evidently if such cases occur, then there is a clear lack of compliance to GDPR provisions and further legal analysis (outside the scope of this work) is highly recommended.

frameworks and libraries, which developers can use for training AI systems, and specialised high-performance hardware.

Finally, there are **the end users** that make use of AI systems, including **service consumers**. This might be companies, many of which are **model users**. They also include consumers and the general public. End users might also be users of other AI systems as well.

2.3 AI LIFECYCLE PHASES

In this section, we provide a short definition for each stage of the AI Lifecycle and recap the individual steps it involves ("Phase in a Nutshell").

2.3.1 Business Goal Definition

Prior to carrying out any AI application/system development, it is important that the user organization fully understand the business context of the AI application/system and the data required to achieve the AI application's business goals, as well as the business metrics to be used to assess the degree to which these goals have been achieved.

Business Goal Definition Phase in a Nutshell: Identify the business purpose of the AI application/system. Link the purpose with the question to be answered by the AI model to be used in the application/system. Identify the model type based on the question.

2.3.2 Data Ingestion

Data Ingestion is the AI life cycle stage where data is obtained from multiple sources (raw data may be of any form structured or unstructured) to compose multi-dimensional data points, called vectors, for immediate use or for storage in order to be accessed and used later. Data Ingestion lies at the basis of any AI application. Data can be ingested directly from its sources in a real-time fashion, a continuous way also known as streaming, or by importing data batches, where data is imported periodically in large macro-batches or in small micro-batches.

Different ingestion mechanisms can be active simultaneously in the same application, synchronizing or decoupling batch and stream ingestion of the same data flows. Ingestion components can also specify data annotation, i.e. whether ingestion is performed with or without metadata (data dictionary, or ontology/taxonomy of the data types). Often, access control operates during data ingestion modelling the privacy status of the data (personal / non-personal data.), choosing suitable privacy preserving techniques and taking into account the achievable trade-off between privacy impact and analytic accuracy. Compliance with applicable EU privacy and data protection legal framework needs to be ensured in all cases.

The privacy status assigned to data is used to define the AI application Service Level Agreement (SLA) in accordance with the applicable EU privacy and data protection legal framework, including –among other things- the possibility for inspection / auditing competent regulatory authorities (such as Data Protection Authorities). It is important to remark that, in ingesting data an IT governance conflict may arise. On the one hand, data is compartmentalized by its owners in order to ensure access control and privacy protection; on the other hand, it must be integrated in order to enable analytics. Often, different policies and policy rules apply to items of the same category. For multimedia data sources, access protocols may even follow a Digital Right Management (DRM) approach where proof-of-hold must be first negotiated with license servers. It is the responsibility of the AI application designer to make sure that ingestion is done respecting the data providers' policies on data usage and the applicable EU privacy and data protection legal framework.

Data Collection/Ingestion Phase in a Nutshell: Identify the input (dynamic) data to be collected and the corresponding context metadata. Organize ingestion according to the AI application requirements, importing data in a stream, batch or multi-modal fashion.

2.3.3 Data Exploration

Data Exploration is the stage where insights start to be taken from ingested data. While it may be skipped in some AI applications where data is well understood, it is usually a very time-consuming phase of the AI life cycle. At this stage, it is important to understand the type of data that were collected. A key distinction must be drawn between the different possible types of data, with numerical and categorical being the most prominent categories[40], alongside multimedia data (e.g. image, audio, video, etc.)[41]. Numerical data lends itself to plotting and allows for computing descriptive statistics and verifying if data fits simple parametric distributions like the Gaussian one. Missing data values can also be detected and handled at the exploration stage. Categorical variables are those that have two or more categories but without an intrinsic order. If the variable has a clear ordering, then it is considered as an ordinal variable.

Data Validation/Exploration in a Nutshell: Verify whether data fit a known statistics distribution, either by component (mono-variate distributions) or as vectors (multi-variate distribution). Estimate the corresponding statistic parameters.

2.3.4 Data Pre-processing

The data pre-processing stage employs techniques to cleanse, integrate and transform the data. This process aims at improving data quality that will improve performance and efficiency of the overall AI system by saving time during the analytic models' training phase and by promoting better quality of results. Specifically, the term data cleaning designates techniques to correct inconsistencies, remove noise and anonymize/pseudonymise data.

Data integration puts together data coming from multiple sources, while data transformation prepares the data for feeding an analytic model, typically by encoding it in a numerical format. A typical encoding is *one-hot encoding* used to represent categorical variables as binary vectors. This encoding first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the position of the integer, which is marked with a 1.

Once converted to numbers, data can be subject to further types of transformation: re-scaling, standardization, normalization, and labelling[42]. At the end of this process, a numerical data set is obtained, which will be the basis for training, testing and evaluating the AI model.

Since having a large enough dataset is one of the key success factors when properly training a model, it is common to apply different data augmentation techniques to those training datasets that are too small. For instance, it is common to include in a training dataset different scaled or rotated versions of images, which were already in that dataset. Another example of data augmentation technique which can be used when processing text is replacing a word by its synonym. Even in those cases in which the training dataset is large enough, data augmentation techniques can improve the final trained model. Data can also be augmented in order to increase its quantity and the diversity of scenarios covered. Data augmentation usually consists in applying transformations which are known to be label-preserving, i. e. the model should not change its output (namely prediction) when presented with the transformed data items. Data augmentation can serve to improve the performance of a model and in particular its robustness to benign perturbations. One task where data augmentation is used by default is image

[40] The discussion mainly refers to nunerical, tabular data. It needs nevertheless to be mentioned that AI systems may also use other types of data, e.g. speech, images. These are also numerical, but sanity checks have an advanced degree of complexity, for which no data exploration as described here is performed.
[41] Multimedia data are complex data that are very relevant in the context of deep learning.
[42] Re-scaling is used to make sure all variables are expressed on the same scale, as some methods may overlook variables with lower intensity. Standardization is used to change the mean of a distribution of values to 0, while normalization maps data to a compact representation interval (e.g., the interval (0, 1), by dividing all values by the maximum). Labelling (done by human experts or by other AI applications) associates each data item to a category or a prediction.

classification, where data can be augmented by for instance applying translations, rotations and blurring filters.

Data pre-processing in a Nutshell: Convert ingested data to a metric (numerical) format, integrate data from different sources, handle missing/null values by interpolation, densify to reduce data sparsity, de-noise, filter outliers, change representation interval, anonymize/pseudonymize data, augment data.

2.3.5 Feature Selection

Feature Selection (in general feature engineering) is the stage where the number of components or features (also called dimensions) composing each data vector is reduced, by identifying the components that are believed to be the most meaningful for the AI model[43]. The result is a reduced dataset, as each data vector has fewer components than before[44]. Besides the computational cost reduction, feature selection can bring more accurate models. Additionally, models built on top of lower dimensional data are more understandable and explainable. This stage can also be embedded in the model building phase (for instance when processing image or speech data), to be discussed in the next section.

Feature selection in a Nutshell: Identify the dimensions of the data set that account for a global parameter, e.g. the overall variance of the labels. Project data set along these dimensions, discarding the others.

2.3.6 Model Selection / Building

This stage performs the selection/building of the best AI model or algorithm[45] for analysing the data. It is a difficult task, often subject to trial and error. Based on the business goal and the type of available data, different types of AI techniques can be used. The three commonly identified major categories are supervised learning, unsupervised learning and reinforcement learning models. Supervised techniques deal with labelled data: the AI model is used to learn the mapping between input examples and the target outputs.

Supervised models can be designed as Classifiers, whose aim is to predict a class label, and *Regressors*, whose aim is to predict a numerical value function of the inputs. Here some common algorithms are Support Vector Machines, Naïve Bayes, Hidden Markov Model, Bayesian networks, and Neural Networks.

Unsupervised techniques use unlabelled training data to describe and extract relations from it, either with the aim of organizing it into clusters, highlight association between data input space, summarize the distribution of data, and reduce data dimensionality (this topic was already addressed as a preliminary step for data preparation in the section on feature selection). Reinforcement learning maps situations with actions, by learning behaviours that will maximize a desired reward function.

While the type of training data, labelled or not, is key for the type of technique to be used and selected, models may also be built from scratch (although this is rather unlikely), with the data scientist designing and coding the model, with the inherent software engineering techniques; or building a model by combining a composition of methods[46]. It is important to remark that model selection (namely choosing the model adapted to the data) may trigger further transformation of

[43] Machine Learning Models are algorithms trained with historical data that discover patterns and relations, and construct mathematical models using these discoveries.
[44] It is noteworthy that this is not always the case. In particular, in recent deep learning approaches that consider end-to-end deep learning approaches, where no feature processing is done.
[45] Stuart J. Russell and Peter Norvig, "Artificial Intelligence: A Modern Approach", Prentice Hall Press. ISBN:978-0-13-604259-4
[46] By composition of methods we refer to model ensembling that consists in combining the outputs of multiple models to take advantage of the advantages of different approaches, at the cost of a greater complexity.

the input data, as different AI models require different numerical encodings of the input data vectors.

Generally speaking, selecting a model also includes choosing its training strategy. In the context of supervised learning for example, training involves computing (a *learning function* of) the difference between the model's output when it receives each training set data item *D* as input, and *D*'s label. This result is used to modify the model in order to decrease the difference.

Many training algorithms for error minimization are available, most of them based on gradient descent. Training algorithms have their own hyperparameters, including the function[47] to be used to compute the model error (e.g. mean squared error), and the batch size, i.e. the number of labelled samples to be fed to the model to accumulate a value of the error to be used for adapting the model itself.

AI Model Selection in a Nutshell: Choose the type of AI model suitable for the application. Encode the data input vectors to match the model's preferred input format.

2.3.7 Model Training

Having selected an AI model, which in the context of this reference model mostly refers to a Machine Learning (ML) model, the training phase of the AI system commences. In the context of supervised learning, the selected ML model must go through a training phase, where internal model parameters like weights and bias are learned from the data. This allows the model to gain understanding over the data being used and thus become more capable in analysing them. Again, training involves computing (a function of) the difference between the model's output when it receives each training set data item *D* as input, and *D*'s label. This result is used to modify the model in order to decrease the difference between inferred result and the desired result and thus progressively leads to more accurate, expected results.

The training phase will feed the ML model with batches of input vectors and will use the selected learning function to adapt the model's internal parameters (weights and bias) based on a measure (e.g. linear, quadratic, log loss) of the difference between the model's output and the labels. Often, the available data set is partitioned at this stage into a training set, used for setting the model's parameters, and a test set, where evaluation criteria (e.g. error rate) are only recorded in order to assess the model's performance outside the training set. Cross-Validation schemes randomly partition multiple times a data set into a training and a test portion of fixed sizes (e.g. 80% and 20% of the available data) and then repeat training and validation phases on each partition.

AI Model Training in a Nutshell: Apply the selected training algorithm with the appropriate parameters to modify the chosen model according to training data. Validate the model training on test set according to a cross validation strategy.

2.3.8 Model Tuning

Model tuning usually overlaps with model training, since tuning is usually considered within the training process. We opted to separate the two stages in the AI lifecycle to highlight the differences in terms of functional operation, although it is most likely that in the majority of the AI systems they will be both part of the training process.

Certain parameters define high level concepts about the model, such as their learning function or modality, and cannot be learned from input data. These special parameters, often called

[47] In deep learning where possibly highly complex loss functions are designed, and are a key element of the training process.

hyper-parameters, need to be setup manually, although they can under certain circumstances be tuned automatically by searching the model parameters' space[48]. This search, called *hyper-parameter optimization*[49], is often performed using classic optimization techniques like *Grid Search*, but *Random Search* and *Bayesian optimization* can be used. It is important to remark that the Model Tuning stage uses a special data set (often called validation set), distinct from the training and test sets used in the previous stages. An evaluation phase can also be considered to estimate the outputs limits and to assess how the model would behave in extreme conditions, for example, by using wrong/unsafe data sets. It is important to be noted that, depending on the number of hyper-parameters to be adjusted, trying all possible combinations may just not be feasible.

AI Model Tuning in a Nutshell: Apply model adaptation to the hyper-parameters of the trained AI model using a validation data set, according to deployment condition.

2.3.9 Transfer Learning

In this phase, the user organization sources a pre-trained and pre-tuned AI model and uses it as starting point for further training to achieve faster and better convergence. This is commonly the case when few data are available for training. It should be noted that all steps described above (tuning, testing, etc.) also apply for transfer learning. Moreover, since in many cases transfer learning is being applied, one can consider transfer learning as a part of model training phase, given that transfer learning usually serves as a starting point of the training algorithm. To ensure wider scope, we distinguish transfer learning into a distinct phase in the AI lifecycle presented here.

Transfer Learning in a Nutshell: Source a pre-trained AI model in the same application domain, and apply additional training to it, as needed to improve its in-production accuracy.

2.3.10 Model Deployment

A Machine Learning model will bring knowledge to an organization only when its predictions become available to final users. Deployment is the process of taking a trained model and making it available to the users.

Model Deployment in a Nutshell: Generate an in-production incarnation of the model as software, firmware or hardware. Deploy the model incarnation to edge or cloud, connecting in-production data flows.

2.3.11 Model Maintenance

After deployment, AI models need to be continuously monitored and maintained to handle *concept changes* and potential *concept drifts* that may arise during their operation. A change of concept happens when the meaning of an input to the model (or of an output label) changes, e.g. due to modified regulations. A concept drift occurs when the change is not drastic but emerges slowly. Drift is often due to sensor *encrustment*, i.e. slow evolution over time in sensor resolution (the smallest detectable difference between two values) or overall representation interval. A popular strategy to handle model maintenance is *window-based relearning*, which relies on recent data points to build a ML model. Another useful technique for AI model maintenance is *back testing*. In most cases, the user organization knows what happened in the aftermath of the AI model adoption and can compare model prediction to reality. This highlights concept changes: if an underlying concept switches, organizations see a decrease of performance. Another way of detecting these concept drifts may involve statistically

[48] Re-tuning of hyper-paratameters is often a challenging task given that the space of hyper-parameters is usually immense and the process requires a large amount of time and computing resources. Moreover, it needs to be noted that this type of tuning requires frequent re-training of the model.
[49] It should be noted that this process is very expensive computationally, and tends to be limited, especially in deep learning applications where training may take days or weeks.

characterizing the input dataset used for training the AI model, so that it is possible to compare this training dataset to the current input data in terms of statistic properties. Significant differences between datasets may be indicative of the presence of potential concept drifts which may require a relearning process to be carried out, even before the output of the system is significantly affected. In this way, retraining/relearning processes, which may be potentially time and resource consuming, can be carried out only when required instead of periodically, like in the above mentioned window-based relearning strategies. Model maintenance also reflects the need to monitor the business goals and assets that might evolve over time and accordingly influence the model itself.

Model Maintenance in a Nutshell: Monitor the ML inference results of the deployed AI model, as well as the input data received by the model, in order to detect possible concept changes or drifts. Retrain the model when needed.

2.3.12 Business Understanding

Building an AI model is often expensive and always time-consuming. It poses several business risks, including failing to have a meaningful impact on the user organization as well as missing in-production deadlines after completion. Business understanding is the stage at which companies that deploy AI models gain insight on the impact of AI on their business and try to maximize the probability of success.

Business Understanding in a Nutshell: Assess the value proposition of the deployed AI model. Estimate (before deployment) and verify (after deployment) its business impact.

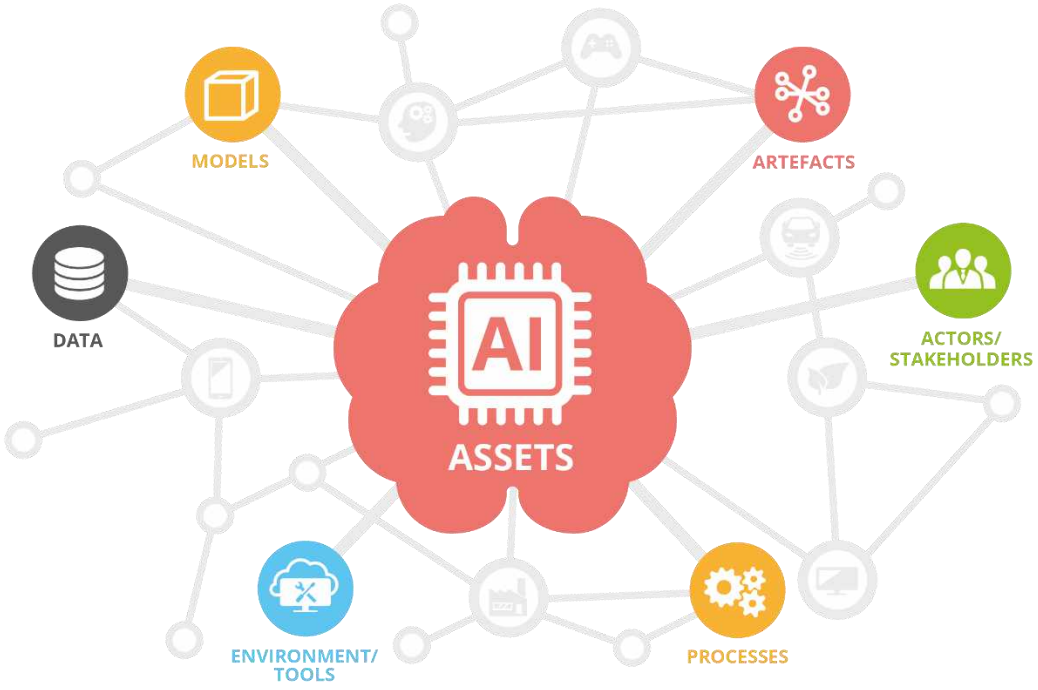
3. AI ASSETS

3.1 METHODOLOGICAL CONVENTIONS

A critical element in threat landscaping is identifying the categories of assets to which threats can be posed. Assets are defined as anything that has value to an individual or organization, and therefore requires protection. In the case of AI, assets are also those that are crucial to meet the needs for which they are being used.

Besides generic assets related to ICT, like data, software, hardware, communication networks, among others, AI implies a set of specific assets, like models, processors, and artefacts that can be compromised and/or damaged either due to intentional as to non-intentional causes.

Figure 3: AI assets' categories



3.2 ASSET TAXONOMY

For each of the stages in the AI lifecycle, the most relevant assets were identified, based on the functional description of specific stages and in order to reflect AI components, but also assets that support the developments and deployment of AI systems. Assets also include processes related to AI given their crosscutting nature. Assets were classified in the following 6 categories (see Figure 3):

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

Figure 4 illustrates the detailed asset taxonomy for AI based on the generic AI lifecycle reference model described in the previous chapter. Moreover, Annex A describes in detail the different assets and Annex C lists the AI lifecycle stage in which they belong.

Figure 4: AI asset taxonomy

Concluding this chapter, it is worth mentioning that due to the complexity of AI and the large scope of the AI ecosystem, as well as the evolving nature of AI systems and techniques, asset mapping is an ongoing task that will need some time to reach a mature stage. This is due to a variety of reasons/issues regarding the nature of AI systems (plethora of different techniques and approaches, different application deployment scenarios, associated fields such as facial recognition and robotics, etc.). An additional challenge involves the complexity and scale of the AI/ML supply chain and all the implications that it implies for the asset and threat landscape[50]. These challenges will be sufficiently managed in future assessment of AI threats.

[50] See https://stiftung-nv.de/ml-supplychain

4. AI THREATS

AI enables automated decision-making and facilitates many facets of daily life, bringing with it enhancements of operations and numerous other benefits. Nevertheless, AI systems are faced with numerous cybersecurity threats and AI itself needs to be secured since there have already been reported cases of malicious attacks, e.g. AI techniques and AI-based systems may lead to unexpected outcomes and may be tampered with to manipulate the expected outcomes[51][52][53]. It is thus essential to have an understanding of the AI Threat Landscape and to have a common and unifying foundation for understanding the potential of threats and accordingly conduct targeted risk assessments. The latter will support the implementation of targeted and proportionate security measures and controls to counter the threats related to AI.

In this chapter, we describe the threat landscape for AI, first discussing briefly related actors, then the adopted methodology to derive the threats to different assets, followed by a description of threats and their categorization in a generic taxonomy.

4.1 THREAT ACTORS

There are various groups of threat actors that may wish to harm AI systems using cyber means[54].

Cybercriminals are primarily motivated by profit. Cybercriminals will tend to use AI as a tool to conduct attacks but also to exploit vulnerabilities in existing AI systems[55]. For example, they might try to hack AI-enabled chatbots to steal credit card or other data. Alternatively, they may launch a ransomware attack against AI-based systems used for supply chain management and warehousing.

Company **insiders**, including employees and contractors that have access to an organization's networks, can involve either those that have malicious intent or those that can harm a company unintentionally. ***Malicious insiders*** might for example seek to steal or sabotage the dataset used by the company's AI systems. ***Non-malicious insiders*** might instead accidentally corrupt such a dataset.

Nation state actors and other state-sponsored attackers are generally speaking advanced. In addition to developing ways to leverage AI systems to attack other countries (including industries and critical infrastructures) as well as using AI systems to defend their own networks, nation state actors are actively searching for vulnerabilities in AI systems that they can exploit. This might be as a means of causing harm to another country or as a means of intelligence-gathering.

Other threat actors include **terrorists**, who seek to cause physical damage or even loss of life. For example, terrorists may wish to hack driverless cars in order to use them as a weapon.

Hacktivists, who mostly tend to be ideologically motivated, may also seek to hack AI systems in order to show that it can be done. There are a growing number of groups concerned about the potential dangers of AI, and it is not inconceivable that they could hack an AI system to

[51] See https://www.idgconnect.com/news/1506124/deepfakes-ai-deceives , September 2020
[52] See https://thenewstack.io/camouflaged-graffiti-road-signs-can-fool-machine-learning-models/, September 2017
[53] See https://www.media.mit.edu/publications/adversarial-attacks-on-medical-machine-learning/, March 2019
[54] Given the broad nature of AI systems and their deployment in diverse sectors, the listing is generic and does not imply a ranking of the likelihood of a threat actor to attack AI systems.
[55] With AI-as-a-service gaining traction, such systems will be increasingly available to non-technical savvy actors .

garner publicity.

There are also non-sophisticated threat actors such as **script kiddies** that may be criminally or ideologically motivated. These are generally unskilled individuals that use pre-written scripts or programs to attack systems, as they lack the expertise to write their own.

Beyond the traditional threat actors discussed above, it increasingly becomes necessary to include **competitors** as threat actors as well, as some companies are increasingly demonstrating intent to attack their rivals in order to gain market share.[56]

4.2 THREAT MODELLING METHODOLOGY

Thread modelling involves the process of identifying threats and eventually listing and prioritizing them[57]. There exist various methodologies on how to conduct threat modelling, with STRIDE[58] being one of the most prominent ones. In the context of future risk/treat assessments for AI for specific use cases, the threat modelling methodology that we followed involves 5 steps, namely:

1. **Objectives identification**: identify the security properties the system should have.
2. **Survey:** map the system, its components and their interactions and the interdependencies with external systems (as described in Chapter 2 on AI Lifecycle).
3. **Asset identification:** pinpoint the critical assets in terms of security that are in need of protection (as described in Chapter 3 on Assets).
4. **Threat identification:** identify threats to assets that will lead to the assets failing to meet the aforementioned objectives (this is the focus of Chapter 4).
5. **Vulnerability identification:** determine – usually based on existing attacks – whether the system is vulnerable with respect to identified threats[59].

In order to develop the ENISA AI Threat Landscape, we consider both traditional security properties, as well as security properties that are more pertinent to the field of AI. The former include ***confidentiality, integrity, and availability*** with additional security properties including ***authenticity, authorization*** *and* ***non-repudiation***, whereas the latter are more specific to AI and include ***robustness, trustworthiness, safety, transparency, explainability, accountability,*** as well as ***data protection[60].***

The impact of threats to confidentiality, integrity and availability is presented and accordingly, based on the impact on these fundamental security properties, the impact of threats on the additional security properties is mapped as follows:

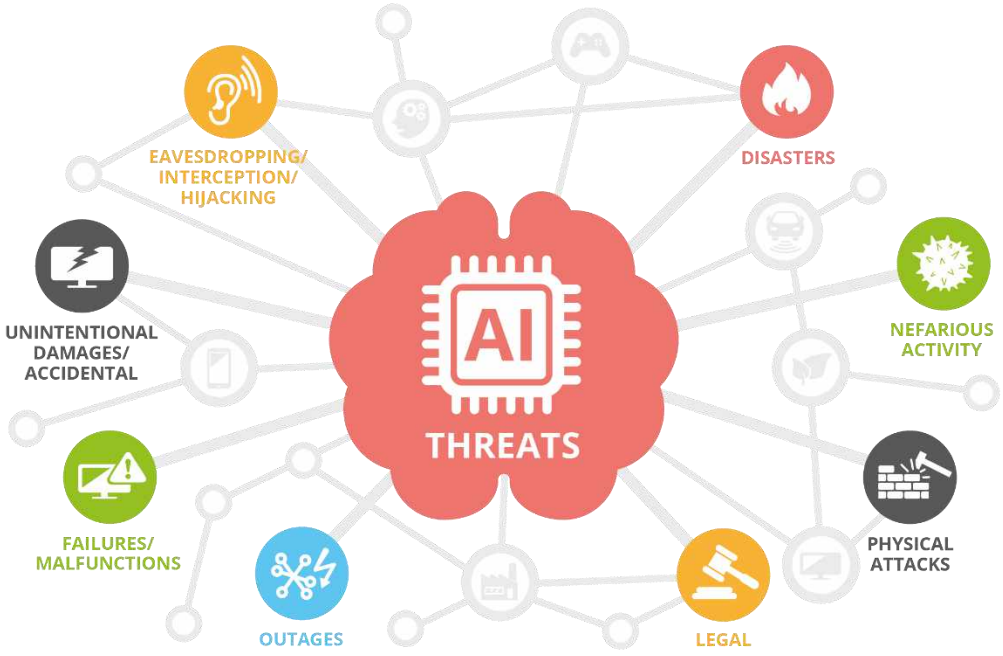
- Authenticity may be affected when integrity is compromised, since the genuineness of the data or results might be affected.
- Authorization may be impacted when confidentiality and integrity are affected, given that the legitimacy of the operation might be impaired.
- Non-repudiation may be impacted when integrity is affected.
- Robustness of an AI system/application may be impacted when availability and integrity are affected.

[56] See, Sailio, M.; Latvala, O.-M.; Szanto, A. Cyber Threat Actors for the Factory of the Future. Appl. Sci. 2020, 10, 4334 https://www.mdpi.com/2076-3417/10/12/4334/htm
[57] See Shostack, Adam (2014). "Threat Modeling: Designing for Security". John Wiley & Sons Inc: Indianapolis.
[58] See https://docs.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20). STRIDE is an acronym that stands for 6 threat categories, namely Spoofing, Tampering with data, Repudiation, Information disclosure, Denial of Service and Elevation of privilege.
[59] Vulnerability identification has not been extensively explored in this report given that specific use cases need to be considered in order to perform this step. Since the report aims to present the AI threat landscape in a domain agnostic manner, vulnerability identification will be pursued in more detail in further works.
[60] The AI specific security properties have been based on the work of the EC AI HLEG on assessment list for trustworthy AI: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

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

It must be noted that the report strictly focuses on AI security and does not further address in any way data protection requirements and/or aspects of GDPR compliance in AI systems and applications. Threats to data protection have been exclusively considered in the context of AI security.

Figure 5: AI Threat Taxonomy



Having introduced the security properties and based on the introduced AI lifecycle reference model and identified assets (see Chapter 3), the next step in the considered methodology entails identification of threats and vulnerabilities. To identify threats we consider each asset individually and as a group and highlight relevant failure modes[61] with respect to the above mentioned security properties. By identifying threats to assets, we are able to map the threat landscape of AI systems. Moreover, the effects of identifying the threat to vulnerabilities of AI systems are also underlined by referring to specific manifestations of attacks. This would lead in the future to the introduction of proportionate security measures and controls.

4.3 THREAT TAXONOMY

The list below presents a list of high-level categorization of threats based on ENISA threat taxonomy[62], which has been used to map the AI Threat Landscape.

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

Figure 5 depicts the AI Threat taxonomy's main categories based on the aforementioned categorization, whereas Annex B describes the 74 identified threats to AI as depicted in the threat taxonomy and lists the affected assets per threat, as well as the potential impact with respect to the aforementioned properties of AI systems. Annex D maps the threats per lifecycle stage (as described in Chapter 2). Some threats repeat in more than one category, because they can occur as both unintended damage and nefarious activity for example. It needs to be noted that the identified threats are specific to the context of AI systems and threats to other elements of the ecosystem have not been fully explored. For example, threats to cloud infrastructure (upon which the majority of AI systems rely) have been briefly highlighted here. The same goes for threats to communication network infrastructures, or the sensors collecting data that feed into AI systems. The interested reader is referred to relevant ENISA threat landscapes that should be used in tandem for comprehensive risk assessments.

Figure 6 details the specific threats under each of the categories.

Annex B describes the 74 identified threats to AI as depicted in the threat taxonomy and lists the affected assets per threat, as well as the potential impact with respect to the aforementioned properties of AI systems. Annex D maps the threats per lifecycle stage (as described in Chapter

[61] See https://link.springer.com/article/10.1186/s40887-018-0025-1
[62] See https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends/enisa-threat-landscape/threat-taxonomy/view

2). Some threats repeat in more than one category, because they can occur as both unintended damage and nefarious activity for example. It needs to be noted that the identified threats are specific to the context of AI systems and threats to other elements of the ecosystem have not been fully explored. For example, threats to cloud infrastructure (upon which the majority of AI systems rely) have been briefly highlighted here. The same goes for threats to communication network infrastructures, or the sensors collecting data that feed into AI systems. The interested reader is referred to relevant ENISA threat landscapes[63] that should be used in tandem for comprehensive risk assessments.

Figure 6: Detailed AI threat taxonomy

[63] See for example https://www.enisa.europa.eu/publications/cloud-computing-risk-assessment, https://www.enisa.europa.eu/publications/enisa-threat-landscape-for-5g-networks and https://www.enisa.europa.eu/publications/baseline-security-recommendations-for-iot

NEFARIOUS ACTIVITY/ABUSE

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models models' code
- Compromise and limit AI results
- Compromising AI inference s correctness data
- Compromising ML inference s correctness algorithms
- Data poisoning
- Data tampering
- Elevation of Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White box , targeted or non targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor insert attacks on training datasets
- Overloading confusing labelled dataset
- Compromising ML training validation data
- Compromising ML training augmented data
- Adversarial examples
- Reducing data accuracy
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List ACL) manipulation
- Compromising ML pre processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI ML results
- Label manipulation or weak labelling
- Model backdoors

PHYSICAL ATTACK

- Errors or timely restrictions due to non reliable data infrastructures
- Model Sabotage
- Infrastructure system physical attacks
- Communication networks tampering
- Sabotage

DISASTER

- Natural disasters (earthquake , flood, fire , etc)
- Environmental phenomena heating , cooling , climate change

FAILURES/MALFUNCTIONS

- Compromising AI application viability
- Errors or timely restrictions due to non reliable data infrastructures
- 3 rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks

EAVESDROPPING/INTERCEPTION/ HIJACKING

- Data inference
- Data theft
- Model Disclosure
- Stream interruption
- Weak encryption

LEGAL

- Corruption of data indexes
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3 rd parties
- Vendor lock in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal informationCorruption of data indexes

OUTAGES

- Infrastructure/system outages
- Communication networks outages

UNINTENTIONAL DAMAGES/ ACCIDENTAL

- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference s correctness data
- Compromising feature selection
- Compromising ML inference s correctness algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training augmented data
- Reducing data accuracy
- Compromise of data brokers providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks

5. CONCLUSIONS

The significance and impact of AI in society nowadays cannot be overstated[64]. It permeates every aspect of our daily lives and therefore it is of paramount importance to ensure the cybersecurity of AI to ensure that AI and the set of associated technologies will be trustworthy, reliable and robust.

Setting a baseline for a common understanding on relevant AI cybersecurity threats such as this threat landscape will be key to widespread deployment and acceptance of AI systems and applications. The AI threat landscape highlights a number of cybersecurity challenges around complexity, technical issues, integrity, confidentiality and privacy. Moving towards a secure digital transformation means different steps for different actors. Accordingly, an AI toolbox should be developed with concrete mitigation measures for the AI threats identified in the landscape based on risk assessments.

Whereas the threats to AI systems may be seen from a horizontal point of view, namely the threats as presented in the AI Threat Landscape report apply to all AI systems/applications (based on their configuration and utilized techniques), conducting targeted risk assessments of AI systems needs to consider the context of use. Different sectors exhibit different degrees of risk that needs to be assessed and accordingly different security measures and controls to be put in place. That being said, the introduction of horizontal methods/methodologies and best practices could be of value in establishing a common baseline and thus promoting a common layer of cybersecurity and trust across sectors. Such a horizontal approach should be based on the AI threat landscape that is also sector-agnostic and introduce security controls and practices as per the aforementioned AI toolbox, in order to promote EU-wide terminology, understanding and mitigation of relevant threats.

Furthermore, the AI threat landscape underlined the need to develop control measures for a variety of threats towards AI and highlighted the fact that there is still research work needed to better foster robust systems and solutions. This includes foresight work by law enforcement, who should proactively assess and predict AI misuse to improve preparedness, such as the Europol, TrendMicro and UNICRI's report on "Malicious Uses and Abuses of Artificial Intelligence"[65].

In addition, a mapping of gaps in the future research directions in the context of AI and cybersecurity should be undertaken to have greater foresight regarding the future of emerging technology and of the interplay between cybersecurity and AI. It is evident from the threat landscape that work needs to be done in the area of automatic formal verification and validation, explainability and transparency, novel security techniques to counter emerging AI threats, to name a few. Research activities are needed in various areas towards developing AI trustworthy algorithms, systems and solutions improving the industrial and security operations and EU markets' competitiveness by developing the 'AI made in Europe' brand as a seal of quality for ethical, secure and cutting-edge AI that can become a worldwide reference.

The advantages that AI technologies bring to the digitalisation of our societies are numerous. AI can support cybersecurity and cybercrime operations with techniques that may be used to

[64] Moreover, the impact of AI cannot be totally anticipated, which put us in the position of, maybe, rewriting all the rules that we have used so far.
[65] See . https://www.europol.europa.eu/newsroom/news/new-report-finds-criminals-leverage-ai-for-malicious-use-%E2%80%93-and-it%E2%80%99s-not-just-deep-fakes

augment/automate cybersecurity operations, such as intelligent firewalls. Moreover, it is essential to secure the diverse assets of the AI ecosystem and lifecycle, assets that reside in complex supply chains and involve cross-border and cross-industry relationships. The security and integrity of the AI supply chain is thus of paramount importance.

In this respect, the significance of leveraging public-private partnerships and fostering the establishment of multi-disciplinary groups of AI cybersecurity experts, such as the ENISA Ad Hoc Working Group on AI Cybersecurity are both essential. Moreover, work such as the one conducted in the context of ETSI ISG SAI is highly important, since security of AI has to date not been widely studied in the context of standardisation.

The complexity and vastness of the AI cybersecurity threat landscape necessitate fostering an EU ecosystem for secure and trustworthy AI, including all elements of the AI supply chain. This is a race that is of particular importance to EU given the long-term strategic objectives concerning AI. The EU secure AI ecosystem should place cybersecurity and data protection at the forefront and foster relevant innovation, capacity-building, awareness raising and research and development initiatives.

ANNEX A - ASSET TAXONOMY DESCRIPTION

Category	Asset	Definition	AI Lifecycle stage
Data	**Augmented Data Set**	An augmented data set is a (usually labeled) data set which has been augmented by adding data produced by transformations or by generative ML models. Augmentation significantly increases labeled data sets' diversity (which is supposed to prevent overfitting) in view of using augmented datasets for training other ML models. In image recognition, data augmentation techniques include cropping, padding, and horizontal flipping.	- Data pre-processing
	Evaluation Data	The evaluation data is used to evaluate the predictive quality of the trained model. The ML system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics.	- Model Tuning
	Labeled Data Set	The term "Labeled Data" refers to a set of scalar or multi-dimensional data items that have been tagged with one or more informative labels, usually for the purpose of training a supervised ML model.	- Data pre-processing
	Metric Data Set	The sorts of numbers we collect when we measure something. Metric data can be ratio scale, interval scale, integer scale and cardinal numbers.	- Feature Selection
	Pre-processed Data Set	The data is pre-processed before feeding it into our ML model.	- Data pre-processing
	Public Data set	Public data set is information that can be freely used, reused and redistributed by anyone with no existing local, national or international legal restrictions on access or usage.	- Data Exploration - Data Ingestion
	Raw Data	Raw data refers to any type of information gathered for AI analysis purposes, possibly after cleaning but before it is transformed or analyzed in any way.	- Data Ingestion
	Testing Data	It is a dataset used to provide an unbiased evaluation of a final ML model fitted on the training dataset. We use testing data to test the model. If the data in the test dataset has never been used in training (e.g. in cross-validation), the test dataset is also called a holdout dataset.	- Model Training
	Training Data	Training data refers to the initial data that is used to develop a ML model, from which the model adapts its internal parameters to refine its rules.	- Model Selection / Building - Model Training - Transfer Learning

Category	Asset	Definition	AI Lifecycle stage
	Validation Data Set	Validation data sets are labelled data sets, which differ from ordinary labelled data sets only in their usage and, usually, in their collection circumstances. Validation data sets are mostly used to perform an evaluation of a ML model in-training, for example by stopping the ML model's training (early stopping) when the error on the validation dataset increases too much, as this is considered a sign of overfitting the model to the training dataset.	- Model Tuning
Models	**Algorithms**	ML algorithms are programs (math and logic) that adjust themselves to perform better as they are exposed to more data. The "learning" part of ML means that those programs change how they process data over time, much as humans change how they process data by learning. So a ML algorithm is a program with a specific way to adjusting its own parameters, given feedback on its previous performance in making predictions about a dataset.	- Model Training
	Data Pre-Processing Algorithms	The data pre-processing employs techniques to clean, integrate and transform the data, resulting in an improved data quality that will improve performance and efficiency by saving time during the analytic models' training phase and by promoting a better quality of results. Specifically, the term data cleaning designates techniques to correct inconsistencies, remove noise and anonymize/pseudonymize data.	- Data pre-processing
	Hyper-parameters	Hyper-parameters define high-level concepts about ML models, such as the frequency of the adjustment of the internal parameters on the part of the training algorithm. They cannot be learnt from input data but need to be set by trial-and-error using model space search techniques.	- Model Tuning
	Training Algorithms	Training algorithms are procedures for adjusting the parameters of ML models. In supervised training, the correct output for each input vector of a training set is presented to the model, and multiple iterations through the training data may be required to adjust the parameters. In unsupervised training, the model parameters are adjusted without specifying the correct output for any of the input vectors.	- Model Selection / Building
	Model	The term ML model designates computer algorithms implementing parametric mathematical models that improve through experience.	- Model Training - Model Tuning - Model Selection / Building - Model Deployment - Model Maintenance
	Model parameters	A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.	- Model Training
	Model performance	The ML model performance is the accuracy and speed of the model's computation that receives inputs from the production-ready environment and outputs the model' classifications, predictions or decisions.	- Model Training - Model Tuning

Category	Asset	Definition	AI Lifecycle stage
	Subspace (Feature) selection Algorithm	Feature selection algorithms are techniques that select a subset of relevant features from an original feature set, in order to increase the performance of ML models trained on the subset. Some feature selection methods used for classification problems are supervised and use class labels as a guide.	- Feature Selection
	Trained models	A trained ML model is a model whose internal parameters have been adjusted by training to reach a minimum of the error function that defines the distance between the actual and expected outputs.	- Model Training - Transfer Learning
	Training parameters	ML model training parameters are quantities adjusted by the learning process by applying training algorithms based on training data. Training parameters values determine the actual classification, prediction or detection function computed by the ML model.	- Model Selection / Building - Model training
	Tuned Model	A tuned ML model is a model where the hyper-parameters affecting the training algorithm operation have been set to maximize the convergence and speed of the training algorithm.	- Model Tuning
Actors	**Cloud Provider**	Cloud providers are third parties that offer computational platforms, and even in some cases tend to offer some data analyses capabilities or “Machine learning as service” (for this please check Model Provider threats below). Cloud provider adds to AI the same attack vectors as to other domains: data breaches, insufficient authentication and authorization, insecure interfaces, etc.	- Data Ingestion - Model training - Model tuning
	Data Engineers	Data Engineers are professionals that prepare the computational infrastructure and mainly focus on the design, management and optimization of the flow of data. They’re usually more intervenient in the first stages of AI Lifecycle: extraction and assembling of data from different sources, transformation, cleaning and loading it in a standardized format and in an adequate repository. Data Engineers must have specialized skills in creating software solutions around data: software engineering, distributed systems, open frameworks, SQL, Cloud platforms, data modelling.	- Data Ingestion - Data Exploration - Data pre-processing - Feature Selection - Model Selection / Building - Model Training - Model Tuning - Model Deployment - Model Maintenance
	Data Owner	Data owner can be a data broker or provider, as described before, or the business owner, who asks for the AI study. In this section, the focus is on the latter.	- Business Goal Definition - Data Ingestion - Data Exploration
	Data Provider/Data Broker	Third parties’ providing data for the AI process.	- Data Ingestion

Category	Asset	Definition	AI Lifecycle stage
	Data Scientists / AI designer/AI developer	Professionals that apply statistics, Machine Learning and analytic approaches to analyse different datasets of different sizes and shapes and solve complex and critical problems. Skills in computer science fundamentals and programming, including experience with languages and database (big/small) technologies are essential.	- Business Goal Definition - Data Ingestion - Data Exploration - Data pre-processing - Feature Selection - Model Selection / Building - Model Training - Model Tuning - Transfer Learning - Model Deployment - Model Maintenance
	End Users	Those inside an organization that use and benefit from the results provided by the AI/ML system/service.	- Business Goal Definition - Data Ingestion - Data Exploration - Model Maintenance - Business Understanding
	Model provider	In the context of transfer and/or federated learning, third parties that provide models (called as “Teacher” models), previously trained and fine-tuned with large datasets that are useful to learn from small datasets and/or by organizations without access to high computational clusters, with GPU.	- Transfer Learning
	Service consumers / Model users	AI/ML users that rely on pre-trained models, or consume them through available services.	- Model Maintenance - Business Understanding
Processes	**Data augmentation**	Techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. It helps reduce overfitting when training a machine learning. Data augmentation usually consists in applying transformations which are known to be label-preserving, i.e. the model should not change its prediction when presented with the transformed data items.	- Data pre-processing - Data Exploration
	Data Collection	It is the process of gathering and measuring information on specific variables (needed for the AI system) from countless different sources.	- Data Ingestion

Category	Asset	Definition	AI Lifecycle stage
	Data Exploration/Pre-processing	Understanding, preparing and cleaning data.	- Data Exploration - Data pre-processing
	Data Ingestion	Data Ingestion is the process related to data transportation from multiple sources to compose multi-dimensional data points. Data can be placed in a storage medium where it can be accessed, used, and analyzed, or, the data stream can be used directly in the ML process.	- Data Ingestion
	Data labelling	It is the process of detecting and tagging data samples. The process can be manual and time-consuming and assisted by software.	- Data pre-processing
	Data Storage	Data can be stored locally, in a distributed file system, in the cloud.	- Data Ingestion
	Data understanding	Data understanding is the knowledge you have about data, data assets, the needs the data will satisfy, its content and location.	- Data Exploration
	Feature selection	During this process the number of dimensions or features of the input vector is reduced, by identifying those that are the most meaningful for the AI/ML model.	- Feature Selection
	Model adaptation – transfer learning / Model deployment	Transfer Learning is the ability to re-use previously learned knowledge to solve new problems faster. Deployment is the process of taking a pre-trained model and making it available to the users. Transfer learning emphasizes the transfer of knowledge across domains, tasks, and distributions that are similar but not the same[66]. By using transfer learning, where a small number of highly tuned and complex centralized models (called Teachers) are shared with the general community, and are customized by users or organizations for a given application using limited training over small specific domain datasets[67].	- Transfer Learning
	Model Maintenance	After deployment, it is necessary to monitor the prediction accuracy to detect possible changes or drift of concepts. A decrease in model performance might be overcome by retraining it using recent data and then redeploy it in production.	- Model Maintenance
	Model selection/building, training and testing	During the training process the selected, or developed, algorithm is trained with input data, this means algorithm parameters, like weights and bias, will be learned from the data. During this stage, the resulting prediction is compared with the actual value for each data instance, the accuracy is evaluated and model parameters adjusted until the best values are found.	- Model Selection / Building

[66] See Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22(10), 1345–1359 (2010)
[67] See Weiss, K., Khoshgoftaar, T.M. & Wang, D. A survey of transfer learning. *J Big Data* **3,** 9 (2016). https://doi.org/10.1186/s40537-016-0043-6

Category	Asset	Definition	AI Lifecycle stage
	Model tuning	Tuning focus on setting up special parameters, often called hyper-parameters. This process can be done manually or automatically by searching the model parameters' space, through a so called hyper-parameter optimization. While during the training process model parameters are tuned, during the tuning process hyper-parameters are adjusted by running the whole training job and looking at the aggregated accuracy.	- Model Tuning
	Reduction/Discretization technique	It is the process of converting a numerical attribute into a symbolic attribute by partitioning the attribute domain.	- Feature Selection
Environment /Tools (hardware/ software)	**Cloud**	It is the on-demand availability of computer system resources, especially data storage (cloud storage) and computing power, without direct active management by the user. The term is generally used to describe data centres available to many users over the Internet.	- Data Ingestion - Model training - Model tuning
	Communication networks	Networks with internet connectivity for communication purposes.	- Data Ingestion
	Communication protocols	Communication protocol is a system of rules that allows two or more entities of a communications system to transmit information via any kind of variation of a physical quantity. The protocol defines the rules, syntax, semantics and synchronization of communication and possible error recovery methods. Protocols may be implemented by hardware, software, or a combination of both.	- Data Ingestion
	Computational platforms	It is the environment in which a piece of software is executed. It may be the hardware or the operating system (OS), even a web browser and associated application programming interfaces, or other underlying software, as long as the program code is executed with it.	- Data pre-processing - Feature Selection - Model Selection / Building - Model Training - Model Tuning - Transfer Learning - Model Deployment - Model Maintenance
	Data exploration tools	The tools used for data exploration. Tools such as visualization tools and charting features are frequently used to create a more straightforward view of data sets than simply examining thousands of individual numbers or names	- Data Exploration
	Data ingestion platforms	It is the platform where data ingestion takes place.	- Data Ingestion
	Database management system	It is the software that handles the storage, retrieval, and updating of data in a computer system.	- Data Ingestion

Category	Asset	Definition	AI Lifecycle stage
	Distributed File System	File System Distribution is a method for storing and accessing files, which allows for multiple users to access to and to share files from multiple machines, or multiple hosts, via a computer network. Control of information access and authorization is critical. In the context of distributed file systems, distributed databases and data stored in the cloud are encompassed.	- Data Ingestion
	Integrated Development Environment	It is a software application that provides comprehensive facilities to computer programmers for software development. An IDE normally consists of at least a source code editor, build automation tools and a debugger.	- Data pre-processing - Feature Selection - Model Selection / Building - Model Training - Model Tuning
	Libraries (with algorithms for transformation, labelling, etc)	Pre-written programs implementing algorithms ready to be used, for: scientific computation, tabular data, Time-Series Analysis, Data Modelling and Preprocessing, deep learning, among others. Their usage saves time and facilitates the development of high-level analytical functions, even for less trained ML developers. Some examples are Apache Spark MLib, Scikit-learn Python, Keras Python, etc.	- Data Exploration - Data pre-processing - Feature Selection - Model Selection / Building - Model Training - Model Tuning
	Machine Learning Platforms	Provide an ecosystem of tools, libraries and resources that support the development of machine learning applications.	- Data Exploration - Data pre-processing - Feature Selection - Model Selection / Building - Model Training - Model Tuning - Model Deployment - Model Maintenance

Category	Asset	Definition	AI Lifecycle stage
Artefacts	**Monitoring Tools**	Tools that are used to continuously keep track of the status of the system in use, in order to have the earliest warning of failures, defects or problems and to improve them.	- Data pre-processing - Feature Selection - Model Selection / Building - Model Training - Model Tuning - Transfer Learning - Model Deployment - Model Maintenance
	Operating System/software	It manages computer hardware, software resources, and provides common services for computer programs.	- Model Deployment - Model Maintenance
	Optimization techniques	Techniques used for optimization in model tuning such as Grid Search, Random Search and Bayesian optimization.	- Model Tuning
	Processors	A processor is the part of a computer that interprets commands and performs the processes the user has requested.	- All stages
	Visualization tools	Any program, utility, routine or function that performs an operation by dragging and dropping icons or by "drawing" the solution. Visual tools are the norm in virtually every graphics-based application.	- Data Exploration
	Access Control Lists	An access control list (ACL) is a table that represents which access rights each user has to a particular resource, such as a file directory or individual file. In an organization's Active Directory, the ACL of a resource specifies the organization's access intent for that resource. An ACL has an entry for each user account (or user group) with access privileges, and each resource has a security attribute that identifies its access control list. The most common privileges include the ability to read a file (or all the files in a directory), to write to the file or files, and to execute the file (if it is an executable file, or program). Collecting data for AI applications requires checking read/write permissions on ACLs regarding people as well as 'things', and taking into account increasingly stringent safety and data privacy regulations. Managing ACL permissions and access rights via groups (and groups of groups) is a standard technique for managing access to IT resources. User accounts will inherit all access permissions to resources that are set on the group of which they are (direct or indirect) members.	- Data Ingestion

Category	Asset	Definition	AI Lifecycle stage
	Composition artefacts: AI models compositions	Compositions of AI models (also called ensemble systems) put together multiple AI models, typically via majority voting, in order to reduce the outputs' variance and improve the accuracy of the overall composition with respect to the ones of individual components. Ensemble systems have been successfully used to address a variety of problems, such as feature selection, confidence estimation, missing data and concept drift from non-stationary distributions, among others.	- Data pre-processing - Feature Selection - Model Selection / Building - Model Training - Model Tuning - Model Deployment - Model Maintenance
	Data and Metadata schemata	A data schema is a skeleton structure, often depicted by means of schema diagrams, that defines how data is organized and the relations among them. It also formulates all the constraints that are to be applied on the data. In turn, metadata schemata define the overall structure for the metadata. They describes how the metadata is set up, and usually rely on standards for common components like dates, names, and places. Discipline-specific metadata schemata are used to collect the specific metadata needed by a discipline.	- Data Ingestion - Data Exploration - Data pre-processing
	Data displays and plots	A data display (or data plot) is a graphical technique for representing a data set as a graph, highlighting the relationship between two or more variables. Data plots provide a visual representation of the relationship between variables, helping human experts to quickly gain an understanding which may not come from lists of values. Data plot techniques include, among others, scatter and spectrum plots, histograms, pie charts, probability and residual plots, box and block plots.	- Data Exploration
	Data Governance Policies	Data governance policies are sets of guidelines ensuring that data and information assets are managed consistently and used properly. They articulate the principles, practices and standards that organizations consider necessary to ensure they hold high-quality data and that data assets are adequately protected. A data governance policy is typically a composite artefact, including individual policies for data quality, access, security and privacy. It also specifies the organizational roles and responsibilities for implementing those policies and the methodology to be used for monitoring compliance with them.	- Data Ingestion
	Data Indexes	Data indexes are special data structures that store a small portion of a data set in a form which is easy to traverse or to search into. Indexes store the value of a specific field or set of fields, ordered by the value of the field. The ordering of the index entries supports efficient equality matches and range-based query operations.	- Data Ingestion - Data Exploration - Data pre-processing

Category	Asset	Definition	AI Lifecycle stage
	Descriptive Statistical Parameters	Descriptive statistical parameters are the quantities that characterize the probability distribution of a statistic or a random variable. They can be regarded as a numerical characteristic of statistical populations. Parametric probability distributions include the normal or Gaussian distribution, the Poisson distribution, the binomial and the exponential family of distributions. For instance, the family of normal distributions has two parameters, the mean and the variance: if those are specified, the distribution is known exactly. Statistical parameters are sometimes unobservable; in this case it is the data scientists' task to infer what they can about the parameter, based on a random sample taken from the population of interest.	- Data Exploration
	High-Level Test cases	High Level Test Cases (HLTCs) are inputs used to test AI models. HLTCs are the union of four different datasets: the classic training, validation and test datasets (the latter being often a subset of training dataset), and a held-out dataset. HLTCs also include some specific inputs of interest. AI models' testing is the procedure for (i) assessing the AI model's performance on each dataset composing the HLTCs and comparing it to a pre-determined minimum acceptable threshold (ii) computing the model's outputs corresponding to some specific inputs of interest. The rationale for the latter is that when a ML model shows good aggregate performance, it can be hard to notice whether its performance is acceptable on specific types of inputs.	- Business Goal Definition - Model deployment
	Informal/ Semi-formal AI Requirements, GQM (Goal/Question/Metrics) model	Semi-formal Requirements are often used to specify functional and non-functional requirements for AI systems. Functional requirements model the domain of interest, the AI problem to be solved, and the task to be executed by the AI system. Non-functional requirements include architectural (hardware) and code (software) components. For example, a CPU-based environment might not be sufficient for large ML training loads and GPUs (cloud-based or on-premises) could be required. Requirements on network bandwidth and storage are also relevant. Since AI can involve handling sensitive data such as patient records, financial information, and personal data, security requirements (usually, regarding the Confidentiality-Integrity-Authenticity (CIA) triad) are also important for AI systems. Goal Question Metrics (GQM) models complement the non-functional requirements with metrics such as computing performance/storage capacity (for the architectural component) and source code and complexity level (for the code component).	- Business Goal Definition
	Model Architecture	It defines the various layers involved in the AI/ML lifecycle and involves the major steps being carried out in the transformation of raw data into training data sets capable for enabling the decision making of a system.	- Model Selection / Building - Model deployment
	Model hardware design	It may be viewed as a 'partitioning scheme,' or algorithm, which considers all of the system's present and foreseeable requirements and arranges the necessary hardware components into a workable set of cleanly bounded subsystems with no more parts than are required.	- Model Selection / Building - Model deployment

Category	Asset	Definition	AI Lifecycle stage
	Model frameworks, software, firmware or hardware incarnations.	Model frameworks include all software, firmware and hardware components required to train and deploy an AI model. Within model frameworks, developers use ML libraries (e.g. Keras or TensorFlow) to describe their ML model's structure and implement the corresponding training algorithms. These libraries rely on math libraries like NumPy to handle complex matrix operations used for the gradient descent and loss function calculations. In turn, math libraries build on lower level libraries such as Basic Linear Algebra Subroutines (BLAS). To speed-up model training and inference, software frameworks typically rely on one or more graphics processing units (GPUs) with corresponding GPU-enabled libraries. Deployment in firmware moves the AI model to the (read-only) memory of a Microcontroller Unit (MCU), which can be embedded into industrial systems. Developers who need increased performance turn to Field Programmable Gate Arrays (FPGAs) that embed memory blocks to reduce the memory access bottleneck that limits performance in these kinds of compute intensive operations. Deployment in hardware deploys the ML model as a custom hardware chip. Specialized AI hardware will eventually provide significant performance enhancements for ML models.	- Transfer Learning - Model Deployment - Model Maintenance
	Use Case	A specific situation in which the ML model could potentially be used.	- Business Understanding
	Value proposition and business model	Value propositions are promises of value to be delivered from organizations to stakeholders via a service or a product, or expectations on the part of the latter of the value (benefit) they will receive from the product or service. Business models provide the rationale of how organizations will create and deliver the value.	- Business Understanding

ANNEX B – THREAT TAXONOMY DESCRIPTION

Threat Category	Threat	Description	Potential impact	Affected assets
Nefarious Activity/Abuse	**Access Control List (ACL) manipulation**	In AI data collection scenarios, group-based ACLs for datasets may fail when the nesting of large groups is changed. Given a data set and a group of sources Sensor_Group_A which has been granted access to update it, it is easy to check if an individual user or sensor is a member of *Sensor_Group_A* and inherits the corresponding permissions. However, if *Sensor_Group_A* is joined as a member to many other groups, inherited permissions become difficult to check for each sensor and escalation of privileges of untrusted sources may result. The threat involves Implicit privilege elevation attacks take advantage of group nesting modifications to upscale access permissions for specific users.	Integrity	Artefacts
	Adversarial examples	Targeting the inference phase of ML and deep learning systems that AI is based on is one of the most prominent and highly publicized threats. Adversarial examples refer to data that include perturbations that are imperceptible to the human eye, but that can have an impact on the effectiveness and performance of ML models.	Integrity Availability	Model Data
	Backdoor/insert attacks on training datasets	Threaten the ML model's integrity by trying to introduce spurious inferences. Attackers introduce special trigger patterns in part of the training data, and presenting the trigger in the inference phase will cause targeted misclassifications. For example, an attacker can introduce in the training data of an image classifier connected to a surveillance camera an example including a certain pixel pattern and the label "policeman". Once the classifier is trained and deployed, the attacker wears a t-shirt with that pattern and passes by the camera with a gun in hand without triggering any alarm.	Integrity	Data
	Compromising AI inference's correctness - data	This type of threats refers to possible exploitations involving either data manipulation, or selection bias in raw data, or modification of labels and deletion or omission of labelled data items. It may also refer to compromising AI correctness by insertion of adversarial data (poisoned/manipulated) in augmented data sets, as well as by means of interruption of training or modification of model parameters.	Integrity	Data
	Compromise and limit AI results	This type of threat can emerge due to involuntary or unintentionally actions from Data Owners, that may hide data due to business secrets or by not recognizing its value; by AI/ML	Integrity Availability	Model Actors

Threat Category	Threat	Description	Potential impact	Affected assets
		designers and engineers, that can intentionally tamper or, due to lack of experience, miss to include data. This threat may also be related to AI/ML service users not being able to understand the model capabilities and/or results.		Artefact
	Compromising ML inference's correctness – algorithms	Threats to the availability of the ML training algorithm, as well as threats that aim at compromising the training algorithm to adversely affect the desired accuracy.	Integrity Availability	Model
	Compromising ML pre-processing	Flaws or defects of the data and metadata schemata greatly influence the quality of the analysis by applications that use the data. In AI applications, a flawed schema will negatively impact on the quality of the ingested information. Flaws often result from of inconsistencies in the use of modeling methodologies, but may also depend on intentional schema poisoning, i.e. any manipulation of a schema intended to compromise the programs that ingest or pre-process data that use it. It is also possible for adversaries to mount Schema-based denial of service attacks, which modify the data schema so that it does not contain required information for subsequent processing.	Integrity Availability	Data Artefacts
	Compromising ML training – augmented data	Threats to augmented datasets due to inconsistency with the training set they are derived from, and specifically when highly diverse, automatically generated data are added to a data set of collected data, which are very consistent but highly representative of their application domain, so there would be no need to limit overfitting. Enriching data always entails some risks. This threat can lead to non-satisfaction of functional requirements, i.e. poor inference.	Integrity Availabilityonfidentiality	Data
	Compromising ML training – validation data	This threat refers to shortening the training of the ML model dramatically by compromising the integrity of the validation dataset. It also includes, generation of adversarial validation data that are quite different from genuine training set data	Integrity Availability	Data
	Compromise of data brokers/providers	This threat refers to compromising data brokers/providers to influence the machine learning process as they can deliberately or accidentally manipulate the data sent to the AI process, in several different ways: poisoning-via-insertion of malicious data; deleting registries to eliminate features either by changing the data, removing part of it or adding new registries. In addition, sometimes the mere data availability prevails over any consideration on data quality, with the risk that learning models are fed with data streams that do not reflect the statistical characteristics of a phenomenon and determining likely biases in the subsequent decisional processes.	Integrity Availability	Actor

Threat Category	Threat	Description	Potential impact	Affected assets
	Compromise of model frameworks	Model frameworks fail when are misconfigured or offer additional attack vectors with respect to traditional software, firmware and hardware environments. The ML platform's data volume and processing requirements mean that the workloads are often handled on the cloud, adding another level of complexity and vulnerability. Moreover, the threat of backdoors in libraries is also evident, similarly to the potential threats of attacks on model input and output data can be performed on the hardware, firmware, Operating System (OS) and software level.	Integrity Availability	Artefacts Environment/tools
	Corruption of data indexes	Data indexes threats manifest when their content becomes corrupted. Corruption may be the result of an attack, or due to system crashes or loss of network connectivity during index replication. The same events may cause interruptions of index construction tasks, bringing a partially built (and therefore defective) index to production. Also, running out of storage capacity during indexing or replication may cause an entire data index to be deleted. Denial-of-service attacks to indexes intentionally corrupt data indexes to decrease the performance of data access. Additionally, timing attacks to indexes use access time to public items before and after inserting them (which depend on the index content) to infer the presence and size of inaccessible data items.	Integrity Availability Confidentiality	Artefacts
	Data poisoning	This threat relates to the injection of erroneous/tampered/wrong data in the training set or the validation set by either getting legitimate access or illegitimate one through exploiting poor authentication/authorization mechanisms. The aim is to adversely affect operation of the AI system.	Integrity Availability	Process Environment/tools Model
	Data tampering	Actors like AI/ML designers and engineers can deliberately or unintentionally manipulate and expose data. Data can also be manipulated during the storage procedure and by means of some processes like feature selection. Besides interfering with model inference, this type of threat can also bring severe discriminatory issues by introducing bias.	Availability Integrity	Process Environment/tools Model
	DDoS	Distributed Denial of Service attacks may be utilized by adversaries to reduce the availability of online IT systems and distributed file systems (e.g. cloud storage) used to support AI systems and their operation.	Availability	Environment/tools
	Elevation-of-Privilege	These threats refers to exploiting trained and tuned models to gain access to parameters values and even to understand whether some data was part of the data set used.	Confidentiality Availability	Model Data
	Insider threat	AI designers and developers may deliberately expose data and models for a variety of reasons, e.g. revenge or extortion. Integrity, data	Confidentiality Integrity	Actors

Threat Category	Threat	Description	Potential impact	Affected assets
		confidentiality and trustworthiness are the main impacted security properties.	Availability	
	Introduction of selection bias	Data owners may introduce selection bias on purpose when publishing raw data in order to adversely affect inference to be drawn on the data.	Integrity Availability	Data
	Label manipulation or weak labelling	This threat refers to supervised learning systems, which not infer correctly due to wrong or imprecise data labels. If adversaries can only modify the training labels with some or all knowledge of the target model, they need to find the most vulnerable labels. Random perturbation of labels is one possible attack, while additionally there is the case of adversarial label noise (intentional switching of classification labels leading to deterministic noise, an error that the model cannot capture due to its generalization bias).	Availability Integrity	Processes Data
	Manipulation of data sets and data transfer process	These threats are seen in context of storage of data sets in infrastructures provided by third parties, which make them remotely accessible. The threat refers to manipulation and tampering of the data stored and manipulation of the data transfer process.	Confidentiality Integrity	Environment/tools
	Manipulation of labelled data	Threats to labelled data items occur when enough labels and data are deleted/omitted, when a sufficient number of spurious labelled data is included into the data set, or when enough labels are modified Since the labelled data set is used for the purpose of training a ML model in the supervised setting, all such modifications affect the model training and inference (e.g., shifting the model's classification boundary)	Integrity	Data
	Manipulation of model tuning	Adversaries may fine-tune hyper-parameters and thus influence the AI system's behaviour. Hyper-parameters can be a vector for accidental overfitting. In addition, hard to detect changes to hyper-parameters would make an ideal insider attack. The usage of default hyper-parameters may increase transferability of adversarial attacks. When automatic optimization is done, the AI/ML might also be compromised in case the optimization algorithm is manipulated by adversaries.	Integrity Availability	Process Model
	Manipulation of optimization algorithm	Optimization algorithms are often used in the context of processes like Model Tuning to setup hyper-parameters values. Accordingly, nefarious abuse of such algorithms by adversaries may lead to erroneous tuning of models.	Availability	Models Processes
	Misclassification based on adversarial examples	This threat involves manipulation of model parameters or use of adversarial examples during inference to force misclassification of model results. This type of threats is related to the Actors category, as they have access to models and data sets, such is the case of Cloud providers, model providers and model users.	Integrity Availability	Actors Processes

Threat Category	Threat	Description	Potential impact	Affected assets
		Misclassification can also be instigated by Processes assets, such is the case of using adversarial examples during training and transfer learning stages, as well as during the inference stage.		
	ML model confidentiality	This threat refers to exploitation of the ML model to leak (in its outputs or otherwise) some information about its internal parameters or performance to un-authorized parties.	Confidentiality	Model
	ML Model integrity manipulation	This threat refers to manipulation of the ML model by delivering output values that were not generated based on its internal parameters or by delivering overtly biased or useless outputs (e.g. constant, or undistinguishable from random noise).	Integrity	Model
	Model backdoors	It is often the case that 3rd parties provide models (called as "Teacher" models), previously trained and fine-tuned with large datasets that are useful to learn from small datasets and/or by organizations without access to high computational clusters. The resulting models may be subject to backdoor threats that expose their inner working (breach of confidentiality), impact their operation (integrity breach) or degrade/cancel their performance (impact on availability).	Integrity Availability Confidentiality	Processes
	Model poisoning	This threat refers to a legitimate mode file being replaced entirely by a poisoned model file. In the context of AI as a Service, with many types of data and code being uploaded on cloud infrastructures, this threat may be realized by exploiting potential weaknesses of cloud providers.	Integrity Availability	Models Actors
	Model Sabotage	Sabotaging the model is a nefarious threat that refers to exploitation or physical damage of libraries and machine learning platforms that host or supply AI/ML services and systems.	Availability Integrity	Environment/tools Model
	Online system manipulation	This is related to model replacement by a malicious backdoored model, used for targeted or non-targeted attack, which can be exploited by Actors like Cloud Providers during Processes like model training or transfer.	Confidentiality Integrity	Model
	Overloading/confusing labelled dataset	Append attacks target availability by adding random samples to the training set to the point of preventing any model trained on that data set from computing any meaningful inference. The threat may lead to overfitting or underfitting the labelled dataset.	Availability	Data
	Reducing data accuracy	This threat refers to the reduction of the degree of data accuracy, by directly modifying the data or by mixing datasets with highly different degrees of quality.	Integrity Availability	Data
	Reduce effectiveness of AI/ML results	Users can make erroneous usage of AI services, either for not having a good understanding about the model capabilities or by not being able to	Integrity Availability	Processes Actors

Threat Category	Threat	Description	Potential impact	Affected assets
		understand when changes in the process imply model maintenance, and possibly re-training procedures. End-users can modify the input data to the model that results in “de-training” the model.		
	Sabotage	Sabotage involves intentionally destroying or maliciously affecting the IT infrastructure that supports AI systems.	Availability	Environment/tools
	Scarce data	AI relies on the availability of consistent and accessible data. This threat involves data scarcity (deliberately created by an adversary) that may compromise AI viability and/or compromise and limit its results. This can be exploited deliberately (for nefarious activities) or unintentionally during Data Ingestion.	Availability	Data Processes
	Transferability of adversarial attacks	ML and deep learning models are mostly based on an inductive approach to problem solving, as opposed to the deductive approach of traditional mathematical modelling. This means that experience matters and not always it can be given for granted that ML models can be smoothly transferred and applied in a new scenario and for a new AI application. This threat refers to adversarial examples that may be transferred to AI/ML applications and Environment/Tools like AI/ML libraries and machine learning platforms.	Integrity	Process Environment/tools
	Unauthorized access to data sets and data transfer process	These threats are seen in context of storage of data sets in infrastructures provided by third parties, which make them remotely accessible. The threat refers to unauthorized access of the data stored and unauthorized access to the inner workings of the data transfer process.	Confidentiality Integrity	Environment/tools
	Unauthorized access to models’ code	This threat refers to machine learning libraries and machine learning platforms being manipulated to inject malicious code that will exploit users models and gain access to datasets.	Confidentiality Integrity	Environment/tools Model
	White-box, targeted or non-targeted	This threat refers to misclassification to a specific target class or to a different class rather than the correct one. This type of threat is mainly associated with Processes assets like Model Selection/Building, Training, Testing, Transfer Learning and Model Deployment and can be exploited by Actors such as Model Providers.	Integrity Availability	Processes
Unintentional Damage	**Bias introduced by data owners**	Data Owners may try to hide information that will be fed to the AI systems as part of their business interests. Moreover, they are also people that may be biased themselves, as they tend to be far from raw data and may be incapable of giving good data to the models. This type of threat can severely affect trustworthiness of AI systems.	Availability Integrity	Actors

Threat Category	Threat	Description	Potential impact	Affected assets
	Compromising AI inference's correctness - data	This type of threats refers possible exploitations involving either data manipulation, or unintentional selection bias in raw data, or modification of labels and deletion or omission of labelled data items. It may also refer to compromising AI correctness by insertion of adversarial data in augmented data sets, as well as by means of interruption of training or modification of model parameters.	Integrity	Data
	Compromise and limit AI results	This type of threat can emerge due to involuntary or unintentionally actions from Data Owners, that may hide data due to business secrets or by not recognizing its value; by AI/ML designers and engineers, that can intentionally tamper or, due to lack of experience, miss to include data. This threat may also be related to AI/ML service users not being able to understand the model capabilities and/or results.	Integrity Availability	Model Actors Artefact
	Compromising ML inference's correctness – algorithms	Threats to the availability of the ML training algorithm, as well as threats that aim at compromising the training algorithm to adversely affect the desired accuracy.	Integrity Availability	Model
	Compromising ML training – augmented data	Threats to augmented datasets due to inconsistency with the training set they are derived from, and specifically when highly diverse, automatically generated data are added to a data set of collected data, which are very consistent but highly representative of their application domain, so there would be no need to limit overfitting. Enriching data always entails some risks. This threat can lead to non-satisfaction of functional requirements, i.e. poor inference.	Availability Integrity	Data
	Compromising feature selection	This threat refers to performance degradation of feature selection algorithms by delivering feature sets that are strongly predictive only for some for some classes, neglecting other features that are needed to discriminate difficult classes.	Integrity Availability	Model
	Compromise of data brokers/providers	This threat refers to compromising data brokers/providers to influence the machine learning process as they can deliberately or accidentally manipulate the data sent to the AI process, in several different ways: poisoning-via-insertion of malicious data; deleting registries to eliminate features either by changing the data, removing part of it or adding new registries. In addition, sometimes the mere data availability prevails over any consideration on data quality, with the risk that learning models are fed with data streams that do not reflect the statistical characteristics of a phenomenon and determining likely biases in the subsequent decisional processes.	Integrity Availability	Actor

Threat Category	Threat	Description	Potential impact	Affected assets
	Compromise of model frameworks	Model frameworks fail when are misconfigured or offer additional attack vectors with respect to traditional software, firmware and hardware environments. The ML platform's data volume and processing requirements mean that the workloads are often handled on the cloud, adding another level of complexity and vulnerability.	Integrity Availability	Artefacts Environment/tools
	Compromise privacy during data operations	Data modification or erroneous handling during Processes like Data Exploration' or Pre-Processing may lead to unintentional data breaches respectively and accordingly lead to legal concerns over privacy breaches.	Confidentiality	Data
	Disclosure of personal information	At all stages of the AI lifecycle, disclosure of personal information (either directly or by means of correlation) is a noteworthy threat. The threat is particularly manifested in the absence of verified data accuracy of sources, lack of data randomization, lack of pseudonymity mechanisms , etc.	Confidentiality	All assets
	Erroneous configuration of models	This type of threat materializes when models are used recklessly by end users (but also AI experts) without proper consideration of contextual factors that may not fit with the phenomenon being analyzed. If there is a mismatch between the goal and the model this may result in biases and discriminations or bad performance in general. Lack of expertise and proper knowledge of AI models' operation are the main cause of these erroneous configurations.	Integrity Availability	Processes Actors
	Label manipulation or weak labelling	This threat refers to supervised learning systems, which not infer correctly due to wrong or imprecise data labels. Messing with the labels may introduce the same effects of threats that are pertinent to adversaries attacking the labelling process.	Availability	Processes Data
	Lack of sufficient representation in data	Raw data assets fail when they are not sufficiently representative of the domain or unfit for the AI business goal, e.g. due to sample size and population characteristics. Data size does not always imply representativeness. If data selection is biased towards some elements that have similar characteristics (selection bias) then even a large sample will not deliver representative data. Assessment of data representativeness cannot be done a priori; it is only possible after identifying the targeted population and the purpose for collecting the data. Selection bias can be alleviated by employing balanced sampling techniques. Correction of existing data sets does, however, require information regarding the existence and nature of the bias; when selection bias is unknown to the AI model designer, no correction-based approaches to the inference process are possible.	Availability	Data

Threat Category	Threat	Description	Potential impact	Affected assets
	Manipulation of labelled data	Unintentional threats to labelled data items occur when enough numbers of labels and data are deleted/omitted by mistake, when a sufficient number of spurious labelled data is included into the data set, or when enough labels are modified Since the labelled data set is used for the purpose of training a ML model, all such modifications affect the model training and inference (e.g., shifting the model's classification boundary).	Integrity Confidentiality	Data
	Misconfiguration or mishandling of AI system	AI designers and developers may unintentionally expose data and models or may even misconfigure an AI system by mistake. Data confidentiality and trustworthiness are the main impacted security properties.	Confidentiality	Actors
	Mishandling of statistical data	This may happen, for instance, if maximum likelihood predictions are drawn from the sample, correctly reflecting the way the majority of individuals express a specific parameter that may not mirror the way the minority will be affected by the prediction. Also, other forms of unintended bias may take place. For instance, in ranking algorithms even if parameters under analysis may be ranked fairly and in the correct order, the rewards allocated to each "slot" (click through rates, impressions or any other sort of share of "good" to allocate) may not be fairly distributed, with limited possibility to rebalance such uneven situations.	Availability Confidentiality	Data
	ML Model Performance Degradation	The performance of an AI's system ML model may degrade due to the data governance policy, by omission or by corruption due to system crashes or loss of network connectivity.	Availability Confidentiality	Process Model
	Online system manipulation	This is related to model replacement by a backdoored model by mistake, used for targeted or non-targeted attack. This can be the result of unintended actions by Actors like Cloud Providers during Processes like model training or transfer.	Confidentiality Integrity	Model
	Reducing data accuracy	This threat refers to the reduction of the degree of data accuracy, by directly modifying the data or by mixing datasets with highly different degrees of quality.	Integrity	Data
Legal	**Compromise privacy during data operations**	Data manipulation or erroneous handling during Processes like Data Exploration' or Pre-Processing may lead to intentional or unintentional data breaches respectively and accordingly lead to legal concerns over privacy breaches.	Confidentiality	Data

Threat Category	Threat	Description	Potential impact	Affected assets
	Corruption of data indexes	Data indexes threats manifest when their content becomes corrupted. Corruption may be the result of an attack, or due to system crashes or loss of network connectivity during index replication. The same events may cause interruptions of index construction tasks, bringing a partially built (and therefore defective) index to production. Also, running out of storage capacity during indexing or replication may cause an entire data index to be deleted. Denial-of-service attacks to indexes intentionally corrupt data indexes to decrease the performance of data access. Additionally, timing attacks to indexes use access time to public items before and after inserting them (which depend on the index content) to infer the presence and size of inaccessible data items.	Integrity Availability	Artefacts
	Disclosure of personal information	At all stages of the AI lifecycle, disclosure of personal information (either directly or by means of correlation) is a noteworthy threat. The threat is particularly manifested in the absence of verified data accuracy of sources, lack of data randomization, lack of pseudonymity mechanisms , etc.	Confidentiality	All assets
	Lack of data governance policies	When personal data are processed, the existence of data governance policies is part of data controller's accountability. The GDPR promotes the implementation of data protection by design measures as a way to be effective in the implementation of data protection principles, and in situation of high risks requires the implementation of a data protection impact assessment (DPIA). Data controllers should identify measurable goals and performance indicators that give evidence, also in a quantitative way, of their level of compliance with the principles and implement a DPIA as default option.	Integrity Confidentiality	Artefacts Data
	Lack of data protection compliance of 3rd parties	Third parties are frequently used in providing and processing data, either directly or by means of libraries and models that they provide. This threat refers to the lack of compliance of the third parties with respect to applicable data protection regulations.	Confidentiality	Actors
	Profiling of end users	Labeling may lend itself to a potential threat to anonymity and privacy by acting as a form of profiling.	Confidentiality	Data
	SLA breach	In the context of 3rd party dependencies, breach of contractual obligations and Service Level Agreements (SLAs) may lead to degradation of performance or even unavailability of the AI system to perform its operation.	Availability	Environment/tools

Threat Category	Threat	Description	Potential impact	Affected assets
	Vendor lock-in	When considering third parties to AI systems, e.g. cloud providers, data storage, AI libraries, etc. the threat of vendor lock-in involves the reliance on a sole third part provider without realistic alternative. While this might not constitute necessarily a cybersecurity threat, the lack of backup and overprovisioning might hamper operations.	Availability	Environment/tools
	Weak requirements analysis	AI requirements may fail when they are built in isolation from the social circumstances that make AI applications necessary. Specifically, functional requirements of AI systems about executing AI tasks with the needed accuracy may fail by not taking into account the impact of the corresponding inherent bias. Non-functional requirements of AI systems may fail by not considering the severity of information leaks and disclosures that can happen in virtualized high-performance execution environments, or when using untrusted software libraries. This is particularly dangerous for AI systems working in domains like healthcare, biotechnology, financial services and law.	Availability	Artefacts
Failures or malfunctions	**Compromising AI application viability**	This type of threat refers to lack of understanding of what AI/ML are and how to succeed with the business models.	Availability	Artefacts
	Compromising ML pre-processing	Flaws or defects of the data and metadata schemata greatly influence the quality of the analysis by applications that use the data. In AI applications, a flawed schema will negatively impact on the quality of the ingested information. Flaws often result from of inconsistencies in the use of modeling methodologies.	Integrity	Data Artefacts
	Corruption of data indexes	Data indexes threats manifest when their content becomes corrupted. Corruption may be the result of an attack, or due to system crashes or loss of network connectivity during index replication. The same events may cause interruptions of index construction tasks, bringing a partially built (and therefore defective) index to production. Also, running out of storage capacity during indexing or replication may cause an entire data index to be deleted.	Integrity Availability	Artefacts
	Compromise of model frameworks	Model frameworks fail when are misconfigured or offer additional attack vectors with respect to traditional software, firmware and hardware environments. The ML platform's data volume and processing requirements mean that the workloads are often handled on the cloud, adding another level of complexity and vulnerability.	Integrity	Artefacts Environment/tools
	Errors or timely restrictions due to non-reliable data infrastructures	This type of threat is related to data and computational exposure and/or inadequate capacity that may expose data and compromise privacy preservation.	Integrity Availability	Environment/tools

Threat Category	Threat	Description	Potential impact	Affected assets
	Inadequate/absent data quality checks	Given the importance of data and the need for data to hold markers of their quality (e.g. in terms of sample size, variances, applied data collection methodologies, real vs synthetic data provenance), the lack of or the inadequacy of data quality checks may lead to poor performance of an AI system.	Availability Confidentiality	Data Processes
	Label manipulation or weak labelling	This threat refers to supervised learning systems, which not infer correctly due to wrong or imprecise data labels. If adversaries can only modify the training labels with some or all knowledge of the target model, they need to find the most vulnerable labels. Random perturbation of labels is one possible attack, while additionally there is the case of adversarial label noise (intentional switching of classification labels leading to deterministic noise, an error that the model cannot capture due to its generalization bias).	Confidentiality Integrity	Processes Data
	Lack of documentation	This threat generally manifests over the course of time. In AI systems, model selection should be made in a framework of accountability and trust and “black-box” approaches should be avoided. At any stage the choice of algorithm parameters should be justified and duly documented. Discarded alternatives should be disclosed and the consequences of model under-fitting or overfitting should be clearly explained. This set of parameters and design choices are important to be able to identify potential errors (intentional or unintentional). Failure to properly maintain documentation of the AI system threatens to indirectly limit its failsafe operation.	Integrity Availability	Processes
	ML Model Performance Degradation	The performance of an AI’s system ML model may degrade due to the data governance policy, by omission or by corruption due to system crashes or loss of network connectivity.	Availability	Process Model
	Poor resource planning	The proper functioning of an AI system may be compromised by the lack of adequate computational resources (storage capacity, transmission speed, computational power). This is particularly relevant in real time application and in the health sector. In order to deliver the expected beneficial outcome, it is essential that these resources are correctly dimensioned and allocated, and that the final user is aware of such infrastructural constraints. It is part of the informational accountability of the developer to provide user, and with prominent means, the list of resources to arrange, and their configuration settings, necessary to avoid failures and impacts on the functioning of an AI system.	Integrity Availability	Artefacts Environment/tools
	Scarce data	AI relies on the availability of consistent and accessible data. This threat involves data scarcity that may compromise AI viability and/or compromise and limit its results. This can be exploited deliberately (for nefarious activities) or unintentionally during Data Ingestion.	Availability	Data Processes

Threat Category	Threat	Description	Potential impact	Affected assets
	Stream interruption	This threat relates to the interruption of data streams, during processes like data ingestion and training. The lack of data and data interruption, in case of stream processing, can cause failures in an AI/ML system	Confidentiality Integrity Availability	Processes
	Weak data governance policies	In AI applications, data governance policies have been known to fail for defective data metrics, absence of documentation and lack of adaptability. Specifying data quality metrics for ML training is not straightforward, often leading to lack of numerical targets and insufficient documentation of data governance policies. Failure to monitor/record AI data usage (e.g. for training, testing or validation), and to update data governance policies based on AI systems achievements and failures is another common pitfall.	Confidentiality Integrity	Artefacts Data
	Weak requirements analysis	AI requirements may fail when they are built in isolation from the social circumstances that make AI applications necessary. Specifically, functional requirements of AI systems about executing AI tasks with the needed accuracy may fail by not taking into account the impact of the corresponding inherent bias. Non-functional requirements of AI systems may fail by not considering the severity of information leaks and disclosures that can happen in virtualized high-performance execution environments, or when using untrusted software libraries. This is particularly dangerous for AI systems working in domains like healthcare, biotechnology, financial services and law.	Integrity Confidentiality Availability	Artefacts
	3rd party provider failure	Failures or malfunctions of 3rd party providers, e.g. cloud providers, data storage providers, AI as Service providers, etc. may lead to unavailability of an AI system and improper or delayed operation.	Availability Confidentiality	Environment/tools
Eavesdropping **Interception** **Hijacking**	**Data inference**	This threat may be exploited by the Data Providers and Model Providers, and can lad to inference of data. Evidently, in the case of personal data, such inference raises concerns in terms of privacy and/or discrimination.	Confidentiality	Data
	Data theft	This threat may manifest during the transportation of data, during Processes like Data Ingestion and in the context of access to data storage means. In these cases, data may be intercepted and stole.	Confidentiality Integrity	Data
	Model Disclosure	Threat of leaking information about trained and/or tuned models internal parameters and other settings of models. .	Confidentiality	Model
	Stream interruption	This threat relates to the interruption of data streams, during processes like data ingestion and training. The lack of data and data interruption, in case of stream processing, can cause failures in an AI/ML system	Confidentiality Integrity Availability	Processes

Threat Category	Threat	Description	Potential impact	Affected assets
	Weak encryption	In the context of AI, this threat is related to assets of the Processes category, and refers to potential eavesdropping of data or hijacking of communications in the case of data transfers/storage/processing. The threat when materialized may expose data sets and even personal and sensitive information.	Confidentiality Integrity	Data Processes Environment/tools
Physical attacks	**Communication networks tampering**	Tampering of communication networks may lead to their unavailability and thus is a major threat that may be exploited by adversaries. The corresponding outages may lead to delays in decision-making, delays in the processing of data streams and entire AI systems being placed offline. Moreover, side-channel attacks may expose private and sensitive information that traverses communication networks.	Confidentiality Availability	Environment/tools
	Errors or timely restrictions due to non-reliable data infrastructures	This type of threat is related to data and computational exposure and/or inadequate capacity that may expose data and compromise privacy preservation.	Integrity Availability	Environment/tools
	Infrastructure/system physical attacks	Physical attacks against infrastructure (IT and corporate services) that supports AI systems' operation and maintenance are a potential threat. Manifestation of this threat leads to degraded performance and even unavailability. The threat manifests by physically attacking, dismantling or even destroying the actual physical infrastructure.	Availability	Environment/tools
	Model Sabotage	Sabotaging the model is a nefarious threat that refers to exploitation or physical damage to hardware hosting libraries and machine learning platforms that host or supply AI/ML services and systems.	Availability	Environment/tools Model
	Sabotage	Sabotage involves intentionally destroying or maliciously affecting the IT infrastructure that supports AI systems.	Availability	Environment/tools
Outages	**Communication networks outages**	Outages to communication networks may adversely influence the performance and operation of AI systems. Such outages may lead to delays in decision-making, delays in the processing of data streams and entire AI systems being placed offline.	Availability	Environment/tools
	Infrastructure/system outages	Outages to infrastructure (IT and corporate services) that supports AI systems' operation and maintenance. Manifestation of this threat leads to degraded performance and even unavailability.	Availability	Environment/tools

Threat Category	Threat	Description	Potential impact	Affected assets
Disasters	**Environmental phenomena (heating, cooling, climate change)**	Environmental phenomena may adversely influence the operation of IT infrastructure and hardware systems that support AI systems. Climate change in particular has been consistently highlighted in ENISA reports on telecom incident reporting as the main cause of telecom outages. Such outages may lead to delays in decision-making, delays in the processing of data streams and entire AI systems being placed offline.	Availability	Environment/tools
	Natural disasters (earthquake, flood, fire, etc.)	Natural disasters may lead to unavailability or destruction of the IT infrastructures and hardware that enables the operation, deployment and maintenance of AI systems.	Availability	Environment/tools

ANNEX C – MAPPING OF ASSETS TO AI LIFECYCLE

Figure 7: AI lifecycle stages

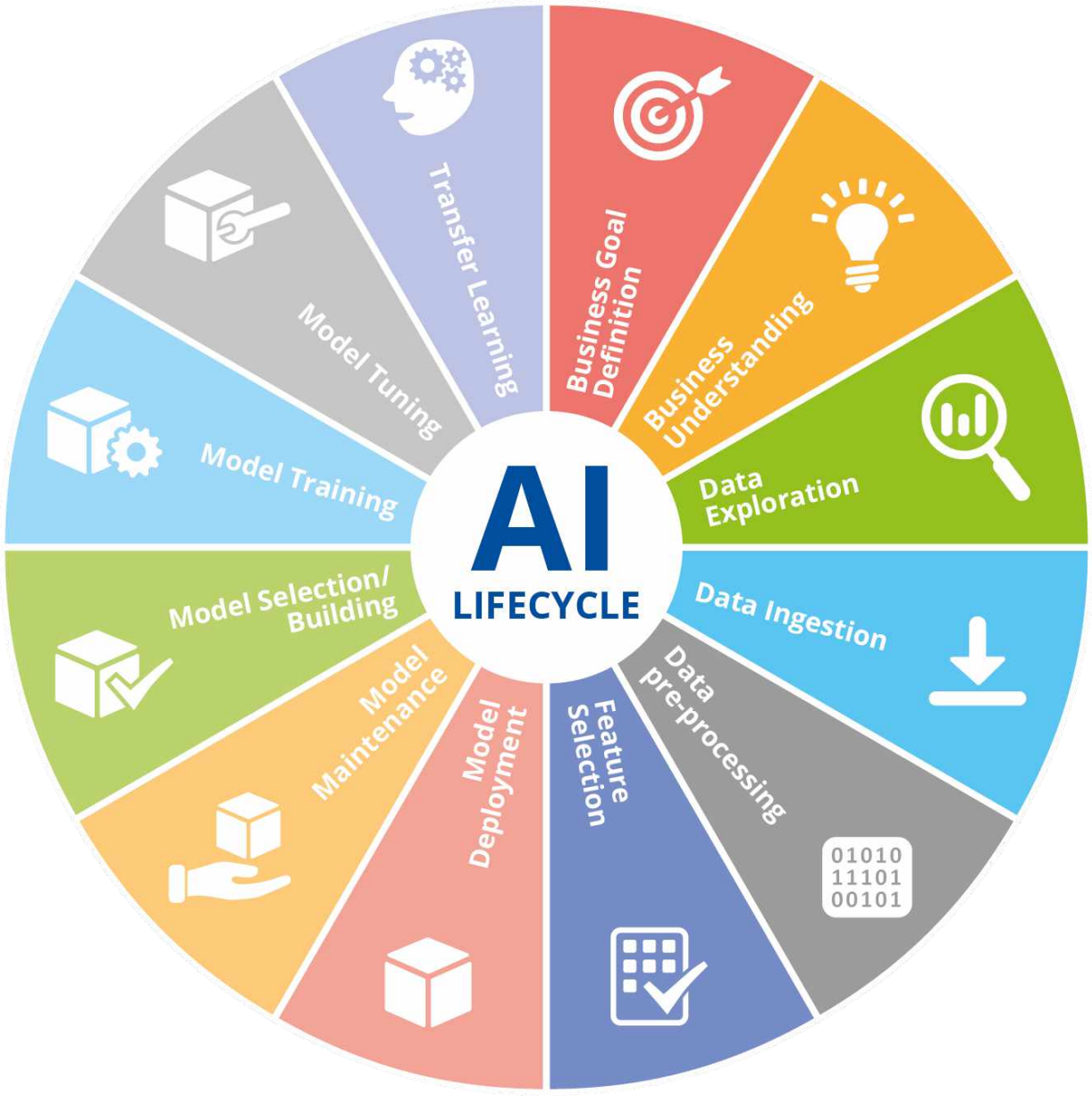
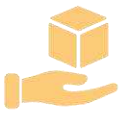


Figure 8: Mapping of assets to AI lifecycle stages



MODEL MAINTENANCE

- Model
- Data Scientists / AI designer/ AI developer
- Data Engineers
- Service consumers / Model users

- End Users
- Model Maintenance
- Computational platforms
- Machine Learning Platforms
- Monitoring Tools
- Operating System/software

- Model frameworks, software, firmware or hardware incarnations.
- Composition artefacts: AI models compositions

MODEL SELECTION/ BUILDING

- Training Data
- Training Algorithms
- Model
- Training parameters
- Data Scientists / AI designer/ AI developer
- Data Engineers

- Model selection/building, training and testing
- Computational platforms
- Integrated Development Environment
- Machine Learning Platforms
- Libraries (with algorithms for transformation, labelling, etc)

- Monitoring Tools
- Composition artefacts: AI models compositions
- Model Architecture
- Model hardware design

MODEL TRAINING

- Training Data
- Testing Data
- Algorithms
- Model
- Model Parameters
- Model Performance
- Training Parameters

- Trained Models
- Data Scientists / AI designer/ AI developer
- Data Engineers
- Cloud Provider
- Cloud
- Computational platforms

- Integrated Development Environment
- Machine Learning Platforms
- Libraries (with algorithms for transformation, labelling, etc)
- Monitoring Tools
- Composition artefacts: AI models compositions

MODEL TUNING

- Validation Data Set
- Evaluation Data
- Model
- Model Performance
- Hyper-parameters
- Tuned Model
- Data Scientists / AI designer/ AI developer

- Data Engineers
- Cloud Provider
- Model Tuning
- Cloud
- Computational Platforms
- Integrated Development Environment
- Machine Learning Platforms

- Libraries (with algorithms for transformation, labelling, etc)
- Monitoring Tools
- Optimization techniques
- Composition artefacts: AI models compositions

TRANSFER LEARNING

- Training Data
- Trained Models
- Data Scientists / AI designer/ AI developer
- Model Provider
- Model adaptation – transfer learning / Model deployment

- Computational Platforms
- Monitoring Tools
- Model frameworks, software, firmware or hardware incarnations

ANNEX D – MAPPING OF THREATS TO AI LIFECYCLE

Figure 9: AI lifecycle stages

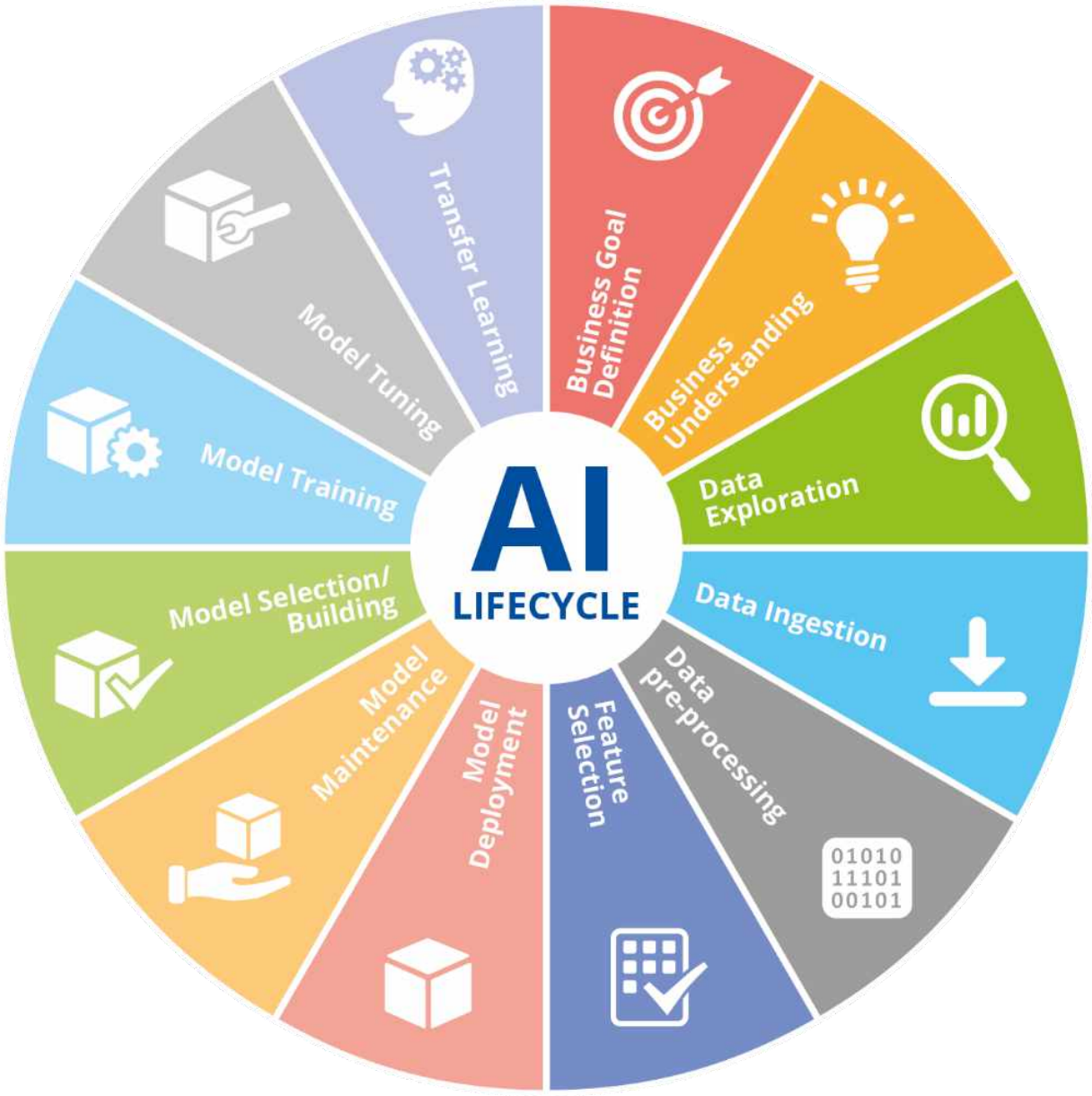


Figure 10: Mapping of threats to AI lifecycle stages

BUSINESS GOAL DEFINITION

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Data poisoning
- Data tampering
- Insider threat
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Model Sabotage
- Compromise of data brokers/ providers
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Compromise and limit AI results
- Misconfiguration or mishandling of AI system
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Disclosure of personal information
- Compromise of model frameworks
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Compromise of model frameworks
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

BUSINESS UNDERSTANDING

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Data poisoning
- Data tampering
- Insider threat
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Model Sabotage
- Compromise of data brokers/ providers
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Compromise and limit AI results
- Misconfiguration or mishandling of AI system
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Disclosure of personal information
- Compromise of model frameworks
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Compromise of model frameworks
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

DATA EXPLORATION

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising AI inference's correctness - data
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor/insert attacks on training datasets
- Overloading/confusing labelled dataset
- Compromising ML training – validation data
- Compromising ML training – augmented data
- Adversarial examples
- Reducing data accuracy
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference's correctness - data
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training – augmented data
- Reducing data accuracy
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Data inference
- Data theft
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

DATA INGESTION

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising AI inference's correctness - data
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor/insert attacks on training datasets
- Overloading/confusing labelled dataset
- Compromising ML training – validation data
- Compromising ML training – augmented data
- Adversarial examples
- Reducing data accuracy
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference's correctness - data
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training – augmented data
- Reducing data accuracy
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Data inference
- Data theft
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

01010
11101
00101

DATA PRE-PROCESSING

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising AI inference's correctness - data
- Compromising ML inference's correctness – algorithms
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor/insert attacks on training datasets
- Overloading/confusing labelled dataset
- Compromising ML training – validation data
- Compromising ML training – augmented data
- Adversarial examples
- Adversarial examples
- Reducing data accuracy
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference's correctness - data
- Compromising feature selection
- Compromising ML inference's correctness – algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training – augmented data
- Reducing data accuracy
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Data inference
- Data theft
- Model Disclosure
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

FEATURE SELECTION

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising AI inference's correctness - data
- Compromising ML inference's correctness – algorithms
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor/insert attacks on training datasets
- Overloading/confusing labelled dataset
- Compromising ML training – validation data
- Compromising ML training – augmented data
- Adversarial examples
- Reducing data accuracy
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference's correctness - data
- Compromising feature selection
- Compromising ML inference's correctness – algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training – augmented data
- Reducing data accuracy
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Data inference
- Data theft
- Model Disclosure
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

MODEL DEPLOYMENT

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising ML inference's correctness – algorithms
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Adversarial examples
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Compromise and limit AI results
- Compromising feature selection
- Compromising ML inference's correctness – algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Disclosure of personal information
- Compromise of model frameworks
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Compromise of model frameworks
- Model Disclosure
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

MODEL MAINTENANCE

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising ML inference's correctness – algorithms
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Adversarial examples
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromising feature selection
- Compromising ML inference's correctness – algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Model Disclosure
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

MODEL SELECTION/ BUILDING

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising AI inference's correctness - data
- Compromising ML inference's correctness – algorithms
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor/insert attacks on training datasets
- Overloading/confusing labelled dataset
- Compromising ML training – validation data
- Compromising ML training – augmented data
- Adversarial examples
- Reducing data accuracy
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference's correctness - data
- Compromising feature selection
- Compromising ML inference's correctness – algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training – augmented data
- Reducing data accuracy
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Data inference
- Data theft
- Model Disclosure
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

MODEL TRAINING

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising AI inference's correctness - data
- Compromising ML inference's correctness – algorithms
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor/insert attacks on training datasets
- Overloading/confusing labelled dataset
- Compromising ML training – validation data
- Compromising ML training – augmented data
- Adversarial examples
- Reducing data accuracy
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference's correctness - data
- Compromising feature selection
- Compromising ML inference's correctness – algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training – augmented data
- Reducing data accuracy
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Data inference
- Data theft
- Model Disclosure
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

MODEL TUNING

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising AI inference's correctness - data
- Compromising ML inference's correctness – algorithms
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor/insert attacks on training datasets
- Overloading/confusing labelled dataset
- Compromising ML training – validation data
- Compromising ML training – augmented data
- Adversarial examples
- Reducing data accuracy
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference's correctness - data
- Compromising feature selection
- Compromising ML inference's correctness – algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training – augmented data
- Reducing data accuracy
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Data inference
- Data theft
- Model Disclosure
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

TRANSFER LEARNING

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models' code
- Compromise and limit AI results
- Compromising AI inference's correctness - data
- Compromising ML inference's correctness – algorithms
- Data poisoning
- Data tampering
- Elevation-of-Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White-box, targeted or non-targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor/insert attacks on training datasets
- Overloading/confusing labelled dataset
- Compromising ML training – validation data
- Compromising ML training – augmented data
- Adversarial examples
- Reducing data accuracy
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers/ providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List (ACL) manipulation
- Compromising ML pre-processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI/ML results
- Label manipulation or weak labelling
- Model backdoor
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference's correctness - data
- Compromising feature selection
- Compromising ML inference's correctness – algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training – augmented data
- Reducing data accuracy
- Compromise of data brokers/ providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3rd parties
- Vendor lock-in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal information
- Corruption of data indexes
- Compromising AI application viability
- Errors or timely restrictions due to non-reliable data infrastructures
- 3rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate/absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre-processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks
- Data inference
- Data theft
- Model Disclosure
- Stream interruption
- Weak encryption
- Errors or timely restrictions due to non-reliable data infrastructures
- Model Sabotage
- Infrastructure/system physical attacks
- Communication networks tampering
- Sabotage
- Infrastructure/system outages
- Communication networks outages
- Natural disasters (earthquake, flood, fire, etc)
- Environmental phenomena (heating, cooling, climate change)

TP-01-20-772-EN-N

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies, and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure, and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found at www.enisa.europa.eu.

ENISA
European Union Agency for Cybersecurity

Athens Office
1 Vasilissis Sofias Str
151 24 Marousi, Attiki, Greece

Heraklion office
95 Nikolaou Plastira
700 13 Vassilika Vouton, Heraklion, Greece

enisa.europa.eu

ISBN 978-92-9204-462-6
DOI 10.2824/238222