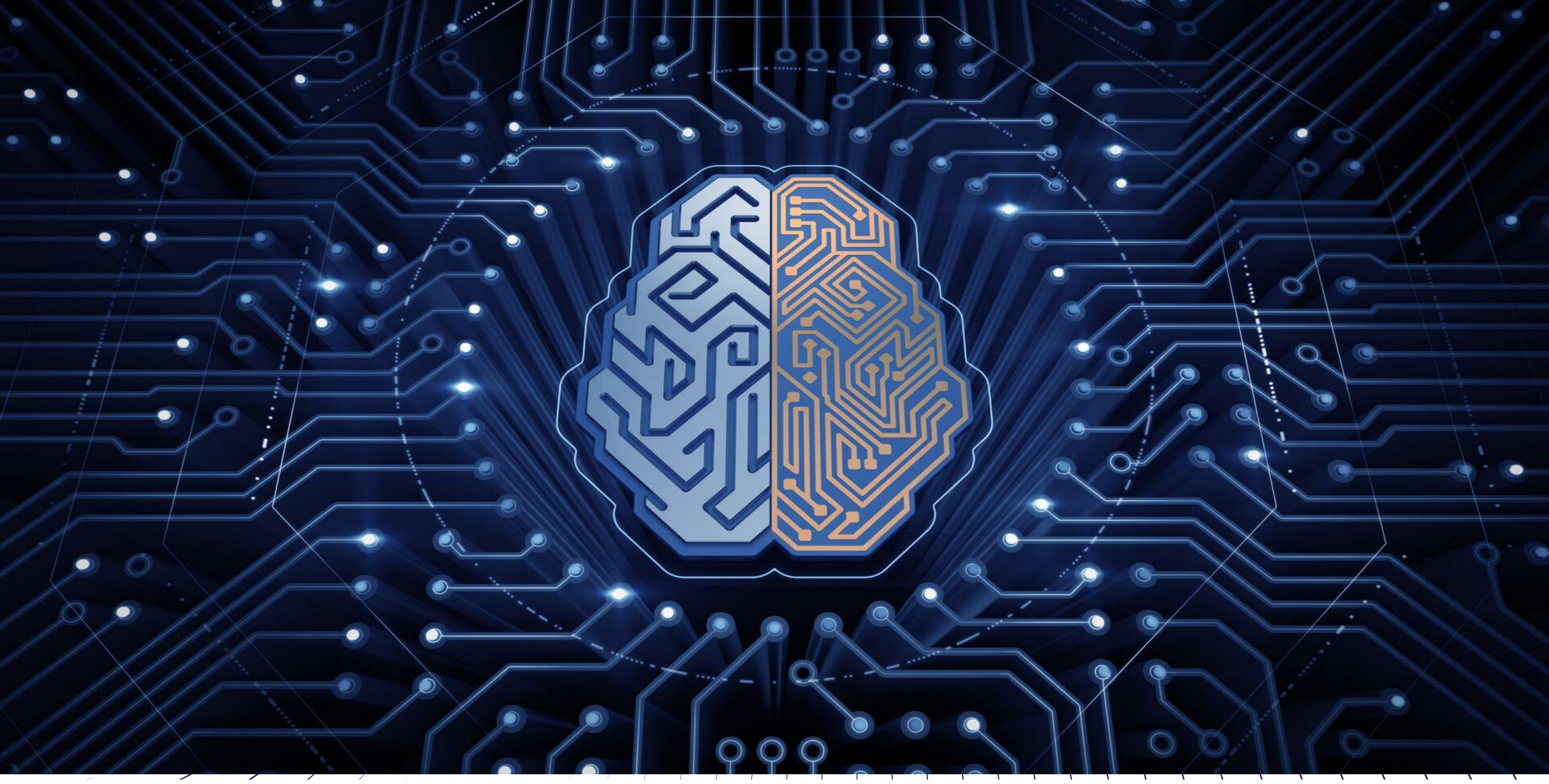


Realising Lawful and Trustworthy AI

An Analysis of the Requirements of Trustworthy AI and EU AI Act

Frederik Olsen Andersen

Cyber Security, 2022-06

Master's Thesis

Student Report
Aalborg University

Realising Lawful and Trustworthy AI

An Analysis of the Requirements of Trustworthy AI and EU AI Act

Frederik Olsen Andersen

Cyber Security, 2022-06

Master’s Thesis

STUDENT REPORT
AALBORG UNIVERSITY

Figures present in this thesis, which have not been copied directly from their respective sources, have been created using either `draw.io` or `miro.com`

AALBORG UNIVERSITY
STUDENT REPORT

Department of Electronic Systems
Aalborg University
http://www.aau.dk

Title:
Realising Lawful and Trustworthy AI: An Analysis of the Requirements of Trustworthy AI and the EU AI Act

Theme:
Master's Thesis

Project Period:
Spring Semester 2022

Project Group:
N/A

Participant(s):
Frederik Olsen Andersen

Supervisor(s):
Henning Olesen
Niels Peter Anglov

Copies: 1

Page Numbers: 78

Date of Completion:
June 10, 2022

Abstract:

This thesis examines two particular EU AI initiatives; the Guidelines for Trustworthy AI made by the EU High-level Experts Group on AI (AI HLEG), as well as the proposed regulation on AI known as the EU AI Act. The AI HLEG's guidelines provide requirements for what constitutes Trustworthy AI, whereas the AI Act ranks AI systems based on the risk they pose to things such as fundamental rights, democracy and safety, and sets out a series of legal requirements for AI systems which are considered high-risk. This thesis aims to analyse how the requirements presented in each of these initiatives can be implemented and realised. The results of this analysis is a series of recommendations for which methods, techniques or practises to best implement the respective requirement. Additionally, these recommendations have been used to make an operational model for realising the analysed requirements.

Contents

Preface

Aalborg University, June 10, 2022

Frederik Olsen Andersen
<foan20@student.aau.dk>

Chapter 1

Introduction

Since the start of the 2000's, the world has seen a significant growth in the development and use of *Artificial Intelligence* (AI) systems in businesses, industry and society. AI systems are used to automate manufacturing, calculate fuel-efficient shipping routes, predict medical conditions and secure computer networks from attackers.

An example use case demonstrating AI systems technological maturity is autonomous vehicles. The absence of a human driver puts great responsibility on the AI system in command. In order to perform at a level where this responsibility is warranted, the AI system needs to be very accurate, precise and reliable in its capabilities. This level of performance often comes at the cost of increased complexity in the AI system [1], leading to several problems with regards to analysing, examining and most importantly verifying that such an AI system functions as intended.

This issue of increased complexity posing a risk to the analysis and examination of AI systems is called the AI *explainability* problem and will be briefly introduced in section 1.2.

The issue of complexity in AI systems is part of one of the more fundamental challenges of AI. Namely the issue of achieving trust in AI. Since the consequences of erroneous behaviour of AI systems may be dangerous to humans, or in some cases even fatal [2], the need for establishing trust in AI is paramount. Using its position as an economic and regulatory powerhouse, the *European Union* (EU) has put forth a series of initiatives for achieving trust in AI. These will be introduced in section 1.3.

However, in order to understand where these issues came from, we first have to understand where AI came from. A brief introduction to the origins of AI will be presented in section 1.1.

1.1 Origins of Artificial Intelligence

In 1958 at the Cornell Aeronautical Laboratory, Frank Rosenblatt invented the *Perceptron* [3]; an image recognition algorithm which was mathematically "trained", using vectors and numerical weights, to reliably tell apart two images, of primarily geometric shapes. The idea for this algorithm was to mimic the interactions observed in natural neurons in order to train the model. The Perceptron is thus considered one of the first applications of *Machine Learning* (ML).

Machine learning is the process of using algorithms and mathematical models to enable a computer to identify patterns in data. Sample data, called *training data* in a ML context, is used to create a data model which can make predictions for a given input, based on patterns observed in the sample data [4]. The process of creating a data model from patterns observed in the training data is called *training* the model, and is what enables a computer to learn. Figures 1.1 and 1.2 show a holistic view of machine learning compared to traditional programming.

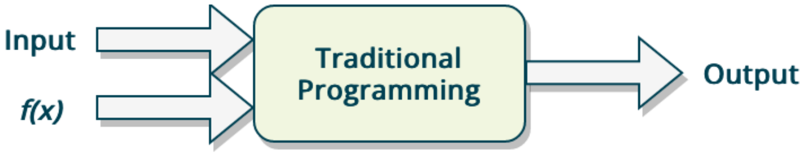


Figure 1.1: Holistic view of traditional programming.

Traditional programming relies on the developer specifying a set of instructions to be performed on the input data in order to generate a desired output. A set of instructions; an algorithm, is equivalent to a mathematical function as all instructions are converted to binary operations on the computer's CPU.

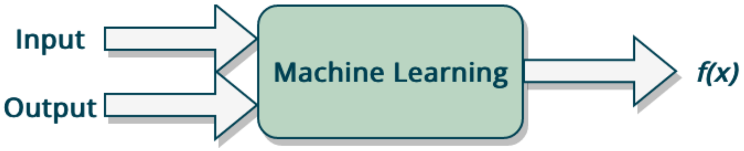


Figure 1.2: Holistic view of machine learning.

In contrast; machine learning learns, through model training, the function that generates the desired output.

Artificial Intelligence is the science of making intelligent machines and computers [5]. The definition of what constitutes intelligent in this context is subject to debate [4], however, one of the most common definitions seeks to mimic the intelligent tasks performed by humans. Under this definition an AI system would require one or more of the following capabilities [4]:

- **Natural language processing** to communicate successfully in a human language.
- **Knowledge representation** to store what it knows or hears.
- **Automated reasoning** to answer questions and to draw new conclusions.
- **Machine learning** to adapt to new circumstances and to detect and extrapolate patterns.
- To interact with the physical world the system would also need:
  - **Computer vision** and **speech recognition** to perceive the world.
  - **Robotics** to manipulate objects and move about.

Many people will recognize examples of AI systems in our everyday lives; personal digital assistants such as Apple's *Siri* and Amazon's *Alexa* and partially autonomous vehicles such as the *Tesla* cars are examples of AI systems which posses one or more of the aforementioned capabilities.

1.2 The AI Explainability Problem

From the humble beginnings of the Perceptron, the field of machine learning has since advanced significantly. Increasingly complex ML models such as deep neural networks, which are exceeding human speed and precision in a multitude of tasks, are being created to solve different problems.

In particular, many ML models are specifically designed for classification tasks, e.g. Image recognition, where precision and accuracy are important metrics for evaluating model performance. As a consequence of model developers striving to increase model performance, models often tend to grow increasingly complex. This inadvertently has lead to many utilized ML models essentially becoming *black box* models, in which the internal logic of the model can not easily be examined and analyzed.

The black box nature of such models means that they only provide the results of their computation, and not usable information on *how* or *why* the model came to a particular conclusion.

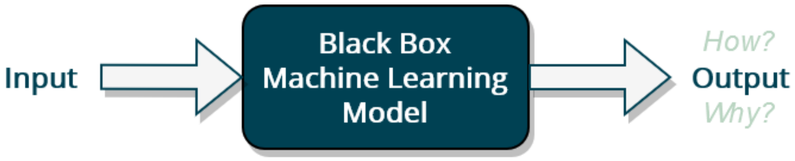


Figure 1.3: Black Box behaviour of complex ML models.

The explainability problem, posed by this issue, has led to the creation of the field of *eXplainable Artificial Intelligence* (XAI). Research in this field aims to provide sound, generalized methods for explaining the logic in black box ML models, as well as provide an understanding of how humans can best interpret the explanations provided by such methods. More information on this topic will be presented in section 2.3.

1.3 Achieving Trust in AI Systems

As the use of artificial intelligence and machine learning in businesses and society increases, the issue achieving trust in AI systems becomes increasingly important. Users interacting with AI systems, will require and expect AI systems to work in the interest of the user and as a force of good.

As a part of the *Coordinated Plan for Artificial Intelligence* program, first launched in 2018 by the *European Commission* and later reviewed in 2021 [6], [7], several initiatives to achieving trust in AI systems have emerged. In particular, these initiatives seek to address the need for guidelines for AI as a means to drive the uptake of AI systems in society.

The program led to the creation of, among other initiatives; the Danish *National Strategy for AI* [8], the *Ethics Guidelines for Trustworthy AI* made by the European Commission High-Level Expert Group on AI (AI HLEG) [9] as well a proposed legal framework for regulating AI systems (European AI Act) [10], [11].

Figure 1.4 shows a timeline of the initiatives in the program.

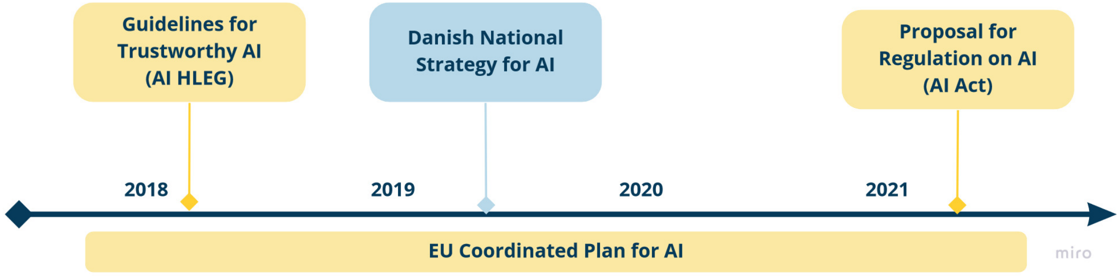

Guidelines for Trustworthy AI (AI HLEG)
Danish National Strategy for AI
Proposal for Regulation on AI (AI Act)
2018
2019
2020
2021
EU Coordinated Plan for AI
miro

Figure 1.4: Timeline of the EU coordinated plan for AI.

1.3.1 Defining Trustworthy AI

In the guidelines for Trustworthy AI the AI HLEG specify that for an AI system to be considered trustworthy, it must be considered as having three components; it should be *lawful*, *ethical* and *robust*.

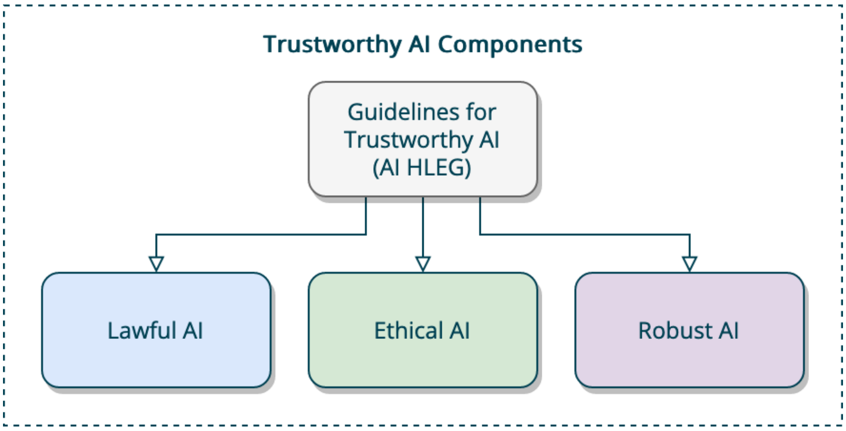

Trustworthy AI Components
Guidelines for Trustworthy AI (AI HLEG)
Lawful AI
Ethical AI
Robust AI

Figure 1.5: Components of Trustworthy AI according to the AI HLEG.

Here it is important to note that the AI HLEG specifically states that they do not cover the legal aspects of Trustworthy AI. Meaning that the Lawful AI component is not regarded in the presented guidelines. Instead the they define their purpose as creating a concise definition of what constitutes Trustworthy AI, as well as provide the guidelines and requirements needed to achieve this.

Looking at a subset of the requirements we see the notion of *transparency* or *explainability* along with topics such as *privacy*, *data protection*, as well as *security* and *robustness* [9]. All of the requirements mentioned in the guidelines will be presented in section 2.1.

Here, security relates to the general security of the general IT system and/or product in which the AI system is placed. In contrast, robustness relates to how resilient a ML model is to attack and manipulation. Figure 1.6 illustrates some of common attacks on ML models specifically mentioned in the AI Act proposal [6]:

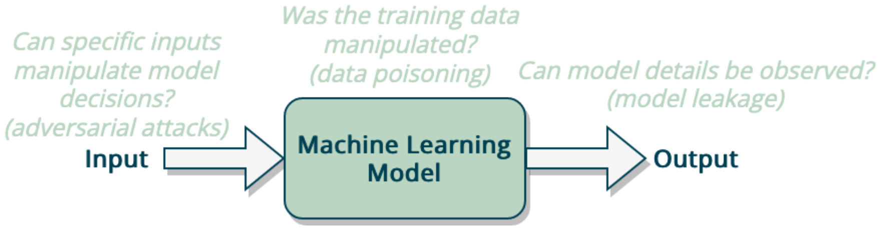

Was the training data manipulated? (data poisoning)
Can specific inputs manipulate model decisions? (adversarial attacks)
Can model details be observed? (model leakage)
Input
Machine Learning Model
Output

Figure 1.6: Attacks on model robustness.

1.3.2 The AI Act

In order to be compliant with the proposed legislation, the AI Act requires providers of high-risk AI systems (more on this in section 2.2) to document the design, development and quality management of the system as well as provide technical documentation for the system. Documentation of, among other things; model behaviour, bias, accuracy, and robustness need to be part of this documentation [11].

The aforementioned requirements are required in order to achieve the European marking of conformity called the *CE* marking. In order to obtain a CE marking for a high-risk AI system, a provider is required to perform a *conformity assessment procedure* in order to document compliance with the new regulation. The process of obtaining and affixing a CE marking to a high-risk AI system generally follows the steps shown in figure 1.7.

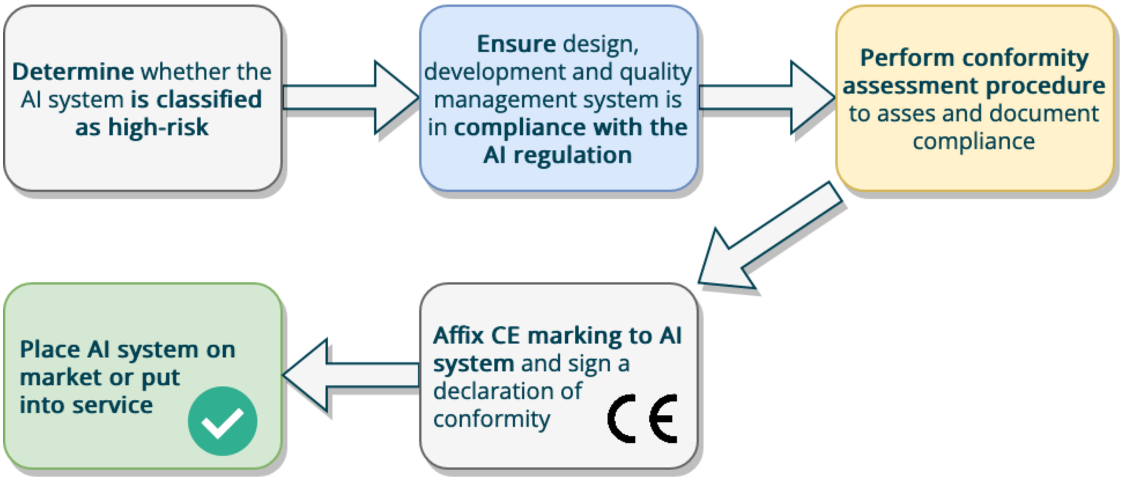

Determine whether the AI system is classified as high-risk
Ensure design, development and quality management system is in compliance with the AI regulation
Perform conformity assessment procedure to asses and document compliance
Affix CE marking to AI system and sign a declaration of conformity
CE
Place AI system on market or put into service

Figure 1.7: CE marking process for high-risk AI systems. – Figure based on [12]

1.4 Problem Formulation

The issue of achieving trust in AI has been addresses by the AI HLEG in their guidelines for Trustworthy AI. These are presented as a set of guidelines (called requirements by the AI HLEG) to implement in order to achieve Trustworthy AI and to foster an awareness of the ethical and trust-related issues of AI. In addition to the guidelines set out by the AI HLEG, the AI Act proposal comes with it own set of requirements for AI systems.

Being able to realise the aforementioned requirements could facilitate the design, development and use of secure, lawful and trustworthy AI systems. This leads to the following problem statement:

> *How can the three components (Lawful, Ethical and Robust) of Trustworthy AI, presented in the AI HLEG Guidelines for Trustworthy AI, be realised?*

with the following sub-question(s):

- Which methods, techniques and/or tools can be used to implement the requirements of the three components?
- Can these implementations be operationalised in a manner which considers the requirements throughout an AI systems life cycle?

1.4.1 Expected Outcome

The expected outcome of this thesis is a series of implementation suggestions for the different requirements analysed as per the problem formulation. One or more implementation suggestions will be presented for each of the analysed requirements. Finally, these will be compiled and presented in a visual figure.

Additionally, it is expected that an operational model for implementing the requirements throughout the AI systems' life cycle will be produced. This will be in the form of a diagram of flowchart.

1.5 Methodology

As can be seen from the problem formulation, the type of problem posed is an analysis problem with no specific pre-determined solution.

The primary methodology used in the creation of this thesis is desktop research. Specifically, the problem formulation was used as a baseline on which the desktop research was based.

The requirements of the AI HLEG as well as the AI Act, have been compiled into lists (will be presented in section 3.3, 3.4 and 3.5) for which implementations suggestions have been produced. These are based on desktop research into literature and the aforementioned sources.

Another way to approach the problem would have been to conduct a qualitative analysis of the compiled requirements, using interviews with experts in the field of AI and ML. This would have provided expert insight into the issue and might have revealed more implementation suggestions as well as tensions between these.

Similarly, interviews with domain experts could have been utilized to evaluate the results of the implementation suggestions derived from the desktop research as well as the operational model mentioned in section 1.4.1.

Chapter 2

State of the Art

This chapter will familiarize the reader with the topics and concept of Trustworthy AI, the AI Act proposal as well as a brief introduction to the field of XAI.

2.1 Trustworthy AI

This section covers the concepts of Trustworthy AI presented in the Danish Strategy for AI [8] and the Ethics Guidelines for Trustworthy AI from the AI HLEG [9].

2.1.1 Danish National Strategy for AI

The Danish national strategy for AI outlines five central initiatives [8]:

1. Principles for responsible development and use of artificial intelligence.
2. Common Danish language resource.
3. More open public-sector data for artificial intelligence.
4. Signature projects in the public sector.
5. Stronger investment in Danish businesses.

As part of the principles for responsible development and use of AI, the strategy presents a set of ethical principles for achieving Trustworthy AI, which will be presented in section 2.1.3. Initiatives 2-5 are not directly relevant for this thesis and are kept for completeness.

2.1.2 EU High-level Expert Group on AI (AI HLEG)

The Ethics Guidelines for Trustworthy AI created by the AI HLEG provides a framework for Trustworthy AI by defining a set of guidelines and requirements for AI systems. It served as one the primary ethical foundations used in the creation of the AI Act proposal. The guidelines state that Trustworthy AI has three key components which should be incorporated into the AI system's life cycle. Specifically, a trustworthy AI system should be [9]:

1. **Lawful**, meaning it complies with all relevant laws and regulations.
2. **Ethical**, meaning it should adhere to ethical principles and values.
3. **Robust**, meaning it should be safe from a technical and a social perspective.

It is important to note that the AI HLEG group does not provide an ordering of these components. Instead they should all be considered equally [9].

2.1.3 Ethical Principles for AI

The aforementioned sources presents the following ethical principles for AI [8], [9]:

EU High-level Expert Group on AI	**Danish National Strategy for AI**
Respect for human autonomy	Self-determination
Prevention of harm	Dignity
Fairness	Responsibility
Explicability	Explainability
	Equality and justice
	Development

Table 2.1: Comparison of ethical principles for AI.

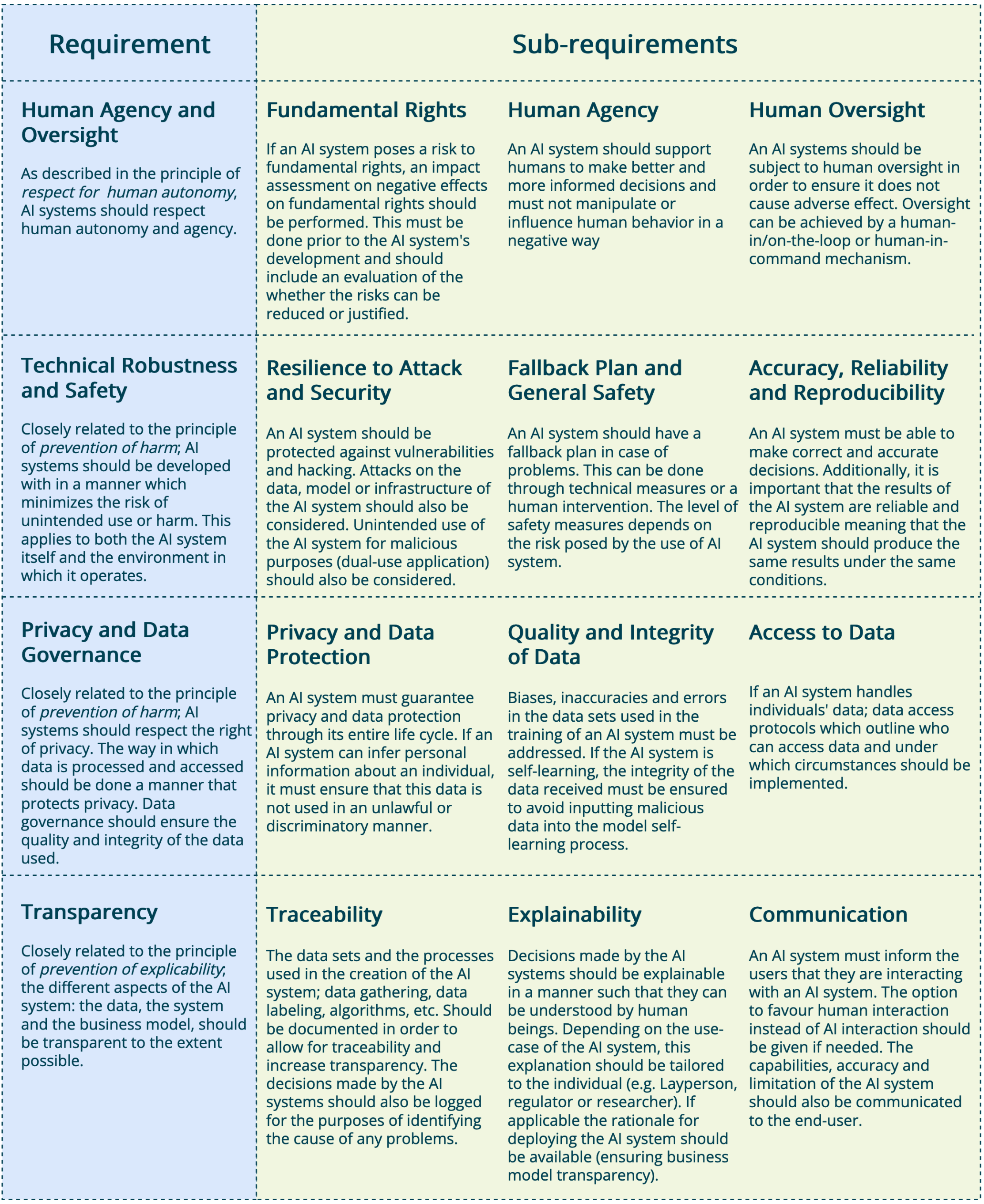
With the following elaborations [8], [9]:

1. (Human Autonomy) AI systems should not have deceive, manipulate or coerce humans. Humans interacting with AI systems must be able to have full self-determination over themselves.
2. (Prevent Harm) AI systems should not cause harm to humans, society, systems or technology.
3. (Fairness) AI systems must be fair and must ensure that individuals and groups are free from unfair bias and discrimination.
4. (Explicability) AI systems should be transparent and the capabilities and purpose of the system should be openly communicated. Additionally, the decisions of an AI system should be explained, to the extend possible, to those affected.

1. (Self-determination) The autonomy of people should have priority in the development and use of AI systems. The use of AI should not remove an individuals self-determination.
2. (Dignity) AI systems should respect human dignity, and should not be used to cause harm or used to infringe on fundamental rights or the democratic process.
3. (Responsibility) All levels related to AI systems should be responsible for the use of the system (i.e. Developers, businesses, users, authorities, etc.)
4. (Explainability) AI systems should be able to explain the decisions it produces.
5. (Equality and Justice) AI systems should not be unfairly biased or discriminate or show prejudice against specific groups.
6. (Development) AI systems should be ethically developed and used for the better progress of society.

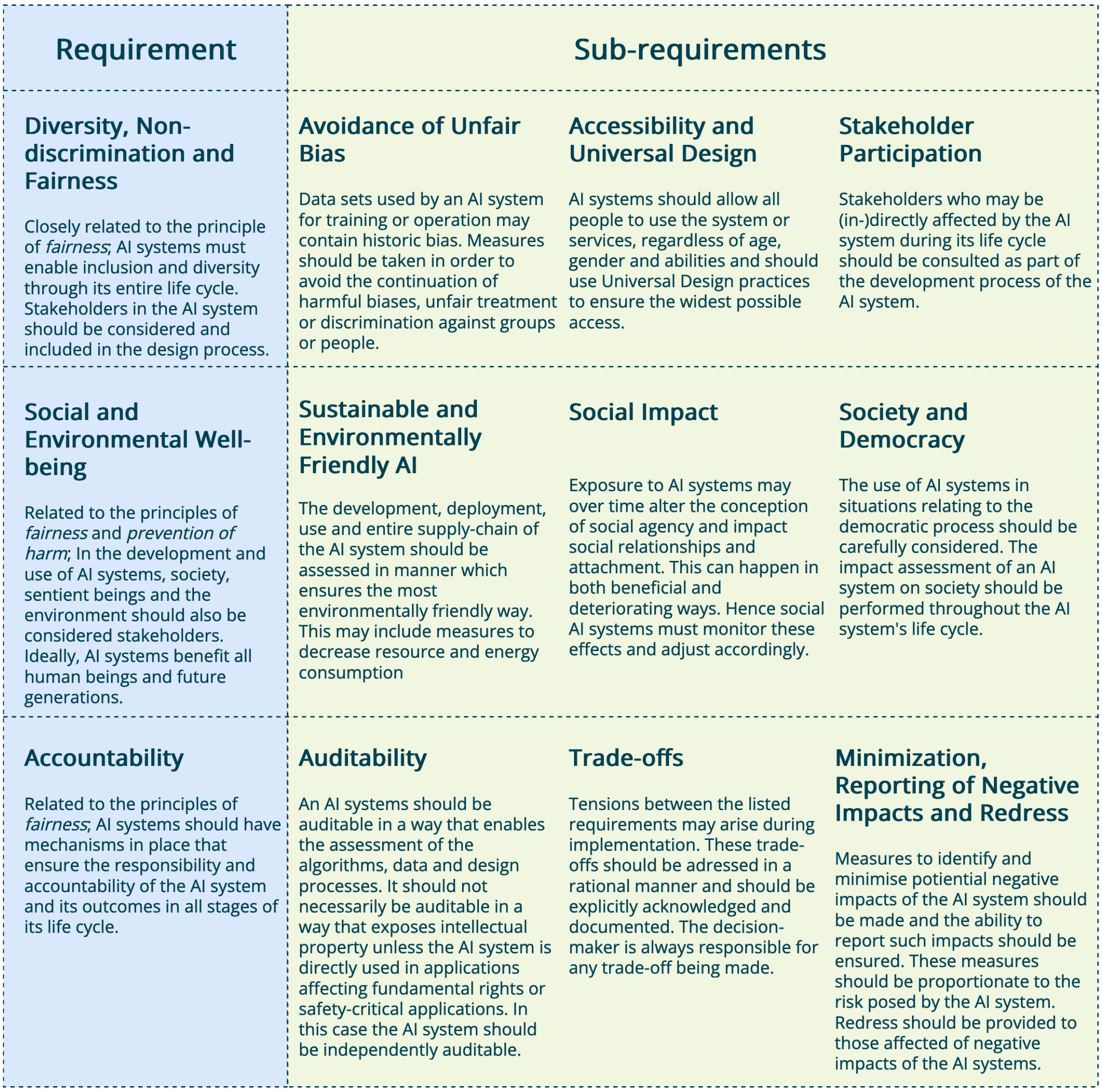
2.1.4 Requirements for Trustworthy AI

Building on the ethical principles described in section 2.1.3, the AI HLEG describes 7 key requirements which should be implemented in order to achieve Trustworthy AI [9].



Requirement	Sub-requirements		
Human Agency and Oversight As described in the principle of *respect for human autonomy*, AI systems should respect human autonomy and agency.	**Fundamental Rights** If an AI system poses a risk to fundamental rights, an impact assessment on negative effects on fundamental rights should be performed. This must be done prior to the AI system's development and should include an evaluation of the whether the risks can be reduced or justified.	**Human Agency** An AI system should support humans to make better and more informed decisions and must not manipulate or influence human behavior in a negative way	**Human Oversight** An AI systems should be subject to human oversight in order to ensure it does not cause adverse effect. Oversight can be achieved by a human-in/on-the-loop or human-in-command mechanism.
Technical Robustness and Safety Closely related to the principle of *prevention of harm*; AI systems should be developed with in a manner which minimizes the risk of unintended use or harm. This applies to both the AI system itself and the environment in which it operates.	**Resilience to Attack and Security** An AI system should be protected against vulnerabilities and hacking. Attacks on the data, model or infrastructure of the AI system should also be considered. Unintended use of the AI system for malicious purposes (dual-use application) should also be considered.	**Fallback Plan and General Safety** An AI system should have a fallback plan in case of problems. This can be done through technical measures or a human intervention. The level of safety measures depends on the risk posed by the use of AI system.	**Accuracy, Reliability and Reproducibility** An AI system must be able to make correct and accurate decisions. Additionally, it is important that the results of the AI system are reliable and reproducible meaning that the AI system should produce the same results under the same conditions.
Privacy and Data Governance Closely related to the principle of *prevention of harm*; AI systems should respect the right of privacy. The way in which data is processed and accessed should be done a manner that protects privacy. Data governance should ensure the quality and integrity of the data used.	**Privacy and Data Protection** An AI system must guarantee privacy and data protection through its entire life cycle. If an AI system can infer personal information about an individual, it must ensure that this data is not used in an unlawful or discriminatory manner.	**Quality and Integrity of Data** Biases, inaccuracies and errors in the data sets used in the training of an AI system must be addressed. If the AI system is self-learning, the integrity of the data received must be ensured to avoid inputting malicious data into the model self-learning process.	**Access to Data** If an AI system handles individuals' data; data access protocols which outline who can access data and under which circumstances should be implemented.
Transparency Closely related to the principle of *prevention of explicability*; the different aspects of the AI system: the data, the system and the business model, should be transparent to the extent possible.	**Traceability** The data sets and the processes used in the creation of the AI system; data gathering, data labeling, algorithms, etc. Should be documented in order to allow for traceability and increase transparency. The decisions made by the AI systems should also be logged for the purposes of identifying the cause of any problems.	**Explainability** Decisions made by the AI systems should be explainable in a manner such that they can be understood by human beings. Depending on the use-case of the AI system, this explanation should be tailored to the individual (e.g. Layperson, regulator or researcher). If applicable the rationale for deploying the AI system should be available (ensuring business model transparency).	**Communication** An AI system must inform the users that they are interacting with an AI system. The option to favour human interaction instead of AI interaction should be given if needed. The capabilities, accuracy and limitation of the AI system should also be communicated to the end-user.

Figure 2.1: Requirements for Trustworthy AI - Part 1.



Requirement	Sub-requirements		
Diversity, Non-discrimination and Fairness Closely related to the principle of *fairness*; AI systems must enable inclusion and diversity through its entire life cycle. Stakeholders in the AI system should be considered and included in the design process.	**Avoidance of Unfair Bias** Data sets used by an AI system for training or operation may contain historic bias. Measures should be taken in order to avoid the continuation of harmful biases, unfair treatment or discrimination against groups or people.	**Accessibility and Universal Design** AI systems should allow all people to use the system or services, regardless of age, gender and abilities and should use Universal Design practices to ensure the widest possible access.	**Stakeholder Participation** Stakeholders who may be (in-)directly affected by the AI system during its life cycle should be consulted as part of the development process of the AI system.
Social and Environmental Well-being Related to the principles of *fairness* and *prevention of harm*; In the development and use of AI systems, society, sentient beings and the environment should also be considered stakeholders. Ideally, AI systems benefit all human beings and future generations.	**Sustainable and Environmentally Friendly AI** The development, deployment, use and entire supply-chain of the AI system should be assessed in manner which ensures the most environmentally friendly way. This may include measures to decrease resource and energy consumption	**Social Impact** Exposure to AI systems may over time alter the conception of social agency and impact social relationships and attachment. This can happen in both beneficial and deteriorating ways. Hence social AI systems must monitor these effects and adjust accordingly.	**Society and Democracy** The use of AI systems in situations relating to the democratic process should be carefully considered. The impact assessment of an AI system on society should be performed throughout the AI system's life cycle.
Accountability Related to the principles of *fairness*; AI systems should have mechanisms in place that ensure the responsibility and accountability of the AI system and its outcomes in all stages of its life cycle.	**Auditability** An AI systems should be auditable in a way that enables the assessment of the algorithms, data and design processes. It should not necessarily be auditable in a way that exposes intellectual property unless the AI system is directly used in applications affecting fundamental rights or safety-critical applications. In this case the AI system should be independently auditable.	**Trade-offs** Tensions between the listed requirements may arise during implementation. These trade-offs should be adressed in a rational manner and should be explicitly acknowledged and documented. The decision-maker is always responsible for any trade-off being made.	**Minimization, Reporting of Negative Impacts and Redress** Measures to identify and minimise potiential negative impacts of the AI system should be made and the ability to report such impacts should be ensured. These measures should be proportionate to the risk posed by the AI system. Redress should be provided to those affected of negative impacts of the AI systems.

Figure 2.2: Requirements for Trustworthy AI - Part 2.

2.2 The European Artificial Intelligence Act (AI Act)

The EU AI Act proposal provides a legislative framework for the development and use of AI in society. This section covers several of the important aspects and legislation presented in the AI Act.

2.2.1 The AI Act Structure

2.2.2 A Risk-based Approach

The AI Act takes a risk-based approach for legislation on AI systems. Specifically, it places AI systems into one of four groups based on the risk the system poses to fundamental rights and safety. Figure 2.3 shows the risk-based categorization of AI systems according to the AI Act [10].

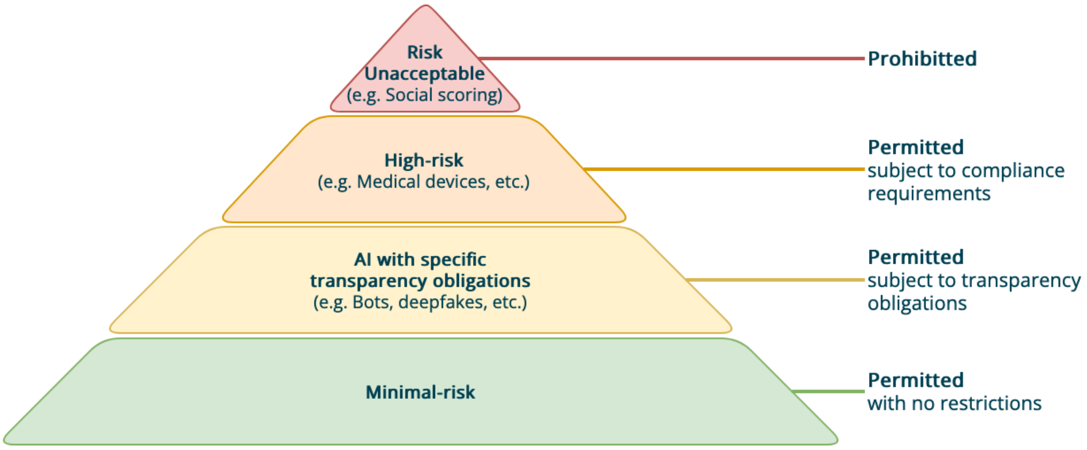

Risk
Unacceptable
(e.g. Social scoring)
Prohibitted
High-risk
(e.g. Medical devices, etc.)
Permitted
subject to compliance requirements
AI with specific transparency obligations
(e.g. Bots, deepfakes, etc.)
Permitted
subject to transparency obligations
Minimal-risk
Permitted
with no restrictions

Figure 2.3: Risk-based differentiation of AI systems. – Figure based on [12]

It should be noted that, depending on the specific use-case, an AI system may fall into both the high-risk and transparency obligations categories. Thus they are not mutually-exclusive.

Prohibited AI Practices

Title II of the AI Act concerns prohibited AI practises under which Article 5 specifies the following prohibited practises [10]:

1. (*Subliminal Manipulation*) AI systems which deploys subliminal techniques to distort a person's behaviour in a manner that causes that person or another person physical or psychological harm.

2. (*Exploitation*) AI systems which exploits vulnerabilities of a specific group of persons due to their age, physical or mental disability to distort their behaviour in a manner that causes that person or another person physical or psychological harm.
3. (*Social Scoring*) AI systems which are used by public authorities or on their behalf to evaluate or classify the trustworthiness of natural persons based on their social behaviour or known or predicted personal or personality characteristics, where this scoring leads to either or both of the following:
   (a) Detrimental or unfavorable treatment of natural persons or groups thereof in social contexts.

   (b) Detrimental or unfavorable treatment of natural persons or groups thereof that is unjustified or disproportionate to their social behaviour.
4. (*Real-time RBI*) The use of real-time remote biometric identification (RBI) systems in publicly accessible spaces for the purposes of law enforcement unless such use is strictly necessary for one of the following objectives:
   (a) Targeted search for potential victims of crime.

   (b) Prevention of substantial or imminent threat to life or safety of natural persons or of a terrorist attack.

   (c) Detection, localisation, identification or prosecution of a perpetrator or suspect of a criminal offence which is punishable in the EU Member State by a custodial sentence of a maximum period of at least three years, as determined by the law of that Member State.

High-risk AI Systems

Title III of the AI Act describes the classification procedure and requirements for high-risk AI system. The classification of high-risk AI systems is based on the intended use of the system as well as the environment in which it operates. The classification rules are listed in section 2.2.3). The requirements for high-risk AI system providers are listed in section 2.2.4.

Transparency Obligations for Certain AI Systems

Title IV of the AI Act describes AI systems which are subject to transparency obligations. These are described in the following [10]:

1. (*Bots*) AI systems which are designed to interact with natural persons must state, unless obvious under the circumstances of use, that the person is interacting with an AI system.

2. (*Emotional Recognition*) AI systems, and users of AI systems, which utilizes an emotional recognition system or biometric classification system, must inform natural persons on which these are used.
3. (*Deep Fakes*) AI systems, and users of AI systems, which generates or manipulates image, audio or video content to resemble persons, objects, places, etc. To appear authentic or truthful, must disclose that the content is artificially generated.

However, these obligations do not apply if the use of such a system is permitted by law to aid in law enforcement. Additionally, as stated in section 2.2.2, these systems may also fall under the high-risk category and are thus subject to, in addition to the transparency obligations, the same requirements as high-risk AI systems.

Low-/Minimal-risk AI Systems

AI systems which does not fall into any of the other categories are classified as minimal-risk and are not subject to any additional legislative requirements provided in the AI Act. These systems, however, are still recommended to adhere to the principles of trustworthy AI, such as the ones described in section 2.1, as well as other relevant guidelines. Additionally, Title IX of the AI Act encourages the drawing up of codes of conduct for voluntary use in minimal-risk AI systems.

2.2.3 Classification of High-risk AI Systems

Title III, Article 6 of the AI Act concerns the classification rules for high-risk AI systems. Specifically, it states that an AI system is considered high-risk if both of the following conditions are met [10]:

1. *the AI system is intended to be used as a safety component of a product, or is itself a product, covered by the Union harmonisation legislation listed in Annex II.*
2. *- the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service of that product pursuant to the Union harmonisation legislation listed in Annex II.*

or the AI system is part of any of the following areas [11]:

1. Biometric identification and categorisation of natural persons.
2. Management and operation of critical infrastructure.
3. Education and vocational training.
4. Employment, workers management and access to self-employment.

5. Access to and enjoyment of essential private services and public services and benefits.
6. Law enforcement.
7. Migration, asylum and border control management.
8. Administration of justice and democratic processes.

Annex II contains a list of Union harmonisation legislation which, among many others; regulations on toy safety, aviation and medical devices. The complete list in the AI Act annexes [11].

2.2.4 Requirements for High-risk AI Systems

The following requirements for high-risk AI systems are presented in the AI Act proposal [10]:

- Risk Management (Article 9)
- Data Governance (Article 10)
- Technical Documentation (Article 11)
- Record-keeping (Article 12)
- Transparency and provision of information to users (Article 13)
- Human Oversight (Article 14)
- Accuracy, Robustness and Cybersecurity (Article 15)

Here shown with their respective article numbers as they appear in the AI Act. The specifics of these will be presented as part of the analysis of each requirement in section 3.5.

2.3 eXplainable AI (XAI)

As was briefly touched upon in the introduction, the field of XAI seeks to research and develop methods for explaining the predictions of ML models. Being able to answer *how* and *why* a ML model came to a particular prediction has numerous advantages. Being able to properly answer these questions:

- Enables **model developers** to **better debug** and **understand** the model and its internals.
- Enables **auditors** to more reliably **audit the ML model specifics** and **ensure compliance**.
- **Facilitates trust** for the end user by **providing an explanation** alongside the prediction.
- **Uncovers** potential **biases** and/or **weaknesses** in the model.
- **Facilitates informed decision making** of end users by connecting explanations with **end user intuition** and **knowledge**.

2.3.1 Overview

Explainability for ML models can be achieved by two general methods:

1. Create or use a self-explaining, transparent machine learning model.
    - Here explainability is intrinsic to the model as the model itself provides the explanation.
2. Create or use a black box model and apply explanation methods afterwards.
    - Here explainability is provided *Post hoc*, meaning that the explanation is provided after model training.

Scope of Explainability

The *scope* of explainability refers to the level of explainability provided by the explanation method [13]:

- **Local** explanation methods provides an explanation for a model prediction on a specific input.
- **Global** explanation methods provides an explanation for how the model makes predictions, regardless of input.

Most current research has been focused on providing local explanations. In one of the latest surveys on XAI, 46 out of 50 papers surveyed fell into this category [14].

Model-specific vs. Model-agnostic Explainability

An explanation method can either be *model-specific* or *model-agnostic* [13]:

- **Model-specific** methods are constrained to explaining only a specific model or class of models.
- **Model-agnostic** methods can be used on any machine learning model after it has been trained.

It is important to note that model-specific methods are not necessarily created on a per-model basis but can instead be generalized to work on a specific model type or class of model. For example; a method which only works on all neural networks is considered model-specific by definition. Intrinsically explainable models such as *decision trees* are also model-specific, as the model itself provides the explanation.

Taxonomy

Figure 2.4 illustrates the taxonomy of explainability for ML models.

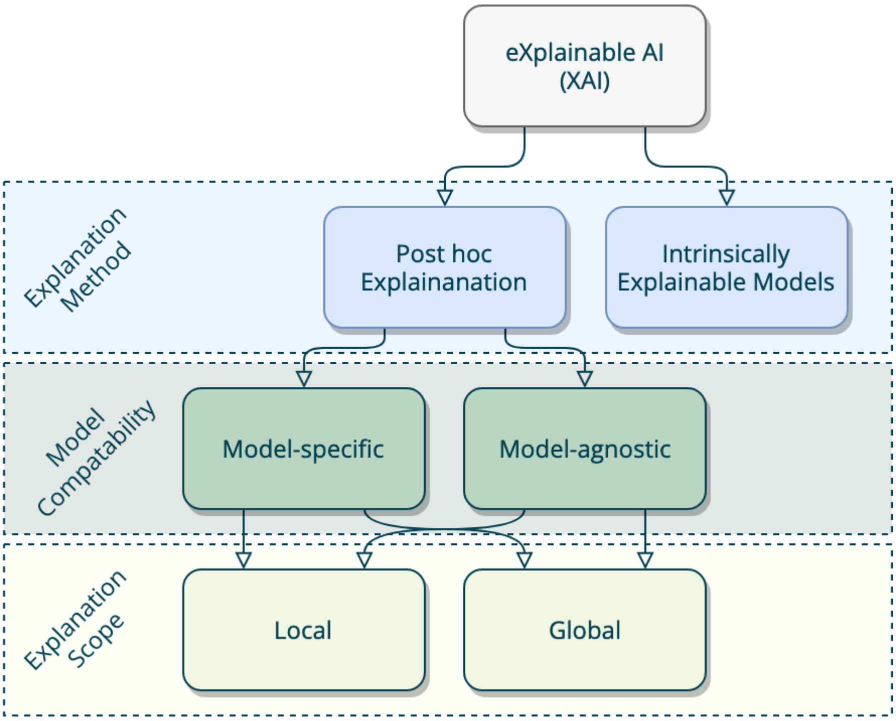


Figure 2.4: XAI Taxonomy

Self-Explaining Models

There exists several self-explaining or intrinsically explainable ML models and algorithms. Some of these include; *linear regression*, *k-nearest neighbors*, *decision rules* and *decision trees* [13]. These models are transparent by design, meaning they can be examined, audited and explained without any additional explanation methods being required.

A particularly intuitive explainable ML model is the decisions tree. This type of model presents all model decision boundaries as splits in nodes of the tree leading to new leaf nodes. The leaf nodes can contain either another split or a final prediction. Figure 2.5 illustrates a decision tree.

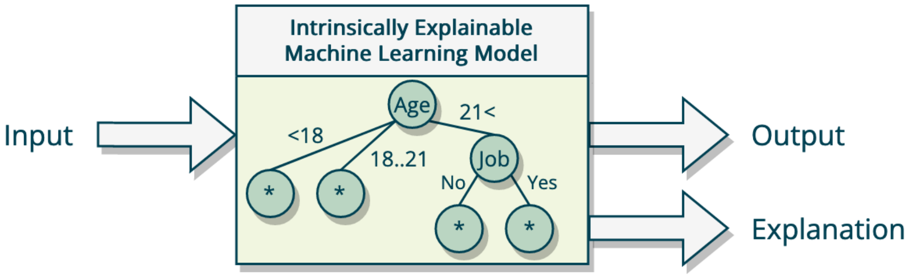


Figure 2.5: Intrinsically explainable model here illustrated using a decision tree.

By tracing a prediction from the leaf node to the root an explanation for a particular prediction can be obtained by the observing splits performed. This allows for explanations of outcome (i.e. Local explainability). Similarly, as the entire model can be examined and visualized as a tree structure, the model in its entirety can be explained (i.e. Global explainability).

Post hoc Explanation Methods

Post hoc explanation methods are applied after training of the ML model and creates explanations based on observed model behaviour. While several classes of post hoc explanation methods exists, one of the more intuitive classes use some form of *input perturbation* mechanism. Figure 2.6 illustrates a holistic view of this class of methods.

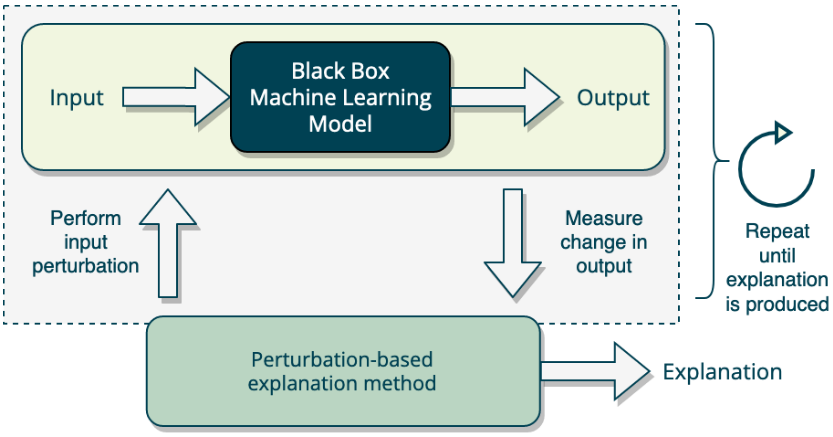

Input
Black Box Machine Learning Model
Output
Perform input perturbation
Measure change in output
Repeat until explanation is produced
Perturbation-based explanation method
Explanation

Figure 2.6: Post hoc explanation using input perturbation.

These methods perform perturbations (i.e. small changes) on the input data and measure the corresponding effect on the output in order to measure which features in the input data affects the output. Generally, the perturbation process is repeated a set amount of times (e.g. n samples) using different perturbations in order to better evaluate the effect on the output.

The SHAP method presented in section 2.3.2 is an example of a perturbation-based post hoc explanation method.

2.3.2 SHapley Additive exPlanations (SHAP)

SHAP is a local model-agnostic explanation method which computes *Shapley values*, a method from *coalition game theory*, and uses these values to explain the prediction of an input.

In coalition game theory multiple players collaborate to win a game and receive a payout. A (+/-) change in payout (called gain) is subject to player cooperation. Shapley values is a method for calculating how to fairly distribute the payout among the players based on their contribution to the total payout.

In the context of machine learning the terminology of coalition game theory translates as follows [13]:

- The **game** is the prediction task for a single instance (i.e. Input).
- The **gain** is the prediction for a single game minus the average prediction for all instances.
- The **players** are the feature values of the instance that collaborate to receive the gain (i.e. Predict a certain value or output).

The SHAP paper represents the Shapley value explanation as an *additive feature attribution method*. The formal definition of the class of additive feature attribution methods is as follows:

Definition 1: Additive feature attribution methods have an *explanation model* that is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{2.1}$$

where g is the explanation model, z' is a vector where $z' \in \{0, 1\}^M$ and M is the number of simplified input features, and $\phi_i \in \mathbb{R}$ is the feature attribution of the feature i [15].

Summing the attribution of all features approximates the output $f(x)$ of the original prediction model for a given input x.

Definition 1.1: Explanation models are defined as any interpretable approximation of the original prediction model.

Let f be the original prediction model and g the explanation model. Explanation models use *simplified inputs* x' which map to the original input x through a mapping function $x = h_x(x')$ [15].

Definition 1.2: Local explanation methods try to ensure $g(z') \approx f(h_x(z')$ whenever $z' \approx x'$ [15].

The SHAP paper defines the following desirable properties for this class of methods:

Property 1: Local accuracy

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{2.2}$$

The explanation model $g(x')$ matches the original prediction model $f(x)$ when $x = h_x(x')$. Additionally, $\phi_0 = f(h_x(0))$ represents the model output with all simplified inputs turned off/missing [15]. This property ensures that the accuracy of an explanation matches the original model locally for any input.

Property 2: Missingness

$$x'_i = 0 \Rightarrow \phi_i = 0 \tag{2.3}$$

The missingness property constrains features $x'_i = 0$ to have no attribution impact [15]. This means that if a feature $x_i = 0$, the feature will not appear in the provided explanation.

Property 3: Consistency Let $f_x(z') = f(h_x(z'))$ and $z'\ i$ denote setting $z'_i = 0$. Then, for any two models f and f', if;

$$f'_x(z') - f'_x(z'\ i) \geq f_x(z') - f_x(z'\ i) \tag{2.4}$$

for all inputs $z' \in \{0, 1\}^M$, then;

$$\phi(f', x) \geq \phi(f, x) \tag{2.5}$$

[15].

This property says that if a model changes such that the contribution of a feature value changes or stays the same, the Shapley value ϕ changes or stays the same [13].

The presented definitions and properties can be summarized in a trivial intuition. SHAP looks at each feature individually and combined in sets with other features, then looks at all of the combined feature contributions using the principles of Shapley values. Once computed, the values are used to tell how much each feature contributed to the particular prediction. Being a local explanation method means that each explanation needs to be done on a per-prediction basis and only is accurate for the particular prediction. The intuition of SHAP is visualised in figure 2.7.

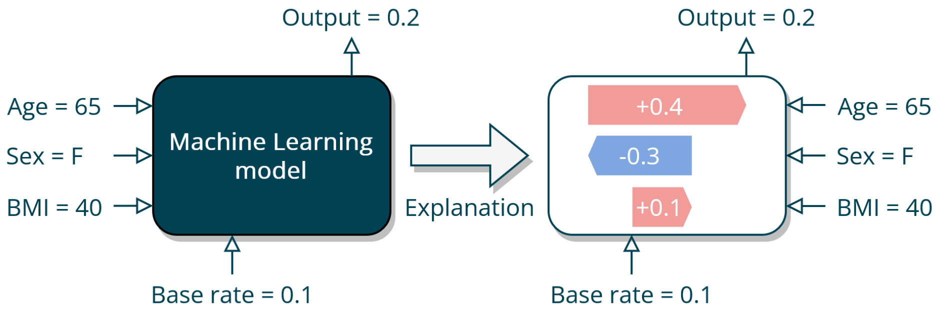


Figure 2.7: SHAP overview.

Chapter 3

Analysis

This chapter covers an analysis of guidelines and requirements presented in the framework for Trustworthy AI provided by the AI HLEG and the proposed AI Act respectively.

The requirements have been analysed and desktop research have been performed to find suggestions for implementation of the individual requirements.

3.1 Realising Trustworthy AI in the Context of the AI Act

For the purposes of simplifying the requirements into more manageable units, the sub-requirements presented in 2.1.4 will be split into two categories; ethical- and robustness sub-requirements. This enables us to analyze and discuss implementation suggestions for each sub requirement separately and place each sub-requirement under their respective section; namely Ethical- (section 3.3) and Robust AI (section 3.4).

Looking at each requirement in isolation is ill-advised specifically by the AI HLEG [9] as all requirements overlap in some form. All requirements and the manner in which they overlap, as well as any tensions between them, should be considered throughout the life cycle of the AI system. The separation of the requirements performed in this chapter is performed only for the purposes of individual examination and analysis of the requirements and should, for the aforementioned reasons, not be taken as more than what is stated.

The AI HLEG definition of the components required for Trustworthy AI will be used as it provides a clear definition of what is required for realising Trustworthy AI. By these definitions, realising the three components; Lawful-, Ethical- and Robust AI, should result in a trustworthy AI system. Figure 3.1 illustrates this rationale.

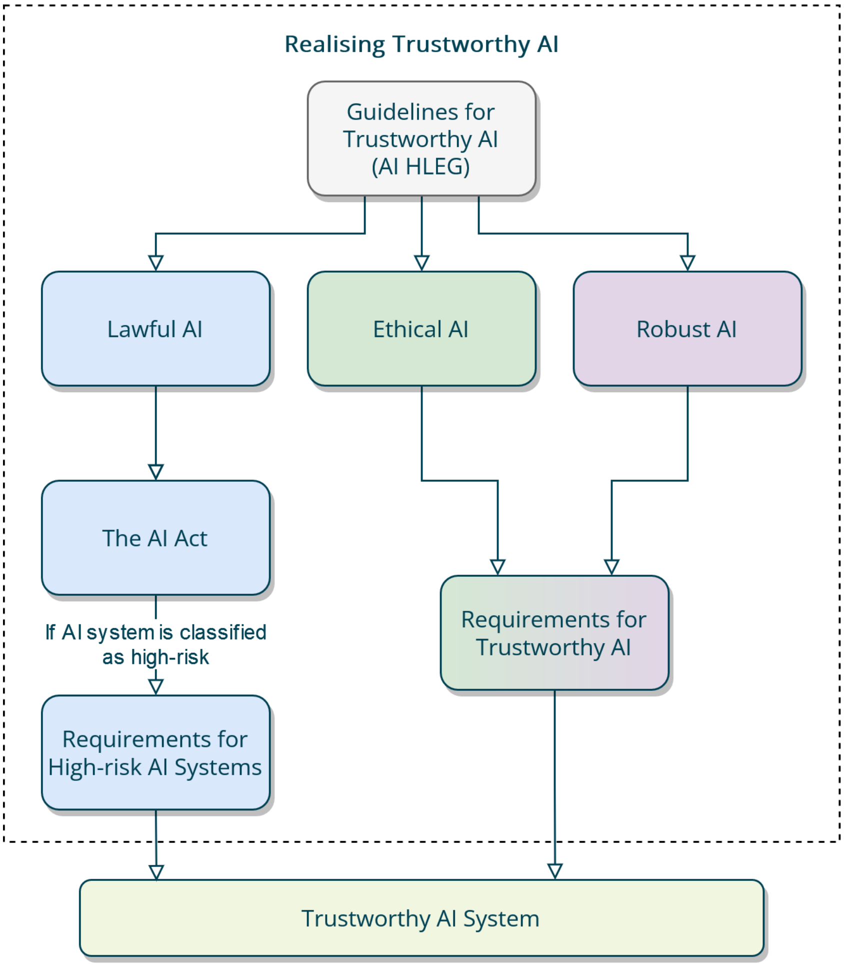

Realising Trustworthy AI
Guidelines for Trustworthy AI (AI HLEG)
Lawful AI
Ethical AI
Robust AI
The AI Act
If AI system is classified as high-risk
Requirements for High-risk AI Systems
Requirements for Trustworthy AI
Trustworthy AI System

Figure 3.1: Realising Trustworthy AI.

However, it is important to note that the guidelines and requirements presented by the AI HLEG are meant for all organisations striving to produce trustworthy AI systems, regardless of whether or not they are subject to legal requirements or regulation.

In contrasts, the AI Act requires (if adopted by the European Parliament) developers of AI systems, which are considered high-risk, to conform with the requirements specified in the proposal.

To summarize; an AI system which conforms to the AI Act requirements, but not the requirements for Trustworthy AI can not by itself be considered a trustworthy AI system. In contrast, an AI system which conforms to the requirements for Trustworthy AI, but is not considered high-risk, and thus not subject to the requirements of high-risk systems in the AI Act, can be considered trustworthy if it has no other legal obligations.

This is because conformity with the AI Act as well as other regulations produces a lawful AI system, which is not necessarily trustworthy. However, because of the first component of Trustworthy AI, a trustworthy AI system must be lawful. Figure 3.2 illustrates this concept.

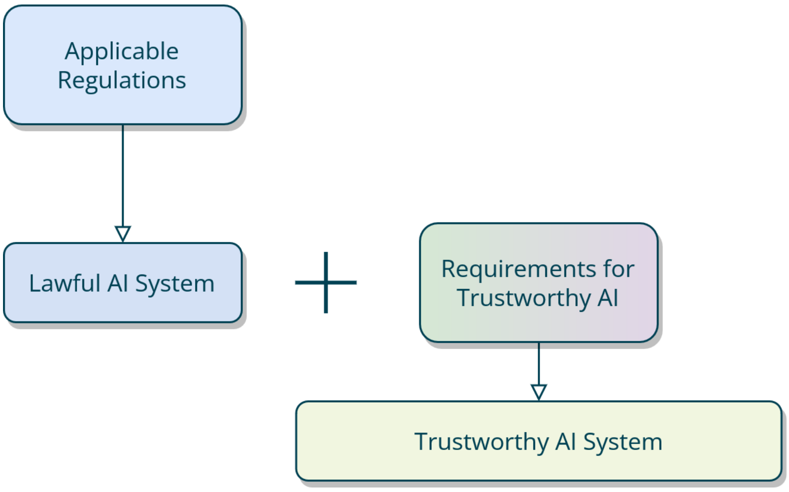


Figure 3.2: Caption

3.2 Lawful AI

As described in section 2.1.2 the first component of Trustworthy AI is that it should be lawful. This means that if conformity with the AI Act can be ensured for an AI system, it can be concluded that the AI system is lawful in a European context. Naturally, additional legislation and regulations may apply, most notably the GDPR [16], however, for the purpose of this thesis only the AI Act will be considered in the context of the Lawful AI component.

The analysis of the requirements of the AI Act as well as suggestions for the implementation of these will be presented in section 3.5

3.3 Ethical AI

As described in section 2.1.2 the second component of Trustworthy AI is that it should be ethical. This entails adherence to the ethical principles listed in section 2.1.3 as well as the requirements in section 2.1.4. The next section covers an analysis of the ethical principles followed by an analysis of the ethical sub-requirements in the section following.

3.3.1 Ethical Principles for Trustworthy AI

Both the ethical principles described by the AI HLEG and the Danish national strategy for AI, presented in section 2.1.3, are rooted in the fundamental rights of the EU Charter and are thus similar in their extend and definitions.

Mapping the ethical principles from the Danish national strategy for AI to those presented by the AI HLEG it is observed that they overlap (**AI HLEG**, *DK AI Strategy*):

1. **Respect for Human Autonomy**
    - *Self-determination* is by definition a part of human autonomy.
2. **Prevention of Harm**
    - *Dignity* is related to prevention of harm by stating that AI systems should not harm the dignity, fundamental rights of humans or harm the democratic process.
    - *Responsibility* is related to the principle of prevention of harm by dissuading actors from using AI to harm by holding stakeholders of an AI system responsible.
    - *Equality and Justice* is related to the principle of prevention of harm through the requirement for justice for the same reasons as the principle of responsibility.
3. **Fairness**
    - *Dignity* is related to the principle of fairness by stating that AI systems should respect human dignity and fair treatment of all.
    - *Equality and Justice* are directly related to the principle of fairness through the requirement of equality.
4. **Explicability**
    - *Explainability* is directly linked to the principle explicability as it states that a decision of an AI system should be explainable to those affected.

Since the ethical principles presented in the Danish strategy for AI is encompassed in those from the AI HLEG, stakeholders of an AI system need not to focus on adhering to these principles separately. Admittedly, both of the ethical principles are provided more as guidelines to foster an ethical mindset to the development and use of AI and should be viewed as such.

3.3.2 Ethical Sub-requirements for Trustworthy AI

The ethical sub-requirements have been compiled by analysing whether they are rooted in an ethical concept or principle. An example of this is *Stakeholder Participation*, which is considered an ethical sub-requirement due to it being rooted in the ethical principle of *fairness*. The ethical sub-requirements are presented in figure 3.3.

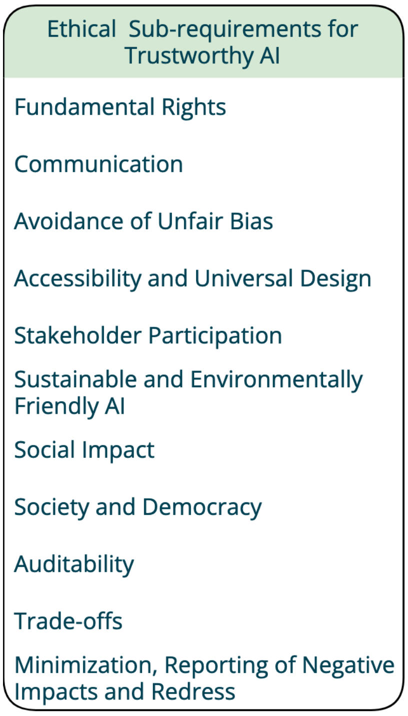

Ethical Sub-requirements for Trustworthy AI
Fundamental Rights
Communication
Avoidance of Unfair Bias
Accessibility and Universal Design
Stakeholder Participation
Sustainable and Environmentally Friendly AI
Social Impact
Society and Democracy
Auditability
Trade-offs
Minimization, Reporting of Negative Impacts and Redress

Figure 3.3: Ethical Sub-requirements.

Fundamental Rights

Ensuring that an AI system respects fundamental rights and adheres to this requirement is not a straightforward task. Several aspects needs to be taken into account such as the impact of the AI system in question, as well as the the manner in which the system is developed.

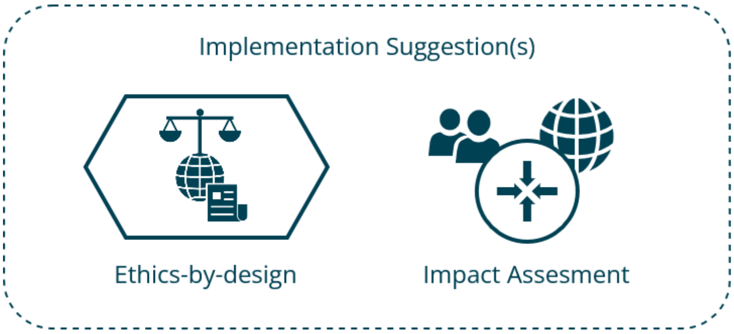
Naturally, similarly to the Lawful AI component, any AI system must respect and comply with any relevant legislation and regulations, such as the AI Act, GDPR [16], etc. This is important to consider, as these regulations are put in place to secure fundamental rights in first place.

It should be noted that this is considered through a European-centric context and legislation alone may not always respect fundamental rights in other regions or internationally. An example of this is the use of AI for *Social Scoring*, which is prohibited by the AI Act, however, liberally used in countries such as China.

The AI HLEG along with the EU Agency for Fundamental Rights (FRA) recommends, along with ensuring regulational compliance, that an impact assessment of the AI system on fundamental rights should be performed prior to or during the design and development of the AI system [9], [17].

If the results of the impact assessment show unacceptable negative impact on fundamental rights these should be mitigated through changes in the design of the system. If no acceptable mitigation can be achieved, the AI system should not continue development in its current form [9].

An approach to solve this issue is to implement an *ethics-by-design* or *rule-of-law-by-design* approach in which the fundamental rights are considered and incorporated during the design process.



Communication

This requirement closely relates to Title IV of the AI Act (see section 2.2.2) which sets out transparency obligations for specific AI systems. However, being compliant with Title IV is not necessarily enough to satisfy this requirement. This is evident as the sub-requirement does not state specific types of AI systems and instead is general in intend. This means that user communication should be considered in the development of all types of AI systems.

As stated in the sub-requirement, the capabilities, accuracy and limitations of the system should be communicated to the user. An approach to implement this is to create and maintain proper documentation designed for the users of the AI system. This, however, requires training of the users in order properly foster an understanding of the AI systems capabilities.

Ideally the AI system should provide this information directly to the user when necessary such as when a decision or prediction is provided by the system. The level and depth of communication is naturally dependant on the intended use and complexity of the system. Good user experience (UX) and interaction design practises could facilitate some of the user communication.

Care has to be taken in order to strike a good balance between training and ensuring users knowledge prior to using the system and communicating directly as needed through some form system interaction. This is important as a user may over-/underestimate the capabilities and limitations of the AI system if presented with the information only at a possibly time-critical moment, such as when a prediction is provided.

It should be noted that this considered through the context of the user being a professional working with the AI system. In the case of a layperson interacting with an AI system this sub-requirement should be considered in that context.

Additionally, as is mandated by the GDPR Article 22, paragraph 1 [16]:

> *The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*

the user (here considering citizens) of the AI system may request human interaction. This should be considered in conjunction with the *Human Agency + Human Oversight* sub-requirement presented in section 3.4. While the GDPR is a part of the scope of Lawful AI, it should nevertheless be considered in conjunction with this and other requirements.

Indeed, these solutions are solely reliant on the intended purpose of the AI system. Not all AI systems may need or require the same level of communication, if any.

Implementation Suggestion(s)
GDPR
Documentation

Avoidance of Unfair Bias

Bias and the avoidance thereof is closely related to the principle of fairness. Often bias and fairness are used to describe the same concept, namely equal and fair treatment and procedures. Bias is a natural product of human cognition and social interaction and may manifest itself in positive, neutral and negative ways. The question of whether or not it is possible to completely eliminate bias is philosophical in nature, however, measures to reduce negative and unfair bias in the context of AI systems can be taken.

A recent survey [18] examines bias and fairness in ML, and demonstrates that bias can take many forms depending on where in the ML pipeline it is introduced. The company *FiddlerAI*, which specialises in creating responsible AI solutions, has created a figure to contextualise these different types of bias [19]. These are presented in figure 3.4.

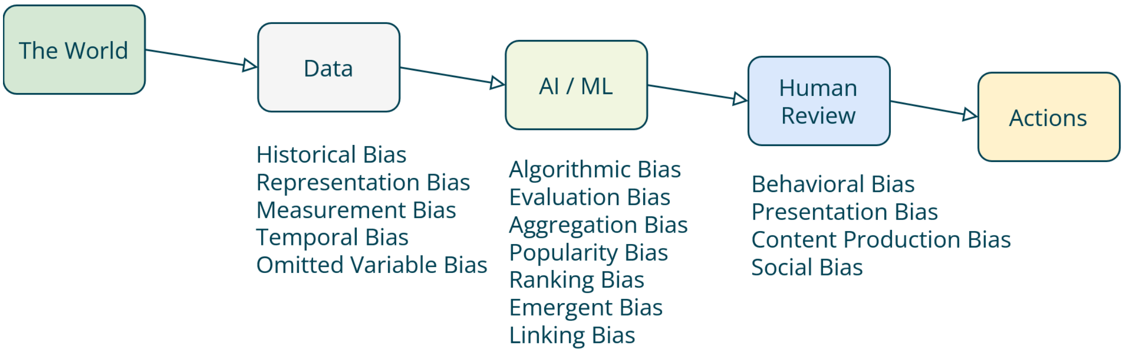


Figure 3.4: Different types of bias in the context of AI. – Figure based on [19].

The authors of the survey highlights some of the challenges to bias and fairness research. One of the main challenges presented is that a general definition of what constitutes fairness has yet to be synthesized [18]. This is problematic as fairness as a means to mitigate bias may have different meanings in different settings and for different people. To accommodate this issue, the authors provide several definitions and metrics for fairness as well as methods for implementing these.

The bottom line of bias management is that the measures needed to eliminate or reduce bias depends on the type of bias in question. To address some of the different types of bias, IBM has developed the *AI Fairness 360* (AIF360) toolkit [20], [21], which enables developers to asses ML models with regards to bias and fairness. Several of the algorithms and methods for mitigating bias presented in the survey [18] have been incorporated into the this toolkit [21].

Managing bias is a challenging issue, with no current methodologies or tools addressing or mitigating all types of bias. Awareness of bias and the different forms it can take should be considered an important factor. This is notably the case, as awareness and diversity can foster a mindset and provide perspective which can ultimately reduce bias throughout the AI system by simply observing its presence [9].

Implementation Suggestion(s)
Bias Management
Diversity

Accessibility and Universal Design

Making an AI system accessible to as wide a range of persons and groups directly relates to the principle of fairness. Depending on the intended use of the AI system, care should be taken to include a user-centric approach the interaction between human and AI system. This is particularly important if users are expected to interact with the AI system without any prior training.

Universal Design is a design framework which aims to facilitate the design of inclusive, accessible and universally usable products and environments. It is defined as follows [22], [23]:

> *Universal design is design that's usable by all people, to the greatest extent possible, without the need for adaptation or specialized design.*

The framework is based on the principles of *accessible*, *inclusive* and *usable* and includes guidelines for achieving these through the design process [22].

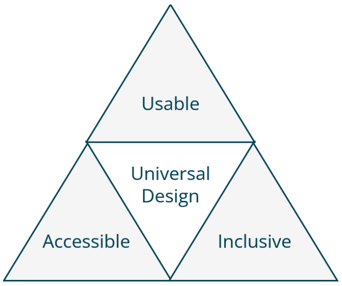
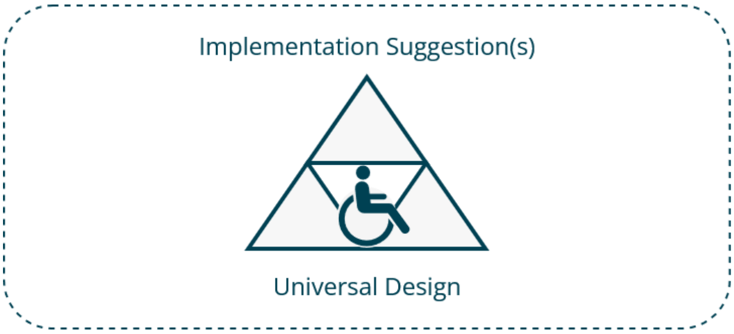


Figure 3.5: Principles of Universal Design.

As stated earlier, the extent to which Universal Design and accessibility measures needs to be implemented should be based on how the AI system is intended to be used, interacted with and by whom. If the users of the AI system is expected to have completed training prior to use, such as may be implemented in the communication sub-requirement described earlier, developers may lenient in their approach to accessibility. This is because the user is expected to be trained in its operation and accessibility measures may thus be superfluous.

An example of this is to consider a medical AI system which predicts the probability of certain lifestyle diseases and thus requires the user to be a medical professional. This means that the developers may expect a certain level of proficiency by the AI systems intended users and thus provide specialized training.



Stakeholder Participation

This sub-requirement promotes inclusivity and participation, and is rather self-evident in its intent. A natural implementation of this sub-requirement is to facilitate stakeholder participation and create social dialogue. This can be done through arranged stakeholder meetings or through the use of an analytics company both during the design phase, deployment and during market operation of the AI system.

Sustainable and Environmentally Friendly AI

This sub-requirement can be approached in two different ways depending on whether the AI system is purely software-based or constructed into or utilising specialised hardware.

If purely software-based the AI system environmental impact should be considered in terms of energy usage in particular. Additionally, if the AI system requires the utilisation of cloud computing resources or purchase of computer hardware, this part of the supply-chain needs to be assessed as well.

If the AI system is part of or utilises specialised hardware, care should be taken in order to make the integration, production and energy usage as efficient as possible and with as little of an environmental impact throughout the entire supply-chain. This is particularly important, as the production of the specific hardware may, depending on the hardware, energy usage and amount of raw materials required, put a strain on resources and the environment.

Implementation Suggestion(s)

Supply-chain Analysis

Social Impact

Relating closely to the requirement of fundamental rights and stakeholder participation, this sub-requirement can be addresses through an assessment of the AI system's social impact.

The AI HLEG notes that AI systems have the potential to change social agency and impact social relationships and attachments [9]. This change may bring benefits, but also the potential for psychological and societal consequences to humans.

An example of this is illustrated by the advent of social AI's which may be designed to simulate or mimic human company, conversation and emotions in order help socially inept and/or lonely individuals, which may result in user attachment to the AI. The consequences of human-machine attachment have yet to be documented, but should nevertheless be considered during the design and development of such a system.

As such, we see that this sub-requirement deals with both small, local social impacts as well as the potential impact of AI on humankind on a grand scale. Thus it is crucial to consider the true scope of social impact when performing an impact assessment.

Implementation Suggestion(s)

Impact Assesment

Society and Democracy

The use of AI systems may (un-)intentionally affect the democratic process and society in manners which are not always obvious to the end user.

An example of this is the data retention algorithms used, in particular, by services such as the *feeds* on websites/apps such as Facebook, Instagram, TikTok, etc. These algorithms are specifically designed to maximise user retention and are thus susceptible to promoting unhealthy content. This was the case both during the *Donald Trump 2016 presidential election*, in which the algorithms used by Facebook promoted content which would confirm users current biases and thus affect the democratic process by shielding people from other sources of content

For this reason, special considerations for the potential impact should be made for AI systems which are social in nature or may influence the democratic process or social dynamics.

This is related to the principle of fairness and the sub-requirement of avoidance of unfair bias. A technical solution to this is to insert a mechanism into the algorithm which partially distributes content to all users and groups, regardless of the specific group or user's interests or preferences. However, this implemented suggestion is specific to the avoidance of unfair bias sub-requirement and should be considered as part of for that requirement.

Implementation Suggestion(s)

Impact Assesment

Auditability

Auditability is a sub-requirement which borders on the verge between what lies within the scope of Lawful- and Ethical AI. The degree of auditability measures introduced into an AI system is dependant on the regulatory requirements for the particular AI system. However, the data used for model training as well as the design process should be documented in a manner which facilitates internal or external review.

Even if not specifically required by law, the organisation responsible for the design, development and deployment of the AI system could implement a governance framework for auditability in order to facilitate this process. Such a framework could be based on or cooperate with a data governance framework for ensuring auditability of the data used by an AI system. This provides auditability for the data used and how it was collected. The models capabilities and what it is optimised for should be documented in order to provide auditability for the design process and how the system operates.

A technical measure to facilitate auditability is the implementation of so called *audit trails* [24] which will be discussed under the *traceability* sub-requirement in section 3.4.

It is important to note that this is far from an exhaustive list of implementation suggestions as the specifics of such an implementation is dependant on the auditability requirements set out by applicable regulations.

Implementation Suggestion(s)
Documentation
Governance

Trade-offs

The AI HLEG notes that inevitable trade-offs may arise when implementing the presented requirements [9]. These should be addressed and any conflicts should be evaluated in regards to the risk they pose to the ethical principles presented in section 2.1.3.

Hence this sub-requirement should be implemented by performing a thorough evaluation of these trade-offs.

Implementation Suggestion(s)

Trade-off Evaluation

Minimization, Reporting of Negative Impacts and Redress

Many of the aforementioned requirements specifically seek to minimize negative impacts before they happen. Using impact assessments to evaluate potential risks may offset or mitigate many of the negative impacts an AI system during the design or development phase.

While this sub-requirement should be implicitly considered during the entire AI system life cycle, it is nevertheless recommended to explicitly consider minimizing the potential negative impacts of the AI system.

Minimizing negative impact naturally requires knowing that such an impact has happened. For this reason, it should be able for users, stakeholders, whisteblowers, etc. To report any such impacts or episodes.

In addition to being able to report negative impacts, measures to ensure redress (i.e. compensation) should be in place [9]. Knowing that compensation is attainable when negative impacts occur generates trust in the the AI system and the organisation behind it.

3.4 Robust AI

As described in section 2.1.2 the third component of Trustworthy AI is Robust AI. This section covers an analysis of the robustness-related sub-requirements.

The European Union Agency for Cybersecurity (ENISA) has in 2020 published a report describing the *Threat Landscape for AI* [25], which outlines several threats specific to AI throughout an AI systems life cycle. Following this report, ENISA published in 2021 a report named *Securing Machine Learning Algorithms* [26] detailing the different ways in which AI systems (more specifically the ML aspects of AI systems) can be secured with regards to the threats presented in the threat landscape report. These sources will be used throughout the analysis of the following sub-requirements.

3.4.1 Robustness Sub-requirements for Trustworthy AI

The robustness sub-requirements have been compiled by analysing whether they are directly influential on the safety and security of the system. En example of this is *Privacy and Data Protection* as this sub-requirement directly affects the security of the data used by the system and thus, if not properly implemented, may affect the safety of any personal data used. The robustness sub-requirements are presented in figure 3.6.

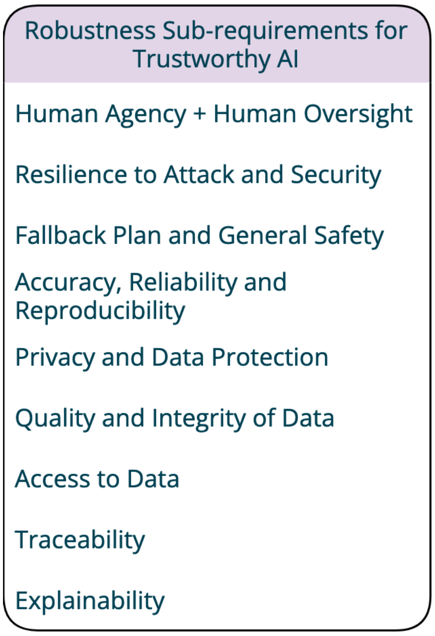

Robustness Sub-requirements for Trustworthy AI
Human Agency + Human Oversight
Resilience to Attack and Security
Fallback Plan and General Safety
Accuracy, Reliability and Reproducibility
Privacy and Data Protection
Quality and Integrity of Data
Access to Data
Traceability
Explainability

Figure 3.6: Robustness Sub-requirements.

Human Agency and Human Oversight

The implementation of this sub-requirement is two-fold:

1. Human Agency should ensure the autonomy and self-determination of users interacting with the AI system.
2. Human Oversight should ensure that the AI system is subject to human involvement in one or more places in the operation of the AI system.

Human Agency is closely linked to the communication sub-requirement described in section 3.3. This is the case as users need to be informed and instructed in the use of the AI system. Otherwise the users may lose autonomy in regards to being able to understand and challenge the results produced by the AI system.

The example of a medical professional using an AI system to evaluate the likelihood of some potentially life-threatening health issue occurring in a patient will be used here. If the medical professional is not properly trained or given the right knowledge or tools to comprehend how or why the AI system came to a particular conclusion, he/she might not be able to sufficiently challenge such a conclusion (i.e. Lose self-determination). This is a problem as the patient may receive improper treatment based on such a decision. This is an extreme example, but nevertheless an interesting one.

Similar to the communication sub-requirement, the GDPR enshrined right to not be subject to a purely automated decision [16] also applies here.

Implementing measures for human oversight can be done using one of the three following mechanisms:

1. *Human-in-the-loop* (HITL) mechanisms puts a human in a decision-making role in one or more places in the decision pipeline of the AI system.
2. *Human-on-the-loop* (HOTL) mechanisms puts a human in an overseeing role in one or more places on the decision pipeline of the AI system.
3. *Human-in-command* (HIC) mechanisms puts a human in the overseeing and decision making role of the AI system. (This mechanism is most relevant in the context of human-hardware interface systems such as vehicles, weapons systems, etc.)

No exact definition of all three mechanisms could be found in literature, hence these definitions are self-made and thus might be subject to interpretation. Creating precise definitions for these mechanisms prove rather difficult as they are, with the exception of HITL mechanisms, poorly described. Additionally, the HITL mechanism makes intuitive sense whereas the exact details of the other mechanisms are rather vague.

Figure 3.7 has been created to present an intuition of the different mechanisms using an autonomous car as example.

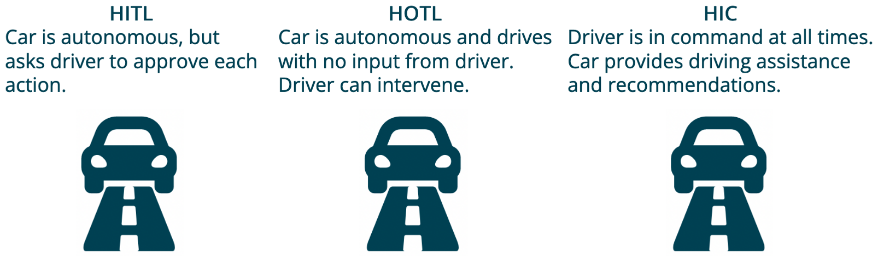

HITL
Car is autonomous, but asks driver to approve each action.
HOTL
Car is autonomous and drives with no input from driver. Driver can intervene.
HIC
Driver is in command at all times. Car provides driving assistance and recommendations.

Figure 3.7: HITL vs. HOTL vs. HIC.

It is very important to note both the benefits and disadvantages of these mechanisms. Having humans present as part of the AI system may increase both accuracy, reliability [27] and trustworthiness but also introduce more bias and decrease fairness [28]. These trade-offs should be evaluated (possibly as part of the *trade-offs* sub-requirement) when evaluating how such a mechanism should be put in place.

Resilience to Attack and Security

This sub-requirement is extensive by nature. This is because all security aspects of the AI system should be considered. This means that the security of the data, ML assets and model, as well as the underlying infrastructure, hardware and software needs to be considered.

Using figure 1.6 as a baseline, we see three types of attacks directly targeted at the ML aspects of the AI system. Namely; data poisoning, adversarial attacks and model leakage.

Here it is important to note that these attacks are not an exhaustive list of threats to AI systems. Many other types of attacks also exist [25], however, the three mentioned are among the most known and studied ML-specific threats.

Data Poisoning Data poisoning is the act of injecting malicious data into the training data. This may introduce a data backdoor, produce erroneous models and reduce system accuracy [29]. Hence, this type of attack may directly impact the *availability* and *integrity* of the AI system [25].

In many ways, a data poisoning attack can be considered the ML equivalent of an *injection* attack.

Data poisoning can be introduced in different ways. It may be introduced at a data supplier in which a malicious employee introduced it before shipping. If the AI system is self-learning based on user-input it may be deliberately introduced by malicious user. Additionally, it may be introduced by a malicious actor during data collection.

Given that data poisoning can be introduced in different ways, the security controls to mitigate the vulnerability varies as well. Table 3.1 presents a subset of the data poisoning specific- vulnerabilities and their respective controls as described by ENISA [26]:

	Vulnerability	**Security Control**
1	Use of uncontrolled data	Control all data used by the ML model
2	No detection of poisoned samples in the training dataset	Use methods to clean the training dataset from suspicious samples

Table 3.1: Data poisoning vulnerabilities and controls.

Relating closely to control 2 of table 3.1; a direct mitigation of data poisoning is data sanitization. This can be done using classic input sanitization, a rule-based approach for data evaluation or even another AI system acting as a filter, depending on the situation [29].

Additionally, another implementation of control 2 is to point out important data and use a HITL mechanism (as described in the previous section) for this specific data [26].

Adversarial Attacks Next are adversarial attacks (also known as adversarial examples). This type of attack targets the model inference and makes the model produce wrong predictions on an input. Hence, this type of attack affects the availability and integrity of AI system [25]. This works by having the input data include perturbations which are imperceptible to the human eye, but which has a great effect on the prediction [25]. Figure 3.8 shows an example of such an adversarial example.

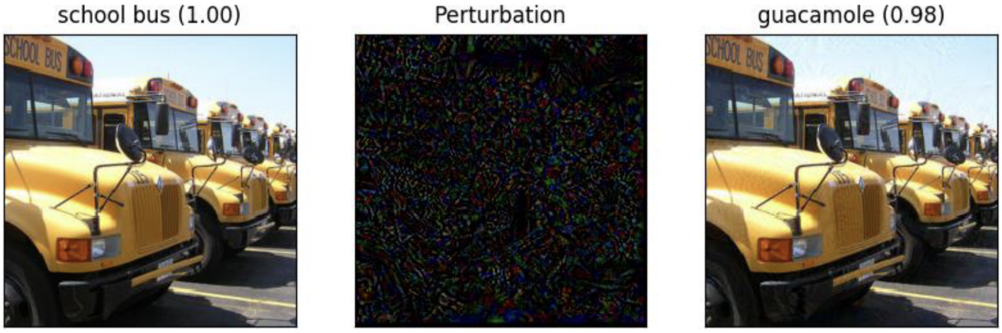


Figure 3.8: An adversarial example [29].

While adversarial attacks on image recognition are easy to visualize it should be noted that this type of attack is not limited to this task. Adversarial examples on other tasks such as object recognition, speech recognition, text summarizing, etc. have been demonstrated [29].

Any trained ML model consists, at its core, of an underlying mathematical model. Adversarial attacks exploit this fact by utilising, among other methods, gradients to find perturbations, which can manipulate the input data in a manner that pushes the data over decision boundaries in the ML model, thus making it generate a wrong prediction.

Given the use of image recognition in high-risk applications of computer visions such as in autonomous vehicles, as described in section 1, the consequences of adversarial attacks on such systems may be severe or even fatal. This concern was demonstrated by a group of researchers by applying a physical perturbation to a US stop sign, thus making the image recognition classifier classify the input as a 45Mph speed limit sign [30]. Figure 3.9 shows this perturbation.

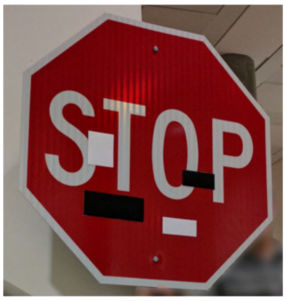


Figure 3.9: Physical perturbation applied to a US stop sign [30].

Methods to increase robustness against adversarial attacks are being researched and proposed, with some providing *certified defense* [31] using methods such as *differential privacy* [32] (More on differential privacy in the privacy and data protection subsection). Being a certified defense means that the method produces a *certificate* which proves that a ML model (classifier) is robust against adversarial attacks within some bound. It should be noted, however, as is stated by the authors of the aforementioned sources, that due to the nature of of how adversarial attacks work it is very hard to completely defend against this type of attack under all circumstances.

Nevertheless, ENISA has created several controls regarding adversarial attacks. Table 3.2 presents a subset of adversarial attack-specific vulnerabilities as well as their respective controls [26]:

	Vulnerability	Security Control
1	Lack of detection of abnormal inputs	Implement tools to detect if a data point is an adversarial example or not
2	Lack of training based on adversarial attacks	Add some adversarial examples to the training dataset
3	Inputs totally controlled by the attacker which allows for input-output-pairs	Apply modifications to inputs

Table 3.2: Adversarial attack vulnerabilities and controls.

ENISA recommends adding a *detector subnetwork* to the ML model to implement control 1 of table 3.2. A detector subnetwork is proposed method for uncovering adversarial examples [33]. This method uses a neural network classifier which is trained to distinguish genuine data from data containing adversarial perturbation. It should be noted that the authors of this method constructed an adversarial attack which was able to fool this detector subnetwork [33] and is thus not necessarily an adequate defence against an advanced adversary. However, given that the detector subnetwork can reliably distinguish adversarial examples which are not specifically designed to fool the detector, it may still be advantageous to include such a method.

A direct technical implementation of control 2 of table 3.2 is to perform *adversarial training*. In this process, adversarial examples are generated and correctly labelled before being merged with the original training data. This increases model robustness to adversarial attacks by performing model training directly on these examples. It should be noted that it is infeasible, due to the nature of these attacks, to address all adversarial examples that may be produced for the ML model. Hence, the implementation of this control might not necessarily reduce the risk enough to be considered mitigated. If adversarial training is implemented, it naturally follows that extensive testing must be performed in order to document the effects of this control.

Model Leakage Finally is the threat of model leakage. This threat refers to the unintentional leakage of model details in the model output data or through some other means. Hence, this is a threat to *confidentiality* of the underlying model of the AI system [25].

Model leakage can happen in different places. This is particularly true, as the model may not necessarily be deployed in the same network or infrastructure as it was trained such as if it is outsourced or deployed to a cloud environment. In this case, the model is entirely available to the hosting party and thus subject to their security level. The risks involved in such a deployment strategy should naturally be considered before model deployment. However, even if securely deployed, model details may still leak as a consequence of the model output itself.

In the case that an attacker has full control over inputs to the model, an attacker may probe the model with series of inputs to receive input-output pairs an perform analysis on these. The results of such an analysis may reveal the decision boundaries and other model specific details, which can then be used for other malicious purposes such as model theft or for the construction of adversarial examples as previously discussed. Furthermore, details about the model used may be involuntarily leaked through sources such as a website, press material, among others.

Table 3.3 presents a subset of model leakage-specific vulnerabilities and controls [26]:

	Vulnerability	Security Control
1	Too much information about the model given in its outputs	Reduce the information given by the model
2	Too much information available on the model	Reduce the available information about the model

Table 3.3: Model leakage vulnerabilities and controls.

The controls presented in table 3.3 are simple in intent, yet not necessarily trivial to implement. Reducing information contained in the model output is generally hard as, since if an output is presented to the user, the output will always contain some information.

ENISA recommends using a method known as *gradient masking* to reduce information contained in model output and implement control 1 [26]. Gradient masking is a defense strategy in which the defender hides or obfuscates gradient information of the ML model [34]. As previously discussed, many types of adversarial attacks use the gradient information which can be analysed from the output to craft adversarial examples. Hence, gradient masking can be considered as a defense for both model leakage as well as adversarial attacks. Many forms of gradient masking exists [34], however, ENISA does not specify particular methods to use[26]. It should thus be considered as part of an evaluation which method to use, if any at all, based on whether applicable or not.

Security-by-design Implementing the aforementioned ML-specific defenses can be done throughout an AI systems life cycle, albeit with significant cost and inconvenience the further in the life cycle the implementation happens. Ideally, an AI system should be developed using, in conjunction with the ethics-by-design approach presented under the fundamental rights sub-requirement in section 3.3, a security-by-design approach. As part of this approach, the aforementioned defenses should be evaluated and considered. Such an evaluation should be risk-based and evaluated as part of a risk assessment. More on risk assessment and management in section 3.5.2.

As stated in the beginning of this section; not only the ML-specific security aspects should be considered, but also the underlying infrastructure, hardware and software. These aspects should also be considered as part of the security-by-design process. However, for the purposes of this thesis, will not be considered directly as an analysis of all these aspects would be notably more extensive and is thus considered out of scope.

Fallback Plan and General Safety

In the case that an AI encounters problems from which in cannot continue operation, adequate safety measures should be in place to either ensure continued operation or safe shutdown.

The AI HLEG considers two straightforward options for continued operation in case of problems [9]:

1. The AI system may switch from a statistical/heuristic approach to a rule-based one.
2. The AI system may incorporate a *human-in-x* approach similar to the ones discussed in the human oversight sub-requirement.

The first options may happen if the AI system must be confident in the produced prediction over a certain threshold. If this is not the case and no prediction satisfying the threshold can be found, it may be unable to continue. In the case of a rule-based approach, a prediction will always be made. This is because in a rule-based approach, decision boundaries are strict and any input data will always terminate to a prediction.

It should be noted that these options only work in the case that the AI system is still functional. In the case of problems leading to a complete shutdown or unavailability, there should exist a contingency plan. This is relevant as unavailability of the AI system may pose a severe business risk. This should be considered as part of a risk assessment.

Accuracy, Reliability and Reproducibility

Given that AI systems are, due to the nature of how machine learning works, not 100% accurate for all prediction tasks, care should be taken to ensure as high an accuracy as possible. This is particularly important if the decisions produced by an AI system can affect human lives, such as in the case of autonomous vehicles.

As stated in section 1.2, the quest for increased accuracy has been an ongoing topic of research and engineering since the inception of machine learning. However, as stated in the same section, the pursuit of increased accuracy has led to increased complexity. For this reason, this sub-requirement should be considered in conjunction with the auditability sub-requirement presented in 3.3 as well as in terms of the risk to the explainability of the AI system. Generally it follows that more accuracy comes at the cost of explainability [29].

In a nutshell; a slightly more accurate model should only be chosen if it does not negatively affect the auditability and explainability of the AI system. Here we see tensions arise between the different requirements; in the case of AI systems having very high requirements for accuracy, such as in the case of autonomous vehicles, a more complex model is preferable due to the increased accuracy provided. However, given that the AI system can directly affect human lives and safety, high requirements for auditability and explainability, to ensure proper operation, entails. Tensions such as this should be carefully considered.

As is the case with accuracy, the intend of reliability is evident; an AI system needs to be reliable in order to be trustworthy. Reliability can be increased through *extensive testing* in which the training data is augmented with edge-case examples of the original data. Figure 3.10 show an example of such an augmentation.

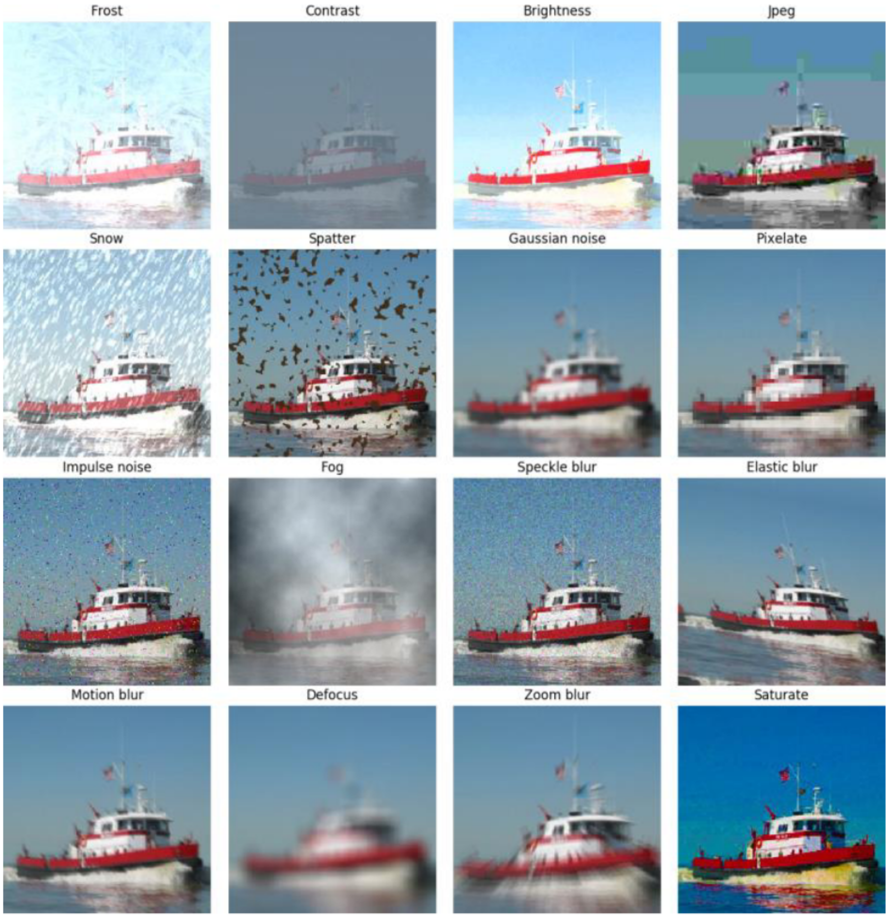


Figure 3.10: Extensive testing using augmented version of the original images to mimic weather conditions and noise [29], [35].

This increases reliability by training the ML model on examples that mimic situations that may occur.

Lastly is the notion of reproducibility. Being able to reproduce data, predictions and situations is needed to facilitate auditability of an AI system. Hence, it is important to consider measures to implement this.

Reproducibility means that the AI system should produce the same results and exhibit the same behaviour if tested again at a later point in time, under the same exact conditions. In the case that an AI system is self-learning, it may not be able to reproduce the same conditions as when a particular behaviour was exhibited. To mitigate this the AI HLEG recommends the use of *replication files* [9], which contain a copy of the AI system state and data at a given point in time. This can facilitate the reproduction of behaviour and outcomes of the AI system.

Privacy and Data Protection

This sub-requirement is extensive by nature. This is because several elements and aspects such as; data governance, the GDPR (EU-context) as well as data and privacy protection methods need to be considered as part of a data protection strategy.

The issue of privacy and data protection relates closely to the vulnerability of data leakage described by ENISA. Data leakage directly affects the *confidentiality* of the data used by the AI system [25]. Data leakage can happen in multiple places in the AI systems life cycle, such as during data handling (data collection, data processing, etc.), model training and during operation.

Table 3.4 presents a subset of data leakage-specific vulnerabilities and controls [26]:

	Vulnerability	**Security Control**
1	The model can allow private information to be retrieved	Ensure that models respect differential privacy Reduce the information given by the model
2	Disclosure of sensitive data for ML algorithm training	Use federated learning to minimise the risk of data breaches

Table 3.4: Data leakage vulnerabilities and controls.

The implementation of the controls presented in table 3.4 are based on methods which are a part of the concept of *privacy preserving machine learning*.

Privacy preserving machine learning (PPML) is a collective term for methods that aim to protect the privacy of data and models used in machine learning [24]. In the following we will look into methods for PPML.

Federated Learning Federated learning is machine learning technique in which multiple clients collaborate to train a ML model [24]. This collaboration is orchestrated by a central server and enables the training of a ML model while keeping the training data decentralized [24]. This process is illustrated in the left side of figure 3.11.

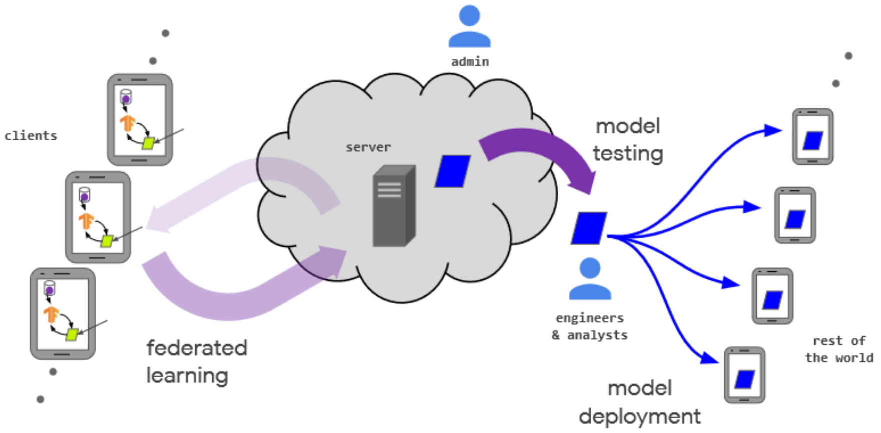


Figure 3.11: Federated learning system life cycle with various actors [36].

This addresses the privacy concern of sensitive data leaking as a consequence of training the model at an untrusted source. It is, however, important to note that federated learning only addresses this specific concern and does not guarantee that the model does not leak private information in other ways or that private data can not be reconstructed based on model output.

To mitigate the threat of sensitive data being present in the model, and being reconstructable from model output, *differential privacy* can be implemented.

Differential Privacy Differential privacy is method (or a set of methods) for generating publicly shareable information about groups within a dataset whilst withholding information about individuals in the dataset [24]. This works by adding a controlled amount of noise to the dataset, which in turn obscures the contribution of each data entry (or individual) in the dataset [24]. This means that the inclusion / exclusion of a given data point or sample in the dataset does not significantly alter the probability of a particular outcome [37]

As previously stated, differential privacy mitigates the threat of sensitive information leaking into the model output. Thus implementing control 1 of table 3.4.

Privacy-preserving Synthetic Data Using synthetically generated data to eliminate the potential of sensitive data leaks is another methods for data protection [38]. The goal is to reproduce the patterns of the original data by creating new data which has the same characteristics as the original dataset. By reproducing the characteristics and distributions of the original data, the model produces the same predictions as it would using the original data.

Generating synthetic data can be done in several ways, with methods using neural networks to generate synthetic data are being developed [37].

An important note is that synthetic data has the benefit of not being subject to the requirements of the GDPR, as the synthetic data does not contain any personal information. This dramatically increases the utility of the data. However, the GDPR is still applicable to the original data, meaning that the original data must be kept and handled in a secure and private manner.

It should, however, be noted that the generation and use of synthetic data is still in its infancy [37] and extensive testing should be performed when using synthetic data in order to ensure that accuracy and reliability is not affected. Similarly, the use of synthetic data does not entail that data or model details may not leak. In fact, the vulnerabilities presented in table 3.3 still apply when using synthetic data.

Privacy-by-design Ideally, an AI system should be developed following a privacy-by-design approach. This should be done in conjunction with the ethics- and security-by-design approached presented under the fundamental rights and resilience to attack and security sub-requirements respectively. The aforementioned methods could technically be implemented during all stages. However, as privacy measures and data protection need to be addressed under the GDPR, they should be implemented during the design and development phase along with the data collection and processing phase.

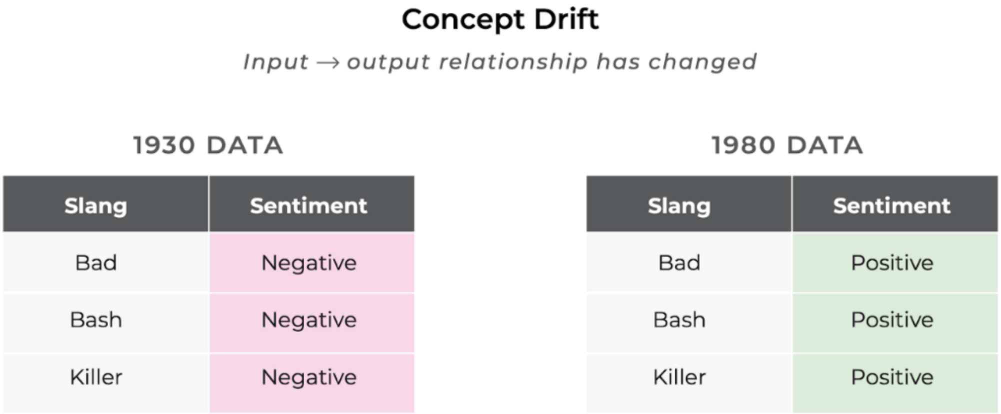
It is important to note that this is not an exhaustive list of all privacy and data protection measures available for AI. The measures were chosen based on their prevalence in literature as well as their appearance in ENISA recommendations.

Quality and Integrity of Data

The quality of the data used to train AI systems are of paramount importance, as a low-quality dataset may contain biases, inaccuracies, errors or even malicious in the data.

Here it is important to consider and document the manner in which the data was obtained. Generally, data can be gathered by the organisation itself or purchased through a data provider. In either case, the data has to be evaluated and sampled. The evaluation and sampling should consider the general integrity of the data and check for inaccuracies, biases and potential security and/or privacy related issues contained in the data.

Additionally, it is important to have some plan for data management both at the data gathering phase as well as during operation of the AI system. This is because of a phenomenon known as *data cascades*; issues and problems present in the data which cause downstream negative effects on model performance and accuracy [39]. Data cascades happen because of errors or changes in data which manifests at a later point in time. An example of a change in data is what is called *concept drift*. Concept drift can take many forms and have varying definitions [40], however, an example by the company *Arize AI* can be seen in figure 3.12.



Concept Drift

Input → output relationship has changed

1930 DATA

Slang	Sentiment
Bad	Negative
Bash	Negative
Killer	Negative

1980 DATA

Slang	Sentiment
Bad	Positive
Bash	Positive
Killer	Positive

Figure 3.12: Intuition of concept drift illustrated using sentiment of slang words [41].

Here the concept of certain slang words have their sentiment change over time. This means that, given the context in which they are used, an AI system may wrongly classify the sentiment of a sentence if the underlying ML model was trained on data which contains the 1930 sentiment [41].

It is important to note that many other metrics for data quality and errors in data beyond what has been mentioned exists [39], [40], however, discussing all such metrics and errors would be very extensive and is thus considered out of scope. Additionally, for the purposes of generating an implementation suggestion for this sub-requirement, it is sufficient to discuss a subset of these and divert focus to their existence rather than their specifics.

Data cascades and concept drift are important to consider, as their effects may have severe consequences.

Take the example of an AI system used for intrusion detection in an online video content providers (such as YouTube) network. In this example case; the *intrusion detection system* (IDS) is trained on network traffic data (possibly purchased from a data provider) which was primarily captured in the early days of the Internet from the mid 1990's to the mid 2000's. In this case we can hypothetically assume that most internet traffic used the *transmission control protocol* (TCP) for communication, as websites for video streaming and such didn't exist yet. Hence, as video streaming became more popular, the use of the *user datagram protocol* (UDP) starts to see a dramatic increase to facilitate the demand for video streaming. This quick rise in UDP traffic to the organisations network may alarm the IDS, as it may be trained to classify excessive UDP traffic as a sign of danger. Thus it may raise and alarm or drop the packets, thus leading to users not being able to use the service. This problem might be a hypothetical and extreme example, but illustrates the problems posed by data cascades and concept drift in a real-world environment and use-case.

To conclude; in order to implement this sub-requirement, there should be awareness of the data used and how it may contain biases, inaccuracies, errors, etc. Which may affect the AI system. In a nutshell; a strategy for data management should be put in place.

Access to Data

As all AI systems rely on good quality data to ensure its performance, all data used by the AI system throughout its life cycle should be considered an important resource which needs to be secured.

For this reason, good access control measures need to be in place. This could and should be considered as part of the organisations current identity management and access policies.

Table 3.5 presents a subset of access-specific vulnerabilities and controls [26]:

	Vulnerability	**Security Control**
1	Poor access rights management	Apply a RBAC model, respecting the least privileged principle
2	Weak access protection mechanisms for ML model components	Ensure ML applications comply with identity management, authentication, and access control policies

Table 3.5: Access vulnerabilities and controls.

Using an *rule-based access control* (RBAC) model to define data access allows for the enforcement of only allowing certain entities have access to the data based on their role. This means that access can be allowed for all carrying the role of data scientists and auditors whilst being disallowed for all other roles. For example; the HR staff may not have any legitimate interest or reason for having access to the AI systems data and should thus, according to the *least privileged principle*, not be allowed access.

These measures should easily integrate into any identity management solution and access policies an organisation might have.

Traceability

Ensuring traceability in an AI system has several benefits as it allows for increased transparency in the AI systems operation. This facilitates the identification of the reasons behind erroneous decisions/predictions by the AI system, which can then be used by developers to prevent future mistake [9]. Additionally, it facilitates auditability by being able to trace the AI systems operation.

Naturally, traceability may originate from good and extensive logging practises. Logs should include all relevant data as to identify problems and to conform with any relevant auditability obligations as determined by a respective analysis of the auditability sub-requirement presented in 3.3.

As hinted at in the auditability sub-requirement, a technical implementation of traceability is the implementation of audit trails. Audit trails are traceable logs, similar to the ones described above, however, with the difference that in addition to logging steps in system operation, the design process and testing results should also be logged [24]. This enables auditors to check whether a system performs as intended and specified in design and testing of the AI system.

The amount of traceability required for the purposes of auditing depends on the context and intended use of the AI system, whereas it is always beneficial to maintain some amount of traceability for the debugging purposes previously described.

Explainability

Explainability is particularly important for improving trust in AI systems and has been one of the main drivers in conceptualising Trustworthy AI as a field of interest. This is because of the complexity-explainability problem mentioned in section 1.2. It naturally follows, that people will not trust systems which they do not understand and thus, by extension, not trust decisions for which they can not be given an adequate explanation.

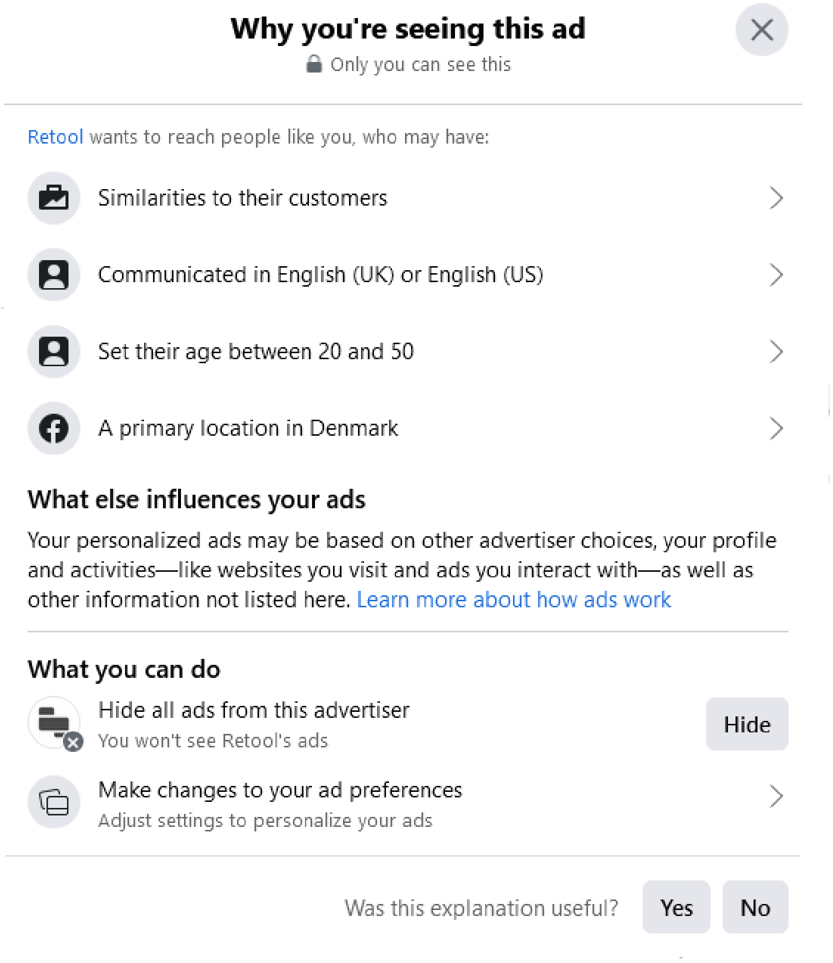
An explanation should always be adapted to the expertise of the stakeholder to which the explanation is relevant (layperson, regulator, etc.) [9]. Naturally, the average citizen/user of an AI system can not be expected to understand concepts such as statistical distributions and feature dependencies. This means that organisations should be mindful of the explanations provided by their implemented explanation methods.

As was introduced in section 2.3, there exists two levels of scope of explanations; local and global. It is important to consider the scope of explainability required for a particular use-case.

For example; a worker overseeing an AI-driven automated assembly line will, in the case of problems, only need an explanation for why that particular error occurred such that it may be fixed. In this case any local-scoped explanation method, which can sufficiently explain why such an error happened, will be adequate.

However, for an auditor it may not be sufficient or even relevant to provide local explanations. This is because an auditor possibly has to audit the entire AI system and how it makes decisions in order to document or certify that the system is operating as specified. Here global explanations may be required.

A real-world example of an AI system providing explanations is shown in figure 3.13. This figure shows a tab from the Facebook web interface which presents information as to why a particular advertisement was shown to the user.



Why you're seeing this ad

Only you can see this

Retool wants to reach people like you, who may have:

- Similarities to their customers
- Communicated in English (UK) or English (US)
- Set their age between 20 and 50
- A primary location in Denmark

What else influences your ads

Your personalized ads may be based on other advertiser choices, your profile and activities—like websites you visit and ads you interact with—as well as other information not listed here. Learn more about how ads work

What you can do

Hide all ads from this advertiser
You won't see Retool's ads

Make changes to your ad preferences
Adjust settings to personalize your ads

Was this explanation useful?

Figure 3.13: "Why am I seeing this ad?"-tab on Facebook.

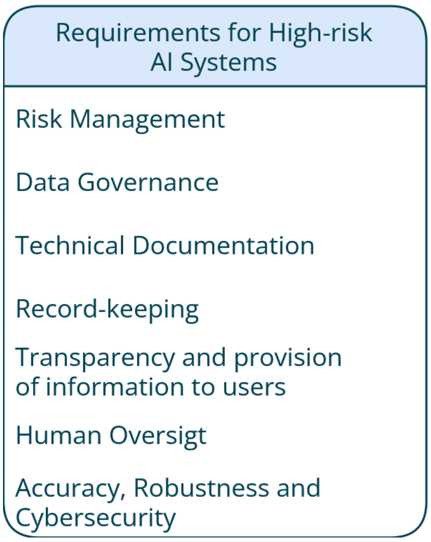
The example presented in figure 3.13 is an example of a local explanation, as it provides an explanation as to why a particular ad was shown. It is currently not known how these explanations are generated as *Meta* has not disclosed this information. However, it is clear, based on the information provided in the tab, that some form of feature importance method (possibly similar to SHAP) must be used in order to display that ad. It directly states that information such as location and age were important for showing that particular ad.

As stated in section 2.3, explainability for ML models can be achieved by either using an intrinsically explainable or through some post hoc explanation method. In the case of the latter, the SHAP method (section 2.3.2) is a prime examples of a method which can be used to provide local and interpretable explanations. Naturally the use of these methods needs to be evaluated and implemented based on whether they fit in providing a good explanation given the intended use of the AI system as well as the people which use it.

The takeaway here is that explanations need to be considered in the context that they are given. For end users the explanations should be considered from the point of the user of the AI system under its intended use. This entails that the explanations provided by the system should be understandable and meaningful in the context which they are provided.

3.5 The AI Act

As stated in section 2.2.4, the following requirements exists for high-risk AI systems according to the AI Act.



Requirements for High-risk AI Systems

Risk Management

Data Governance

Technical Documentation

Record-keeping

Transparency and provision of information to users

Human Oversigt

Accuracy, Robustness and Cybersecurity

Figure 3.14: AI Act requirements for high-risk AI systems.

3.5.1 Scope and Delimitation

Given that the AI Act is extensive in scope, it is not feasible to analyse every single article and paragraph within the bounds of this thesis. This entails some delimitation which is explained in the following.

Several paragraphs of the AI Act specify requirements which are directly linked legislation of specific sectors such as, among others, *credit institutions*, *medical devices*. These requirements will not be handled as they are fulfilled only through the fulfillment of the relevant sectorial legislation and are thus considered out of scope of this thesis.

An article or paragraph may require that a certain requirement or procedure needs to be done continuously throughout an AI system's life cycle. This can naturally not be performed due to the constraints of this thesis and will thus only be commented upon. The same applies for when the requirement of an article or paragraph refers to post-market monitoring of an AI system, as this is not possible for the same reason as stated before.

3.5.2 Risk Management

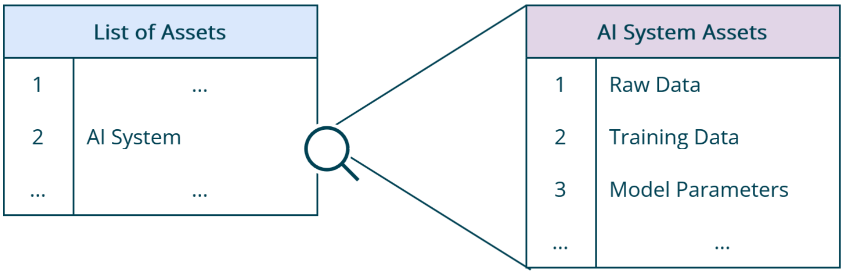
In accordance with chapter II, article 9 of the AI Act; a risk management system must be implemented and documented for the high-risk AI system. A risk management system is considered any risk management strategy or framework which encompasses the risk requirements presented in this section.

The risk management system must adhere to the requirements directed in Article 9, meaning it must [10]:

1. Identify and analyse known and foreseeable risks associated with the high-risk AI system.
2. Estimate and evaluate the risks that may emerge through foreseeable intended and malicious use of the high-risk AI system.
3. Evaluate other possibly arising risks based on the analysis of data gathered from the post-market monitoring system referred to in Article 61.
4. Adopt risk management measures which are in accordance with the following provisions:
    - The risk management measures must take into account the state of the art and relevant standards.
    - The risk management measures must be made such that the risks associated are deemed acceptable when the high-risk AI system is used as intended. These risks must be communicated to the user.
    - The high-risk AI system must be tested to find the most appropriate risk management measures.
        - The testing procedures must ensure that the high-risk AI system performs as intended.
        - The testing of the high-risk AI system must be performed during the development process or at any point prior to putting the AI system on the market.

Given these requirements for risk management it is safe to assume that a proper implementation of any well-known risk management frameworks such as the *ISO/IEC 27005 27005* or *NIST Risk Management Framework (RMF)* will be adequate for the fulfilment of this requirement.

It should be important to note that these frameworks are not AI-specific in their design. This means that an AI system should not be viewed as a single asset, but as a collection of several individual assets; here included is the training data, the model, tools, algorithms used, etc. Figure 3.15 illustrates this concept.



List of Assets	
1	...
2	AI System
...	...

AI System Assets	
1	Raw Data
2	Training Data
3	Model Parameters
...	...

Figure 3.15: Simplified view of AI assets as part of a risk management framework. AI assets based on [25].

NIST is currently working on a AI-specific risk management framework [42], however, it is still only considered a draft with a scheduled release in late 2022 or 2023 and is thus not considered viable for the implementation for this requirement.

3.5.3 Data Governance

In accordance with chapter II, article 10 of the AI Act; the the following data governance and management practises should be considered during development of the high-risk AI system [10]:

1. Design choices regarding data and the collection and use thereof must be documented.
2. Data processing operations such as; annotation, labelling, cleaning, enrichment and aggregation must be documented and considered.
3. Assumptions on the information that the data are supposed to measure and represent should be should be considered and documented.
4. Assessment of the availability, quality and suitability of the data sets that are needed.
5. Data must be examined for possible biases.
6. Identification of possible data gaps or shortcomings and how these can be addressed.

Additionally, the following should also be considered where applicable [10]:

- The data used must be complete and representative of the persons or groups on which the high-risk AI system is intended to be used.
- The data used must take into account the specific geographical, behavioural or functional setting within which the high-risk AI system is intended to be used.

- If personal data is used, privacy-preserving measures such as pseudonymsation, anonymisation or encryption must be used.

For this general requirement we see that it draws several parallels to many of the previously discussed requirements for Trustworthy AI.

For example; bias management has to be taken into account in the same manner as described in bias management sub-requirement in section 3.3.

It is important to note that the consideration regarding privacy-preserving measures needs to be made in the context of EU law, as the GDPR is responsible for most of the handling of personal data in this regard. However, the privacy-preserving measures presented in the privacy and data governance sub-requirement in section 3.4 can be implemented for this requirement as part of this consideration.

Additionally, the concepts presented in the quality and integrity of data sub-requirement apply for the implementation of the identification and addressing of data gaps and shortcomings.

We also see new considerations being present in the requirements. Generally, everything regarding data management practises must be documented.

3.5.4 Technical Documentation

In accordance with chapter II, article 11 of the AI Act; the following requirements for technical documentation of the high-risk AI system applies [10]:

- The technical documentation of a high-risk AI system shall be drawn up before the system is placed on the market or put into service and must be kept up-to date.

The technical documentation must at a minimum contain all that which is specified in Annex IV of the AI [11]:

1. A general description of the AI system including:
    (a) Its intended purpose, the person(s) developing the system the date and the version of the system.

    (b) Where applicable; how the AI system interacts or can be used to interact with hardware or software that is not part of the AI system itself.

    (c) The versions of relevant software or firmware and any requirement related to version updates.

    (d) The description of all forms in which the AI system is placed on the market or put into service.

   (e) The description of hardware on which the AI system is intended to run.

   (f) Where the AI system is a component of products; photographs or illustrations showing external features, marking and internal layout of those products.

   (g) Instructions of use for the user and, where applicable installation instructions.

2. A detailed description of the following elements of the AI system:

   (a) The methods used in the development of the AI system, including which third-party tools or systems have been used.

   (b) The design specifications of the AI system. Here meaning the general logic of the system, key design choices, what the system is designed to optimise for, relevance of different parameters and possible trade-offs.

   (c) The system architecture and the computational resources required to develop, train, test and validate the AI system.

   (d) Where relevant, the training methodologies and training datasets used, including how the data was obtained and how it has been processed.

   (e) An assessment of the human oversight measures needed.

   (f) An assessment of the measures needed to interpret (explain) the outputs of the AI system.

   (g) If applicable; any pre-determined changes to the AI system along with measures to ensure continued compliance.

   (h) Metrics used to measure accuracy, robustness, cybersecurity and compliance with other relevant requirements.

   (i) The risk management system used.

   (j) The foreseeable unintended outcomes and risks to health, safety and discrimination.

This requirement is rather simple in intent as it simply lays out the requirements for what should be contained in the documentation of the AI system. This is required in order to document conformity with the requirements presented in 3.5.

Hence, the implementation suggestion for this requirement is to simply maintain technical documentation which adheres to the aforementioned specifications.

3.5.5 Record-keeping

In accordance with chapter II, article 12 of the AI Act; the high-risk AI system must be designed and developed with capabilities enabling the automatic logging of events [10].

The logging capabilities must ensure level of traceability appropriate to the intended purpose of the AI system [10].

If the system is used for remote biometric identification (RBI), the following must be contained in the produced logs [10]:

1. The period of use of the system (start/end date and time for each use).
2. The database against which the input data was checked.
3. The input data which lead to a match.
4. The identification of persons involved in the verification of the result.

Here it should be noted that, with the exception of RBI systems, what is considered "appropriate" to the intended purpose of the AI system is not directly specified. This means that it is up to the organisation to decide what is considered appropriate logging capabilities.

Similar to the previous requirement for technical documentation, this requirement is simple in intend. It is created to ensure that the operation of the system can be audited.

Hence, the implementation suggestion for this requirement is to implement an adequate logging mechanism which adheres to the aforementioned specifications.

3.5.6 Transparency and Provision of Information to Users

In accordance with chapter II, article 13 of the AI Act; high-risk must designed and developed in such a way that their operation is sufficiently transparent to enable users to use it properly [10].

High-risk AI systems must be accompanied by usage instructions which must contain [10]:

1. The identity and contact information of the provider of the system.
2. The capabilities and limitations of the AI system, including:
    (a) Its intended purpose.
    (b) The level of accuracy, robustness and cybersecurity (referred to in section 3.5.8) against which the high-risk AI system has been tested.

(c) Known or foreseeable circumstances which may lead to risks to health, safety or fundamental rights.

(d) If applicable; The performance on the persons or groups on which the system is intended to be used.

(e) If applicable; specification for the input data to be used.

3. Any pre-determined changes to the AI system or its performance.
4. The human oversight measures (referred to in section 3.5.7) which have been implemented.
5. Any maintenance and care measures needed to ensure the functionality of the AI system.

This requirement is similar intend to the communication sub-requirement presented in section 3.3 as it seeks to provide the user with information about the capabilities and limitations of the AI system.

Given that technical documentation already needs to be maintained, in accordance with section 3.5.4, it should be trivial to draw up usage instructions based on these. This could be done in combination with the suggestions described in the aforementioned communication sub-requirement.

3.5.7 Human Oversight

In accordance with chapter II, article 14 of the AI Act; high-risk AI systems must be designed and developed such that they can effectively be overseen by a human during the period in which the AI system is in use [10].

The intent here being that human oversight may prevent or minimise risks posed by the AI system.

Based on the requirements put forth in the article, human oversight of the AI system must be ensured in either of the following ways [10]:

1. Human oversight is built into the AI system by the provider.
2. Human oversight is implemented by the user of the AI system.

Here it is meant that human oversight mechanism should either be directly implemented into the AI system, such that the operation of the AI system requires human oversight. Being implemented by the user means that the human oversight mechanism is not part of the AI system directly. Instead human oversight is done separately from the AI system.

In either case, the following requirements apply for the implemented human oversight mechanism. Specifically, the human oversight mechanism must enable the individual(s) assigned to the oversight process to do the following [10]:

1. Fully understand the capabilities and limitations of the high-risk AI system and be able to monitor its operation for signs of anomalies, errors and unexpected performance.
2. Remain aware of the possible risk of relying on the output produced by a high-risk AI system (called *automation bias*).
3. Be able to correctly interpret the AI system's output, using the available interpretation tools and methods.
4. Be able to disregard, override or reverse the output of the AI system.
5. Be able to intervene or interrupt the operation of the AI system using a "stop" button or similar procedure.

Here it should be noted that the notion of "interpretation" synonymous with explanation, as it relates to the extend to which an adequate explanation for the AI systems output can be made. Naturally, this entails that such an explanation is interpretable by the user. However, it is safe to assume that measures to provide explanations for the output, such as the ones described in the explainability sub-requirement presented in section 3.4, are usable in this context as well.

The human oversight mechanisms described in the human oversight sub-requirement in section 3.4; HITL, HOTL, HIC are applicable here as well. It should be considered which one of them is the most appropriate of the intended use of the AI system. The mechanism should support the addition of an emergency stop button, as well as the ability to override output, as specified by the aforementioned requirements.

Similar to the record-keeping requirement (section 3.5.5), there are additional requirements for human oversight mechanisms for RBI systems. These are [10]:

1. No action or decision is taken by the user on the basis of the identification resulting from the system unless this has been verified and confirmed by at least two natural persons.

This means that if a RBI system identifies a person, no action must be taken, based on this identification, unless verified and confirmed by two persons.

The requirements for what a human oversight mechanism must contain are clearly specified, and any of the aforementioned human oversight mechanisms, be it a HITL, HOTL or HIC-based mechanism, may be used as long as they conform to the specified requirements.

3.5.8 Accuracy, Robustness and Cybersecurity

In accordance with chapter II, article 15 of the AI Act; high-risk AI systems must be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness and cybersecurity throughout their life cycle [10].

What constitutes an appropriate level is based on the intended use of the AI system, but is not directly specified in the AI Act. Hence, it is up to to interpretation by the organisation developing the AI system.

Specified in the article is the following [10];

1. The AI system must be resilient to errors, faults or inconsistencies that may occur in the system or the environment in which it operates.
2. The AI system must be resilient to attempts by unauthorized third parties to alter the AI systems use or performance by exploiting vulnerabilities in the system.

To summarize; a high-risk AI system must be reliable, robust and secure. This means that the implementation suggestions presented in the resilience to attack and security sub-requirement in section 3.4, can be conveniently used as implementation suggestions for this requirements as well.

However, the implementation of these should keep in mind the following [10]:

1. *"The technical solutions aimed at ensuring the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks."*
2. *"The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent and control for attacks trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws."*

3.6 Compilation of Implementation Suggestions

Based on the analysis performed in section 3.3, 3.4 and 3.5, the implementation suggestions of the requirements in the aforementioned sections have been compiled. The results can be seen in figure 3.16. The implementation suggestions have been color coded to illustrate to what they relate. Figure 3.17 shows legend describing the color coding.

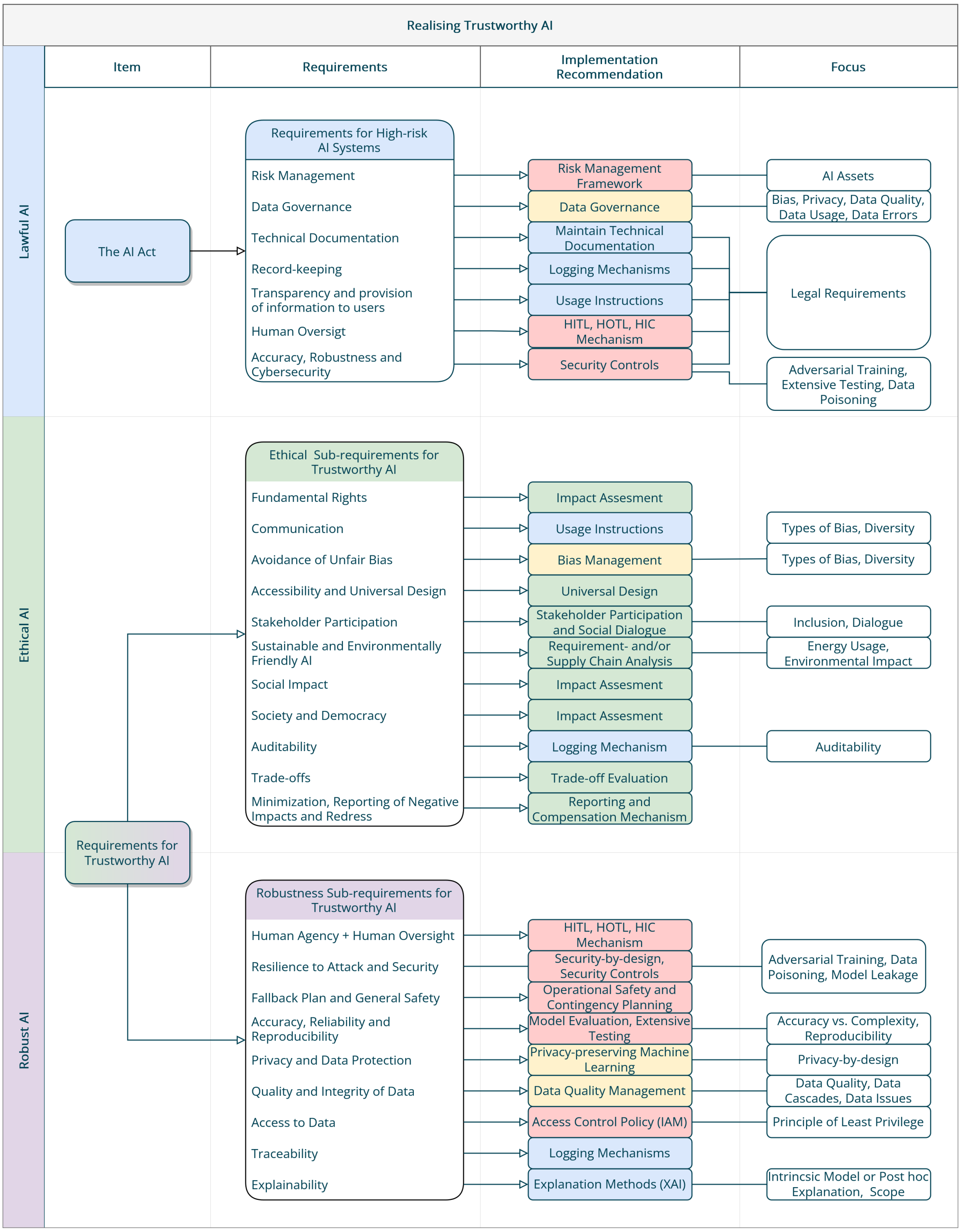

Realising Trustworthy AI
Item
Requirements
Implementation Recommendation
Focus
Lawful AI
The AI Act
Requirements for High-risk AI Systems
Risk Management
Data Governance
Technical Documentation
Record-keeping
Transparency and provision of information to users
Human Oversigt
Accuracy, Robustness and Cybersecurity
Risk Management Framework
Data Governance
Maintain Technical Documentation
Logging Mechanisms
Usage Instructions
HITL, HOTL, HIC Mechanism
Security Controls
AI Assets
Bias, Privacy, Data Quality, Data Usage, Data Errors
Legal Requirements
Adversarial Training, Extensive Testing, Data Poisoning
Ethical AI
Requirements for Trustworthy AI
Ethical Sub-requirements for Trustworthy AI
Fundamental Rights
Communication
Avoidance of Unfair Bias
Accessibility and Universal Design
Stakeholder Participation
Sustainable and Environmentally Friendly AI
Social Impact
Society and Democracy
Auditability
Trade-offs
Minimization, Reporting of Negative Impacts and Redress
Impact Assesment
Usage Instructions
Bias Management
Universal Design
Stakeholder Participation and Social Dialogue
Requirement- and/or Supply Chain Analysis
Impact Assesment
Impact Assesment
Logging Mechanism
Trade-off Evaluation
Reporting and Compensation Mechanism
Types of Bias, Diversity
Types of Bias, Diversity
Inclusion, Dialogue
Energy Usage, Environmental Impact
Auditability
Robust AI
Robustness Sub-requirements for Trustworthy AI
Human Agency + Human Oversight
Resilience to Attack and Security
Fallback Plan and General Safety
Accuracy, Reliability and Reproducibility
Privacy and Data Protection
Quality and Integrity of Data
Access to Data
Traceability
Explainability
HITL, HOTL, HIC Mechanism
Security-by-design, Security Controls
Operational Safety and Contingency Planning
Model Evaluation, Extensive Testing
Privacy-preserving Machine Learning
Data Quality Management
Access Control Policy (IAM)
Logging Mechanisms
Explanation Methods (XAI)
Adversarial Training, Data Poisoning, Model Leakage
Accuracy vs. Complexity, Reproducibility
Privacy-by-design
Data Quality, Data Cascades, Data Issues
Principle of Least Privilege
Intrincsic Model or Post hoc Explanation, Scope

Figure 3.16: Realising Trustworthy AI - Compilation of requirements with implementation suggestions and focus areas.

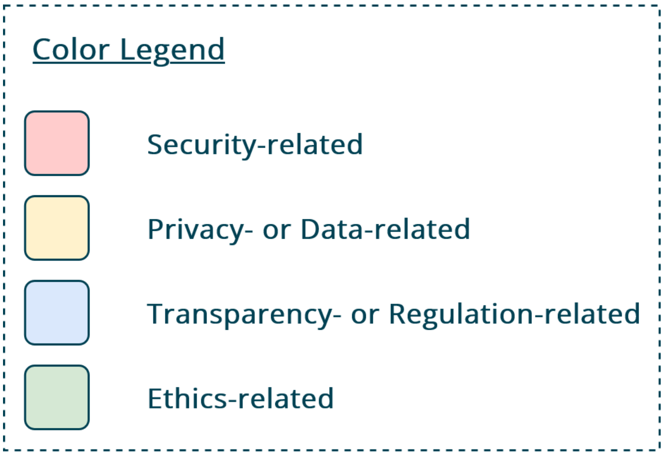


Figure 3.17: Color coding used in figure 3.16.

Chapter 4

Operational Model

This chapter will describe the design and creation of an operational model for realising the guidelines for Trustworthy AI set by the AI HLEG as well as the requirements for high-risk AI systems according to the AI Act.

4.1 Model Design

The operational model has been designed based on the requirements analysed in chapter 3. The general design process to operationalise each requirement in the flowchart is by describing each of the implementation suggestions as a yes/no question to which an implementation suggestion can be provided. Figure 4.1 illustrates this idea for a single requirement.

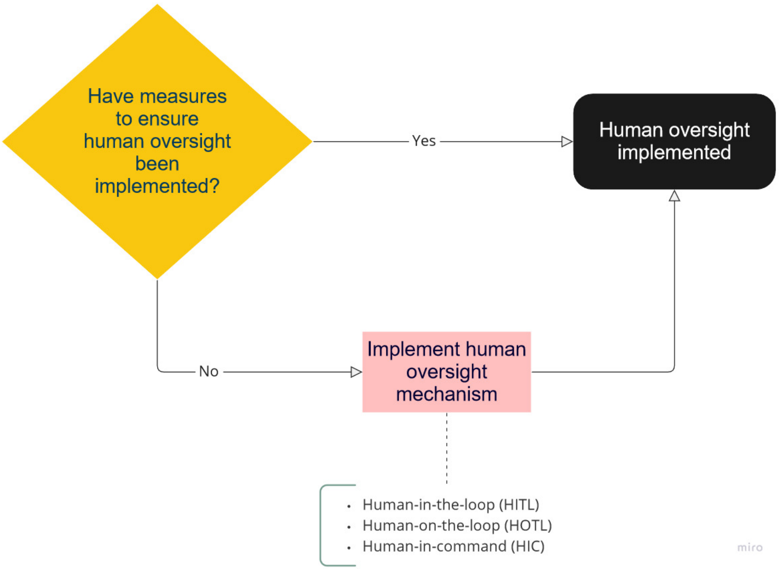
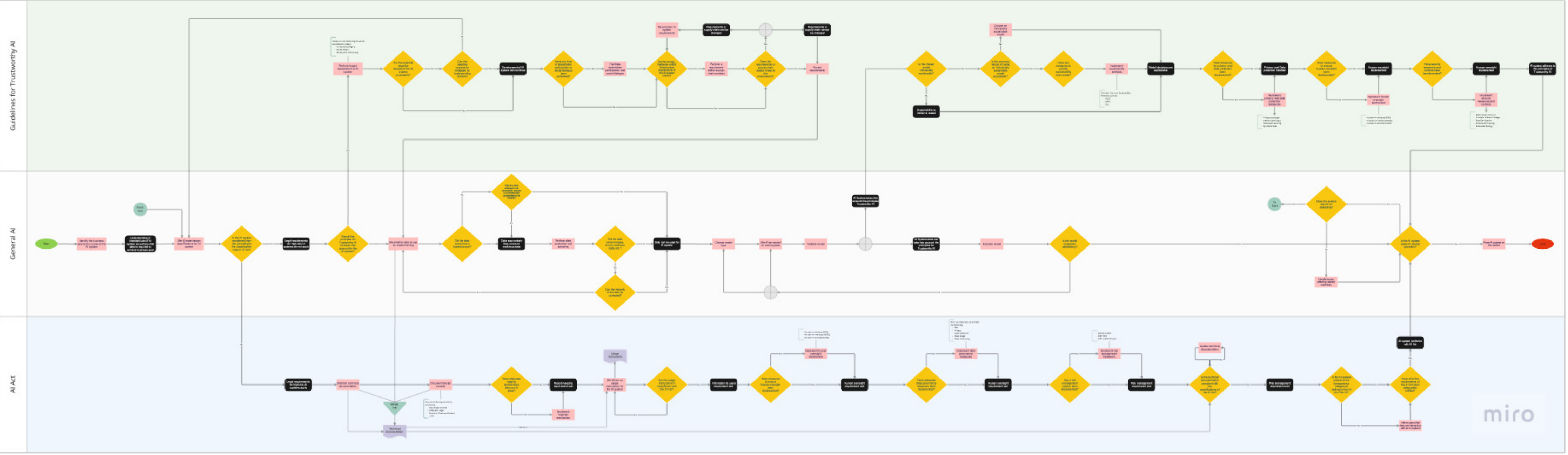

Have measures to ensure human oversight been implemented?
Yes
Human oversight implemented
No
Implement human oversight mechanism
Human-in-the-loop (HITL)
Human-on-the-loop (HOTL)
Human-in-command (HIC)
miro

Figure 4.1: Operational model of the human oversight sub requirement.

Naturally, some creative liberties had to be taken in regards to the ordering of the requirements and the connections points. This is because the operational model is meant as a management tool and thus the ordering of certain elements are not considered crucial to the understanding of the model, nor the implementation of the requirements.

4.2 Model Prototype

The aforementioned design process has resulted in the creation of the following model prototype. It should be noted that given the models large size it may be hard to view in the thesis report or `.pdf` format. For better visibility, the entire model is viewable online on: .


Guidelines for Trustworthy AI
General AI
AI Act
Start
Accept requirements
Choose model type
Validate model
Evaluate model
Privacy and Data protection handled
Human oversight implemented
To Start
End
miro

Chapter 5

Discussion

This chapter will present a short discussion of the different known limitations.

5.1 Limitations

5.1.1 Intended Use

As stated in section 4.1, the implementation suggestions and the operational model are intended to be used as management tools. Hence, it should not be viewed as more.

5.1.2 Implementation Suggestions

As previously stated the implementation suggestions presented are not an exhaustive list of the possible methods which can be used to implement the requirements. It is entirely possible that some of the presented recommendations are not considered best-practises or possibly even inadvisable.

5.1.3 The AI Act

It is important to note that as of writing this thesis the AI Act is still a draft legislation and has not yet been approved by the European Parliament. This means that several of the definitions and specifics of the AI Act, presented in this thesis, may be subject to change in the finalized version of the AI Act.

5.1.4 Legal Specifics of the AI Act

Given that the AI Act is a legal document, the definitions and formulations used within may not be interpreted in a correct manner. This should be done by a professional within the field of law.

5.1.5 Operational Model Evaluation

Given that the operational model for realising Trustworthy AI presented in section 4.2 is currently only a prototype it is important to note that it has not yet been evaluated in any way or form. Naturally, this means that many of the proposed ideas and specifics of the model are subject to change, based on feedback and evaluation. It is not a guaranteed to be a complete set of best practises and implementations.

Chapter 6

Conclusion and Future Work

This chapter will present the conclusion of the problem formulation along with a specification of the primary contributions of the thesis, as well as suggestions for future work.

6.1 Conclusion

This thesis aimed to research and analyse the components of Trustworthy AI and how they can be realised.

In sections 3.3, 3.4 and 3.5, the components Ethical AI, Robust AI and Lawful AI were analysed along with the requirements for Trustworthy AI and the requirements for high-risk AI systems respectively. One or more implementation recommendations for methods, tools or techniques to realise each of the analysed requirements have been provided.

The recommendations established by the analysis of the aforementioned requirements have been compiled into figure 3.16.

The implementation suggestions presented in the aforementioned figure was used as a foundation for the design of an operational model for realising all three Trustworthy AI components.

To summarise; an analysis of the components of Trustworthy AI, and how they can be realised and implemented, has been performed. Additionally, the implementation suggestions were operationalised in a flowchart as demonstrated in section 4.2. Combined, this satisfies the problem formulation along with both sub-questions.

It should be noted that an evaluation of the recommendations and operational model has not been performed. Therefore, it is not possible to conclude to which extent solutions presented are viable. For this we refer to future work section.

6.1.1 Contributions

Two primary contributions have been made as a product of this thesis:

1. First is the analysis of implementation suggestions for the requirements for Trustworthy AI. These have been compiled into a figure, which allows for visualization of the implementation suggestions in the context of the requirements.
2. Second is the operational model. This is intended as a management tool for stakeholders and developers of AI systems

6.2 Future Work

As stated in the section (1.5), the project might have benefited from the use of interviews with domain experts. This could be used to evaluate the accuracy and credibility of the proposed implementation recommendations.

Additionally, the operational model prototype presented in section 4.2 has not been evaluated. This naturally leads to future work to evaluate the effectiveness of the model. Here it might be the case that it needs extended with additional entries or shortened in the case that entries are redundant or superfluous.

As a consequence the AI Act not yet having been approved by the European Parliament, the requirements analysed in this thesis may need to be re-evaluated in the case that changes to the AI Act are made in the future.

Bibliography

[1] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: challenges and prospects," *CoRR*, vol. abs/1812.04608, 2018. arXiv: `1812.04608`. [Online]. Available: `http://arxiv.org/abs/1812.04608`.

[2] Ritzau, "Tesla-ulykker med avanceret autopilot efterforskes i usa," *TV2 Nyheder*, [Online]. Available: `https://nyheder.tv2.dk/udland/2021-09-03-tesla-ulykker-med-avanceret-autopilot-efterforskes-i-usa`.

[3] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, pp. 65–386, 1958. DOI: `10.1037/h0042519`. [Online]. Available: `https://doi.org/10.1037/2Fh0042519`.

[4] S. J. Russell and P. Norvig, *Artificial Intelligence, A Modern Approach*, Fourth. Pearson, 2021, ISBN: 978-0-13-461099-3. [Online]. Available: `https://archive.org/details/artificial-intelligence-a-modern-approach-4th-edition`.

[5] J. McCarthy, "What is artificial intelligence," 2004. [Online]. Available: `http://www-formal.stanford.edu/jmc/whatisai.html`.

[6] European Commission, *Fostering a European approach to Artificial Intelligence, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions*. European Commission, 2021. [Online]. Available: `https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence`.

[7] ——, *Fostering a European approach to Artificial Intelligence, ANNEXES to the Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions*. European Commission, 2021. [Online]. Available: `https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review`.

[8] Ministry of Finance and Ministry of Industry, Business and Financial Affairs, *National Strategy for Artificial Intelligence*. Ministry of Finance and Ministry of Industry, Business and Financial Affairs, 2019. [Online]. Available: `https://en.digst.`

dk/policy-and-strategy/denmark-s-national-strategy-for-artificial-intelligence/.

[9] High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI*. European Commission, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[10] European Commission, *Proposal for a Regulation of the European Parliament and of the Council, LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE(ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. European Commission, 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

[11] ——, *ANNEXES to the Proposal for a Regulation of the European Parliament and of the Council, LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE(ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. European Commission, 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

[12] L. Sioli. "A european strategy for artificial intelligence." (2021), [Online]. Available: https://www.ceps.eu/wp-content/uploads/2021/04/AI-Presentation-CEPS-Webinar-L.-Sioli-23.4.21.pdf?.

[13] C. Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, 2nd ed. 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book.

[14] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," *AACL-IJCNLP 2020*, 2020.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. DOI: 10.48550/arXiv.1705.07874. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[16] European Parliament and Council of the European Unions, *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. European Parliament and Council of the European Union, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[17] European Union Agency for Fundamental Rights, *Artificial Intelligence and Fundamental Rights, Getting the Future Right*. European Union Agency for Fundamental Rights, 2020. [Online]. Available: https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, 2021, ISSN: 0360-0300. DOI: 10.1145/3457607. [Online]. Available: https://doi.org/10.1145/3457607.

[19] M. Reagan. "Understanding Bias and Fairness in AI Systems." (2021), [Online]. Available: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3.

[20] R. K. E. Bellamy, K. Dey, M. Hind, *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, 4:1–4:15, 2019. DOI: 10.1147/JRD.2019.2942287.

[21] IBM Trusted AI Team. "AI Fairness 360 (AIF360)." (2022), [Online]. Available: https://aif360.mybluemix.net/.

[22] S. Burgstahler, "Universal Design: Process, Principles, and Applications.," *DO-IT*, 2009. [Online]. Available: https://www.washington.edu/doit/universal-design-process-principles-and-applications.

[23] The Universal Design Project. "What is Universal Design?" (2022), [Online]. Available: https://universaldesign.org/definition.

[24] M. Brundage, S. Avin, J. Wang, *et al.*, *Toward trustworthy ai development: Mechanisms for supporting verifiable claims*, 2020. DOI: 10.48550/ARXIV.2004.07213. [Online]. Available: https://arxiv.org/abs/2004.07213.

[25] European Union Agency for Cybersecurity (ENISA), *Artificial Intelligence Cybersecurity Challenges, Threat Landscape for Artificial Intelligence*. European Union Agency for Cybersecurity (ENISA), 2020, ISBN: 978-92-9204-462-6. DOI: 10.2824/238222. [Online]. Available: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.

[26] ——, *Securing Machine Learning Algorithms*. European Union Agency for Cybersecurity (ENISA), 2021, ISBN: 978-92-9204-543-2. DOI: 10.2824/874249. [Online]. Available: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.

[27] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, vol. 135, pp. 364–381, 2022, ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2022.05.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X22001790.

[28] B. Green and Y. Chen, "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 90–99, ISBN: 9781450361255. DOI: 10.1145/3287560.3287563. [Online]. Available: https://doi.org/10.1145/3287560.3287563.

[29] R. Hamon, H. Junklewitz and I. Sanchez at the EU Commission Joint Research Centre (JRC), "Robustness and explainability of artificial intelligence," Tech. Rep., 2020. DOI: 10.2760/57493. [Online]. Available: https://ai-watch.ec.europa.eu/publications/robustness-and-explainability-artificial-intelligence_en.

[30] K. Eykholt, I. Evtimov, E. Fernandes, *et al.*, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634. DOI: 10.1109/CVPR.2018.00175.

[31] A. Raghunathan, J. Steinhardt, and P. Liang, *Certified defenses against adversarial examples*, 2018. DOI: 10.48550/ARXIV.1801.09344. [Online]. Available: https://arxiv.org/abs/1801.09344.

[32] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 656–672. DOI: 10.1109/SP.2019.00044.

[33] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, *On detecting adversarial perturbations*, 2017. DOI: 10.48550/ARXIV.1702.04267. [Online]. Available: https://arxiv.org/abs/1702.04267.

[34] H. Xu, Y. Ma, H.-C. Liu, *et al.*, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.

[35] D. Hendrycks and T. Dietterich, *Benchmarking neural network robustness to common corruptions and perturbations*, 2019. arXiv: 1903.12261 [cs.LG].

[36] P. Kairouz, H. B. McMahan, B. Avent, *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[37] J. Yoon, J. Jordon, and M. van der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=S1zk9iRqF7.

[38] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *2008 IEEE 24th International Conference on Data Engineering*, 2008, pp. 277–286. DOI: 10.1109/ICDE.2008.4497436.

[39] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380966. DOI: 10.1145/3411764.3445518. [Online]. Available: https://doi.org/10.1145/3411764.3445518.

[40] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean, "Characterizing concept drift," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 964–994, 2016.

[41] A. Dhinakaran. "The Model's Shipped; What Could Possibly go Wrong?" (2021), [Online]. Available: `https://arize.com/blog/ml-model-failure-modes/`.

[42] NIST. "NIST Artificial Intelligence Risk Management Framework (AI RMF)." (2022), [Online]. Available: `https://www.nist.gov/itl/ai-risk-management-framework`.