Thesis for the qualification of the academic degree
Master of Science (M.Sc.) Business Informatics

Faculty for Computer Science and Mathematics
of the Goethe University Frankfurt

Ethical assessment of AI systems in healthcare: A use case

Submitted by

Frederike Laufenberg
6152509
Master Sc. Business Informatics
born on 9/4/1995 in Rendsburg

Teresa Sophia Gabriele Werthmann
7121241
Master Sc. Business Informatics
born on 30/8/1993 in Marktheidenfeld

First examiner: Prof. Dr. Gemma Roig
Second examiner: Dr. Karsten Tolle
Submitted on 20/5/2021

Abstract

The development of artificial intelligence (AI) systems and their use in a wide variety of areas is increasing and is the focus of various current discussions. In this context, ethical aspects are increasingly being addressed and AI-supported processes are being critically assessed from an ethical perspective. For this purpose, it is necessary to define the requirements of an AI system with regard to conformity with ethical values and norms and to evaluate them depending on the use case.

The aim of this work is to examine an AI system in the healthcare sector for supporting processes in the detection of OHCA (Out-of-Hospital Cardiac Arrest) with regard to ethical issues. Since the present use case was developed and tested in Europe and thus European values and norms should be respected, the focus is placed on ethical principles of European guidelines for an ethically compliant AI system and specifically on the principles of fairness and explicability for a trustworthy AI system.

In order to ethically examine the AI system of the use case, three analysis frameworks were combined and supplemented with a data analysis of the true positives of the clinical use case study. Using the frameworks, basic ethical requirements for a trustworthy AI system were subdivided with increasing levels of detail and their fulfilment was assessed. As a result, verifiable ethical aspects as well as ethical issues for refinement or further consideration were obtained. With the help of the data analysis based on this, it was examined which features influence the different recognition times of an OHCA by the machine learning (ML) model and to what extent fairness and explainability are given.

A full assessment of explicability of AI-supported decision-making cannot be made due to the black box model used and the lack of algorithmic transparency. Consequently, an assessment of fairness in relation to the generation of predictions is also not fully possible. However, the analysis shows that certain patients and caller characteristics shorten the OHCA detection time of the ML model. In particular, calls where the patient is of a higher age group or where the caller is medically educated show a faster detection of OHCA by the ML model. It is also noted that the ML model does not make predictions with a higher accuracy than human medical staff, but in a shorter time. The shorter OHCA detection time of the ML model is particularly noticeable for feature characteristics with a larger sample size. Furthermore, it can be observed that a complete acceptance of the medical staff is not given; only a few predictions of the ML model are taken into account in the assessment of the staff. This may be due to inadequate training of medical staff. The general acceptance of society needs to be further assessed due to limited transparency and communication.

Further research could address the enhancement of the AI system with additional functions or with more time-critical diseases.

Die Entwicklung von Systemen mit Künstlicher Intelligenz (KI) und deren Einsatz in unterschiedlichsten Gebieten nimmt immer mehr zu und wird in diversen aktuellen Diskussionen fokussiert. Dabei wird auch vermehrt auf ethische Aspekte eingegangen und KI-gestützte Prozesse aus ethischer Sicht kritisch beurteilt. Hierfür ist es notwendig, die Anforderungen an ein KI-System hinsichtlich der Konformität mit ethischen Werten und Normen zu definieren und je nach Anwendungsfall zu bewerten.

Das Ziel dieser Arbeit ist es, ein KI-System im Gesundheitsbereich zur Prozessunterstützung bei der Erkennung von OHCA (Out-of-Hospital Cardiac Arrest) auf ethische Problemstellungen hin zu prüfen. Da der vorliegende Anwendungsfall in Europa entwickelt und getestet wurde und somit europäische Werte und Normen respektiert werden sollten, wird der Fokus auf ethische Prinzipien europäischer Richtlinien an ein ethisch konformes KI-System und speziell auf die Prinzipien von Fairness und Erklärbarkeit an ein vertrauenswürdiges KI-System gelegt.

Um das KI-System des Anwendungsfalls ethisch zu untersuchen, wurden drei Analyse-Frameworks miteinander kombiniert und mit einer Datenanalyse der True Positives der klinischen Anwendungsfall-Studie ergänzt. Mit Hilfe der Frameworks wurden grundlegende ethische Anforderungen an ein vertrauenswürdiges KI-System mit steigendem Detaillierungsgrad untergliedert und deren Erfüllung beurteilt. Als Ergebnis lagen verifizierbare ethische Aspekte sowie ethische Problemstellungen zur Nachbearbeitung oder weiteren Betrachtung vor. Mit Hilfe der darauf aufbauende Datenanalyse wurde untersucht, welche Faktoren die unterschiedliche Erkennungszeit eines OHCA durch das Machine-Learning-Modell (ML-Modell) beeinflussen und inwieweit Fairness und Erklärbarkeit hierbei gegeben ist.

Eine vollumfängliche Beurteilung der Erklärbarkeit der KI gestützten Entscheidungsfindung kann aufgrund des eingesetzten Black-Box-Modells und der fehlenden algorithmischen Transparenz nicht erfolgen. Folglich ist auch eine Bewertung von Fairness im Bezug auf die Generierung von Vorhersagen nicht vollständig möglich. Die Analyse zeigt jedoch, dass bestimmte Patienten und Anrufermerkmale die OHCA-Erkennungszeit des ML-Modells verkürzen. Besonders Anrufe, bei welchen der Patient/die Patientin einer höheren Altersgruppe zuzuordnen ist oder bei welchen der Anrufer/die Anruferin eine medizinische Ausbildung absolviert hat, weisen eine schnellere Erkennung des OHCA durch das ML-Modell auf. Des Weiteren ist festzustellen, dass das ML-Modell Vorhersagen nicht mit einer höheren Genauigkeit als menschliches medizinischen Personal trifft, jedoch in einer kürzeren Zeit. Die kürzere OHCA-Erkennungszeit des ML-Modells ist besonders bei Merkmalsausprägungen mit größerem Stichprobenumfang festzustellen. Ferner kann beobachtet werden, dass eine gänzliche Akzeptanz des medizinischen Personals nicht gegeben ist; nur wenige Vorhersagen des ML-Modells werden bei der Beurteilung des Personals berücksichtigt. Dies kann auf eine unzureichende Schulung des medizinischen Personals zurückzuführen sein. Die allgemeine Akzeptanz der Gesellschaft ist aufgrund der eingeschränkten Transparenz und Kommunikation weiter zu prüfen.

Weiterführende Forschung könnte die Erweiterung des KI-Systems um zusätzliche Funktionen oder um weitere zeitkritische Erkrankungen behandeln.

Table of contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AED	Automated External Defibrillator
CRP	Cardiopulmonary Resuscitation
DNN	Deep Neural Network
FDR	False Discovery Rate
FNR	False Negative Rate
FOR	False Omission Rate
FPR	False Positive Rate
ML	Machine Learning
NLP	Natural Language Processing
NPV	Negative Predictive Value
OCT	Optical Coherence Tomography
OHCA	Out-of-Hospital Cardiac Arrest
PEA	Pulseless Electrical Activity (Assumption)
PPV	Positive Predictive Value
RCT	Randomized Clinical Trial
VF	Ventricular Fibrillation (Assumption)

1 Introduction

"We are headed towards a situation where AI is vastly smarter than humans. I think that time frame is less than five years from now": Elon Musk propagated this in mid-2020 in an interview with the New York Times [1]. A provocative statement, the accuracy of which will not be discussed here. However, this statement shows that artificial intelligence (AI) is becoming increasingly important and powerful as AI systems learn autonomously and can take over more and more human tasks.

Even for the general public, "AI" is no longer a new word: The media frequently report on AI systems [2], tech companies are growing [3] and future predictions are being made about the replacement of workers by AI systems [4]. The use and prominence of self-learning algorithms is growing, and AI systems are increasingly encountered in our daily lives [5]. Many AI systems are used by us on a daily routine, such as the translation program Deepl or the assistance systems Alexa and Siri [6]. Personalised advertising [7],[8] with which we are confronted via social media, for example, is also often AI-driven. In the meantime, AI systems are not only replacing such less responsible tasks, but are also making responsible and consequential decisions, such as deciding on a loan approval [9], hiring decisions [10] or medical diagnoses [11]. The increasing power of AI systems, which is also underlined by the citation of Elon Musk, gives rise to a critical, legal and ethical examination of AI and its decisions: In 2019, the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) published a guide for implementing trustworthy AI [12] and on 21/04/2021, a legal framework was presented by the EU, which is now in the legislative process [13]. At this point, it should be noted that the master's thesis assesses a use case that includes an AI system developed in Europe and therefore mainly uses the rights, values and norms of the Western world and Europe in particular as a basis. In addition to the legal anchoring of the consideration and observance of ethical aspects that must be adhered to in the development and implementation process of an AI system, the reputation and acceptance of AI systems by society also plays a role in the successful implementation of such self-learning algorithms. The non-trivial technology behind AI applications creates concerns and mistrust towards AI systems. In addition, people sometimes see AI systems as a danger that, for example, their jobs could be replaced and could develop in a harmful way. Furthermore, AI systems are distrusted because AI systems could possibly make discriminatory and unfair decisions. Therefore, the explainability, transparency and fairness of AI systems is of fundamental importance [14].

In the context of this master's thesis, an ethical assessment - especially of the ethical principles of fairness and explicability of the AI HLEG - of an AI system is carried out, which will be ethically evaluated holistically from 11/09/2020-30/03/2021 by an international team of experts under the direction of Prof. Zicari using Z-Inspection® (see Section 2.4.1). The system on which this assessment is based is an AI system that was developed for the healthcare sector - specifically to support the recognition of Out-of-Hospital cardiac arrest (OHCA) during emergency calls. AI systems are particularly helpful in healthcare industry, and they can possibly save lives through more accurate, faster and better predictions and diagnoses. They can also relieve healthcare staff or reduce costs in the healthcare sector [11]. However, in addition to the number and importance of applications, the need for fairness and explicability is also growing, especially in medicine, as it involves confidential, sensitive and personal data and issues. The next two examples show the relevance of fairness and explicability in the healthcare industry:

Fairness: Artificial intelligences in the medical field can be subject to biases. This can result in discrimination against certain groups, for example. It may predict a higher probability, for instance, of a more accurate diagnosis based on race, gender or age, even though there are no causal medical reasons for this (more on this in Section 2.2).

Explicability: Another important implication for AI models in the health sector is transparency and the associated explainability and interpretability of an AI model. The explainability of AI systems is theoretically underpinned in Section 2.3. The healthcare professional using the AI should be able to understand and interpret the decisions of the model. A patient who receives a certain diagnosis will want to know where the diagnosis comes from. If the decision was made on the basis of an AI, the doctor may not be able to explain the decision or understand it himself [15].

Before the theoretical foundation of the ethical principles, the presentation of the analysis methods and the use case with the subsequent evaluation of the use case are given, a quick introduction to AI is needed here. Because what is often used as a buzzword is actually much more differentiated: Artificial intelligence is not just artificial intelligence. Figure 1 shows the relationships between the terms briefly discussed below.

Artificial intelligence is an umbrella term for algorithms and AI technologies that are trained to perform tasks that actually require human intelligence. The software then imitates human thinking, such as logical reasoning, independent learning and the ability to remember things [15],[16]. *Machine learning* is a widely used form of AI technology. Machine learning algorithms are based on a statistical model that recognises patterns based on the input of training data, on the basis of which an output occurs. A subfield of machine learning is the *neural network* [11],[16]. This is a special form of machine learning and has its origins in neurology. It uses the neural nervous system of the human brain. It is represented by different layers of artificial neurons. Input is provided by training data, which then passes through these different layers. Each layer reacts to different features of the input and subsequently delivers an output [11],[16]. In a *deep neural network* (DNN, a special form of neural network), there are a number of different layers that are used to analyse the input data [11],[16].

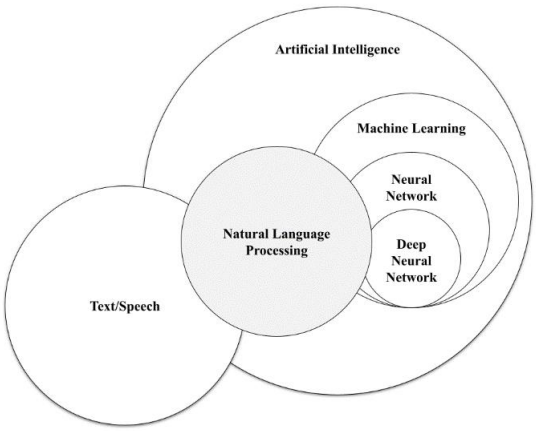


Figure 1: Connections between AI terms (Visualisation with content from [11],[16]).

Due to the complexity and the different layers of the neural networks, most models are *black boxes*. This means that the causality of the output of the systems by humans is no longer comprehensible. This has an impact on the explainability (see Section 2.3) of AI systems.

Natural Language Processing involves the handling with human language. ML technologies are often used for this purpose. As Figure 1 shows, AI technologies - especially ML systems - are tools for dealing with human language, for example classifying it, converting it into text or translating it.

In order to carry out the ethical assessment of the given use case, a theoretical basis is first created in Section 2. For this, we start by defining what constitutes a trustworthy AI system by describing and defining, apart from other things, the four ethical principles and the associated seven requirements for a trustworthy AI system of the AI HLEG (see Section 2.1). Subsequently, the ethical principles that are given special consideration in this master's thesis, fairness and explicability, are first defined in Section 2.2 and 2.3 (with focus on explainability) and their importance is elaborated in order to then explain whether and to what extent these principles can be achieved in an AI system. Section 2.4 presents the frameworks we used. As a superordinate assessment method, we partly use the Z-Inspection® by R. Zicari et. al [17], as this was applied in parallel to our master's thesis on a large scale by an international team of experts. Other analysis methods used within the Z-Inspection® process are the CAE framework (see Section 2.4.2) and the ALTAI evaluation list of AI HLEG (see Section 2.4.3).

The use case is presented in Section 3, starting with the relevance and possible methods for classifying AI systems in the healthcare sector (see Section 3.1). Subsequently, in Section 4.2, the AI system, the technology and the field of application of the underlying use case are briefly introduced and classified on the basis of the methods presented in Section 3.1. In Section 3.3 we will use parts of the Z-inspection® process to further present the use case, its application, the people involved and the stakeholders. In the following, in Section 3.4, two studies are presented in which the AI system has already been tested in different ways using real data and real calls.

In Section 4, the ethical analysis takes place with special consideration of the ethical principles of fairness and explicability of the AI system. In the process, the structure of the Z-inspection® process is used as a guideline: In Section 4.1, claims are made in relation to the use case, which are then substantiated or counter-evidence in the further course of the analysis and assessment. Section 4.2 then analyses the data available to us. Section 5 is used to evaluate the individual claims on fairness (see Section 5.1) and explicability (see Section 5.2) that emerged from Section 4.1, taking into account the data analysis. A conclusion follows in Section 6.

2 Trustworthy AI

As the number of AI systems developed increases, the question of the extent to which AI systems are trustworthy also arises more and more in the literature. More precisely, it is being questioned what an AI system must fulfil in order to be accepted by users and to maintain their trust in the system. AI systems often perform well in development and test environments, but are rarely accepted in practice. Therefore, it is of great importance to raise awareness of the need for trustworthy AI among all stakeholders that are involved in the development. In addition to research and innovation, the focus of developing new AI systems should thus always be on preventing harm and building trust with users. How a trustworthy AI system can be implemented is explained in the following Section using the Ethics Guidelines for Trustworthy AI [12]. Since the present use case was developed and tested in the European Union, we focus on this framework of the European Commission. First, we will describe the approach as a whole and then explain the two selected principles of fairness and explicability (with focus on explainability) in more detail.

2.1 Ethics guidelines for trustworthy AI

In order to establish and analyse the trustworthiness of an AI system, the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) recommends European States (as both the countries and the commission, are oriented towards European values, norms and rights) to follow the Ethics Guidelines for Trustworthy AI [12] developed by the group in 2019. The following Sections describe and link the fundamentals, principles and key requirements covered in this guideline starting with the three fundamentals of trustworthy AI.

2.1.1 Fundamentals of trustworthy AI

According to the guidelines, trustworthy AI consists of the three fundamentals *lawfulness*, *ethical alignment* and *robustness* (see Figure 2), which must be guaranteed at all times as primary goals.

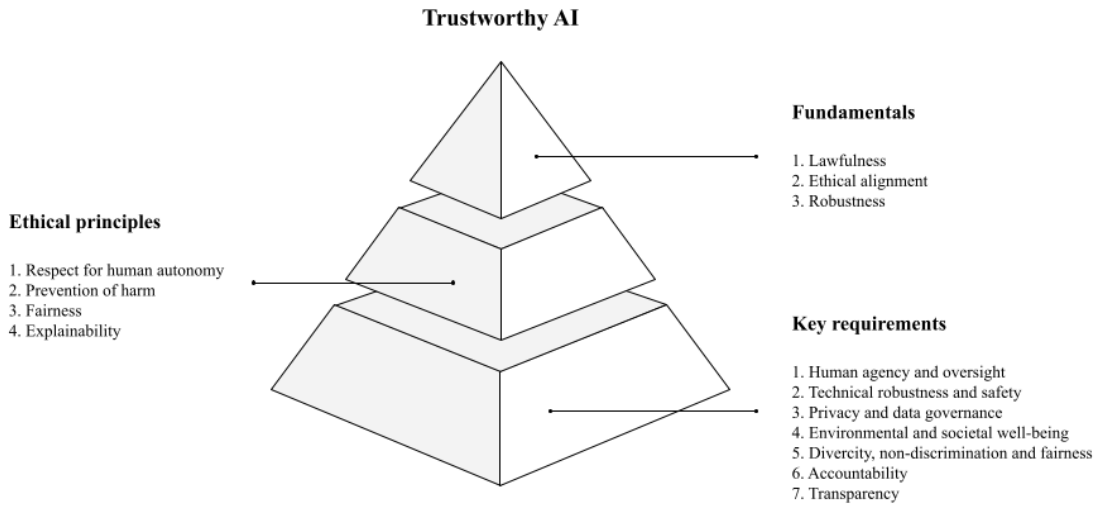


Figure 2: Abstraction levels of trustworthy AI according to AI HLEG [12].

Lawfulness: The AI system should comply with legal requirements.
Lawfulness is an absolutely fundamental requirement for an AI system. Decisions and functionalities of AI systems should be generated in a legally compliant manner. If they are not legal, the AI

system is under no circumstances trustworthy. Depending on the development and deployment region, the respective national or international legal regulations should thus be complied with and monitored. Specifically in the European Union (Union of States of the use case), the EU Charter of Fundamental Rights, the General Data Protection Regulation (GDPR), anti-discrimination directives, and other human rights and EU laws apply [12]. These may relate to general matters, but may also be specific to the use of electronic systems. For example, the GDPR [18] already establishes initial rules specifically for electronic data processing and automated decision making. In particular, Art. 13-15 contain information and disclosure obligations about the purposes of processing, categories, recipients and storage periods of personal data as well as about the existence of automated decision-making and its underlying logic and effects. Whether such automated decision-making may be used at all is defined in Art. 22 [18]. In addition to the before mentioned generally applicable regulations, the corresponding industry-specific regulations should also be observed. Particularly in the healthcare sector (the sector of the given use case), special regulations apply due to the sensitive, personal data, its processing, and the fundamental assurance of damage prevention. Compliance with these cross-industry and industry-specific legal regulations forms the basis for conformity with ethical principles, values, and standards [12].

Ethical alignment: The AI system should follow ethical values and standards.
A trustworthy AI system should not only be legally compliant, but also fulfil values and standards with regard to ethical principles. Compliance with legal requirements can already lead to a reflection on their underlying ethical foundations. Fundamental rights such as those for the preservation of human dignity, freedom and democracy, and those for equality and non-discrimination therefore already convey basic ethical values like physical and mental integrity, respect for self-determination, and fair treatment. AI systems should therefore technically enable respect for these values and norms by, for example, guaranteeing fair access to the system and its resources and avoiding unjustified outcomes. However, since laws are not adapted at the same speed as technical innovations and cannot always fully answer ethical questions, this step also requires an analysis of other ethical norms depending on the use case. Compliance with these ethical principles and norms is in turn a prerequisite for the ethical and robust functioning of AI systems [12].

Robustness: The AI system should be technically and socially robust.
In order to fully establish and maintain the trustworthiness of the AI system, it is necessary to secure the system on a technical and social level. AI systems should generate decisions reliably and securely and ensure damage prevention and low failure rates. This is especially important in healthcare (the industry of the present use case), as minimal security flaws and vulnerabilities can lead to great harm to patients. Unintended functioning or misbehaviour of the AI system that leads to potential harm can have a major negative impact on trust in the AI system. Specific security measures should therefore be considered and implemented in the early stages of development [12].

2.1.2 Ethical principles for trustworthy AI

In order to achieve these three fundamentals, ethical principles must be respected on the first level of abstraction including legal requirements. According to the AI HLEG [12], four ethical principles (see Table 1: *Respect for human autonomy*, *prevention of harm*, *fairness* and *explainability*) can be concretized derived from the two fundamentals of ethical alignment and robustness (the

legal component, together with fundamental rights, forms the basis and legal framework for all development steps). Possible tensions (see Section 2.4.1) that may arise as a result of the principles influencing each other should be resolved.

<table>
<tr><th>Ethical principles</th><th>Key requirements</th></tr>
<tr><td>1. Respect for human autonomy</td><td>1. Human agency and oversight</td></tr>
<tr><td rowspan="2">2. Prevention of harm</td><td>2. Technical robustness and safety</td></tr>
<tr><td>3. Privacy and data governance</td></tr>
<tr><td rowspan="2">3. Fairness</td><td>4. Environmental and societal well-being</td></tr>
<tr><td>5. Diversity, non-discrimination and fairness</td></tr>
<tr><td>4. Explicability</td><td>6. Accountability</td></tr>
<tr><td></td><td>7. Transparency</td></tr>
</table>

Table 1: Ethical principles and key requirements for trustworthy AI according to AI HLEG [12].

Respect for human autonomy: The AI system should preserve human freedom and autonomy.
An AI system should operate and be applicable in a way that does not deceive and manipulate end users, nor compromise their self-determination and personal identification. It should promote the capabilities of users and all stakeholders. The focus of development should therefore be human-centric, and human-machine interaction and the possibility of integrating human decision-making must be ensured. An AI system should also support or augment human work and for this reason also requires human monitoring of processes [12].

Prevention of harm: The AI system should avoid damage of any kind.
Since human interaction with the AI system is to be ensured, the avoidance of material and immaterial damage to users is of great importance. The fundamental rights also demand, as explained previously, e.g. the preservation of human dignity and physical and mental integrity. AI systems must therefore be technically robust and protected against misuse. Particularly vulnerable groups of people should be integrated into the development process and an unequal distribution of power and information (e.g. between companies and consumers) should be investigated in order to analyse possible impact [12].

Fairness: The AI system should ensure a fair operation and interaction.
All project phases (from design to usage) of the AI system should be fair (see Section 2.2). More specifically, fair distribution of resources, such as equal opportunities in the use of the AI system, and avoidance of discrimination and biased outcomes should be ensured. End users should not be deceived, forced, or otherwise restricted in any way. Therefore, tensions or conflicts (see Section 2.4.1) between ethical values that arise during development should be weighed or balanced. Furthermore, users and stakeholders should be able to contest decisions made by the AI system [12]. To achieve this, decision-making processes must be clearly identifiable and explainable.

Explicability: Decisions and results of the AI system must be explainable.
In order to maintain user trust even after the AI system has been deployed, all functions and decisions made by the system must be explainable (see Section 2.3). For this purpose, decision-making processes and their involved logic and actors should be defined and made transparent. In addition, it should be possible to communicate the capabilities, limits and purposes of the AI system. If it is a black box model and therefore only the externally perceivable behaviour can be explained, transparent communication or explainability techniques and measures are of great importance. The extent to which an explanation should be available depends on the respective context of use and on the gravity of the consequences of incorrect decisions [12].

2.1.3 Key requirements for trustworthy AI

The four ethical principles just described can in turn be broken down at a second level of abstraction into seven key requirements for a trustworthy AI system (see Table 1: *Human agency and oversight, technical robustness and safety, privacy and data governance, environmental and societal well-being, diversity, non-discrimination and fairness, accountability* and *transparency*). These requirements must be considered, implemented, and evaluated for fulfilment throughout the entire lifecycle [12].

Human agency and oversight: The AI system should allow human interactions and decisions without negative impairment.
Ethical principle: *Respect for human autonomy*
When using an AI system, users should be provided with information about how it works. They should be able to understand AI systems and make their own decisions. The self-determination of humans should be respected throughout all phases, which is why they must not be manipulated, deceived or pressured by the system [12]. As mentioned earlier, this principle of the right to decide and to be informed is also legally reinforced in Art. 22 GDPR [18]. Also, the generally applicable fundamental rights should be observed. Impacts on these need to be analysed at early stages of development and impact assessments should be prepared based on this. Human supervision should therefore be enabled for guidance and control [12].

Technical robustness and safety: The AI system should be reliable and robust.
Ethical principle: *Prevention of harm*
By respecting fundamental rights, physical and mental integrity should be ensured as the highest goal when using an AI system. Any harm should therefore be avoided. Risks should be assessed accordingly at all stages of development and supplemented with fallback plans (e.g., stopping an operation until it is human controlled). Risks of wrong decisions can be reduced by high precision in outcomes alone. Especially in the healthcare sector (the industry of the use case being analysed), this is of great importance, as wrong decisions can have a significant impact on human lives. Reliability and reproducibility of the AI system is important for high precision, in order to perform tests and compare results and decisions. The system should also be safe in other operational environments and be protected from negative attacks using security measures [12].

Privacy and data governance: The AI system should observe the protection goals of information security.
Ethical principle: *Prevention of harm*
In order to preserve fundamental rights and prevent damage, the three information security goals of confidentiality, integrity and availability should be observed. While confidentiality refers to the accessibility of data and information by exclusively authorized persons, integrity describes the completeness of correct data as well as the intended faultless functioning of the system. Availability describes the permanent accessibility of the AI system and its low downtime [19]. Data (input, processing and output data) must therefore be of a high quality and protection of these must be ensured in all development and usage phases. For example, this is a prerequisite for the principle of fairness; users trust that their data will not be manipulated or used for purposes other than those they have agreed to [12]. This must also be ensured on the basis of Art. 6 GDPR [18]. In particular, the quality of the data when it is collected is crucial in order to avoid data biases and system impairment by harmful data already at this point. Data processing methods and access rights should also always be documented [12].

Environmental and societal well-being: The AI system should take social and environmental responsibility.
Ethical principle: *Fairness, prevention of harm*
Since harm must be avoided and physical and mental integrity must be guaranteed, the impact of the use of the AI system on the user must be examined and evaluated throughout. The AI system should not affect social and environmental well-being as well as democracy in any way. Sustainable research and development should therefore be aimed for. Processes and supply chains must be reviewed with regard to sustainability, and their resource and energy consumption must be controlled, e.g., with the help of eco-balances [12].

Diversity, non-discrimination and fairness: The AI system should be developed for a large user group.
Ethical principle: *Fairness*
When using the AI system, equal access to the system should be ensured at all times. All stakeholders should therefore be involved in all phases from design to use. In combination with complete and fair data collection, non-discrimination can already be ensured during development (incl. algorithm). For this, requirements for the AI system must be clearly defined and communicated, and feedback rounds must be introduced. The goal should be to make a system accessible to all users without restriction, regardless of their age, gender or other personal characteristics. Measures such as the inclusion of accessibility in the design and development process and the provision of appropriate training should therefore be planned [12].

Accountability: The AI system should ensure accountabilities.
Ethical principle: *Fairness*
In addition to measures for taking ecological and social responsibility, measures to ensure accountability must also be guaranteed. Verifiability of decision and results should not only be possible, but also communicated. The AI system should be audited internally and externally, and the audit results and responsibilities should be made publicly available. In the event of possible negative results, a rapid response capability should be ensured. To prevent such negative impacts, early risk analysis and documentation (including impact assessments) could help. If any impact cannot

be assessed in time, legal protection should be in place. This also gives users confidence that legal protection can be ensured in the event of damage [12]. This principle also influences the principle of transparency.

Transparency: Processes of an AI system and their input and output should be documented.
Ethical principle: *Explicability*
Documentation of all processes, data sets and models is a prerequisite for the traceability of decisions and results of the AI system and the actors involved. If wrong decisions are made and damage is caused as a result, these must be verifiable and explainable. In addition to the documentation, the necessary understanding of the auditor or the person explaining is also important here. Protocols, explanations and their wording should be adapted to and focused on the respective target groups. Further, early communication about the use of the AI system should also be considered. The interaction with a system, its limits and functions, as well as the freedom to decide on its use must be communicated to the user in an understandable way [12].

Depending on the assessment method (see Section 2.4), the requirements explained can be concretized in further levels of abstraction and adapted to the respective system. This leads to a detailed evaluation, which reveals issues and flags (see Section 2.4.1) and can serve as a starting point for appropriate adjustments. Since this work deals specifically with a use case from the healthcare industry, such an analysis of the principles and key requirements is of high importance. The use of AI and therefore also automated decisions in the medical context can only be influenced and controlled by consumers and patients to a limited extent. For this reason, and also because of the general need for patient trust in medical devices and treatments, trust-based AI in the healthcare sector is essential. In Section 4.1, the respective levels of abstraction are applied to the present use case for evaluation and, together with the resulting issues and flags, form the basis of the subsequent analysis. For this purpose, the data taken from the Randomized Clinical Trial (RCT) [20] of the use case is used and analysed. In this master's thesis, the focus of the data analysis is on the two ethical principles of fairness and explainability (with focus on explainability), which are defined in more detail and evaluated in terms of their significance and achievability in the following Sections.

2.2 Fairness

The first focus of this master's thesis is on the third ethical principle fairness. As already described in Section 2.1.3, the three requirements of *Environmental and societal well-being*, *diversity, non-discrimination and fairness* and *Accountability* are assigned to this principle. To ensure fairness in an AI system, the aspects of fairness must be observed throughout the life cycle (from design to development to use) of the AI system. In this context, different dimensions can be declared as "fair": These include the access (Is equal access or use of the AI system ensured?), involvement of the different stakeholders, data collection and finally the decisions of the system [12]. This master's thesis mainly focuses on the fair data collection and fair decisions of the system. In Section 2.2.1, fairness is first defined in general terms and later in relation to computer science in order to create a uniform understanding of the term. Section 2.2.2 explains why fairness is important in AI systems (and in particular in the health sector), and Section 2.2.3 discusses the difficulty of achieving a completely fair AI system by describing various biases. Section 2.2.4 shows how the quality or fairness of an AI system can be measured and presents the confusion matrix with its

values and the ratios calculated from them. Finally, Section 2.2.5 explains how fairness or more fairness can be achieved by defining fairness metrics and mitigation algorithms.

2.2.1 How is fairness defined?

In the literature, there is no uniform definition of fairness. Fairness is dealt with in different domains - for example in game theory, law, economics, philosophical science, political science or social psychology [21].

The absence of any kind of prejudice, bias or favouritism towards an individual or group of people describes the concept of fairness in a very rudimentary but powerful way. This means that decisions can be considered fair if they disregard characteristics and attributes, such as age, gender, religion, skin colour etc. of groups or individuals that are irrelevant to the decision-making process [22],[23]. In social psychology, the term fairness can be defined by an individual's subjective perception and judgement of the actions of others. Individual assessments of such actions are based on one's own moral, social, religious, and legal values and thus differ from assessments of others. Whether one feels oneself to be treated fairly can therefore be assessed purely subjectively [24].

In addition to the different domains that affect fairness definitions, different cultures with their accompanying different values, norms and rights also matter. As we are Europeans and the given use case was developed and tested in Europe, fairness is guided by the values, norms and rights of the western world. According to the European Commission, fairness is anchored in the right to non-discrimination, solidarity and justice, compare Art. 20ff of the Charter of Fundamental Rights of the EU [25].

In computer science, the concept of fairness has only been dealt with for a few years [22]. The concept of fairness in this domain is becoming increasingly important, which is why the subject, and the concept of fairness are also being dealt with in computer science. The challenge is that in transferring fairness to informatics, the subjective perception of the individual [24] must be transformed into something measurable, because computers process information via machine commands based on numbers. Above all, it is about the conditions under which an algorithm is fair. Because an AI system represents an algorithm that then sometimes even independently makes consequential decisions about individuals or groups, the consideration and definition of fairness is indispensable. Although a large part of algorithms and thus also the problem of developing fair algorithms originate from the western world, there is no uniform definition in the literature for fairness in computer science [22]. For example, some decisions made by algorithms are considered fair according to some fairness definitions, whereas the same decision is called unfair according to another fairness definition [26]. Furthermore, it is mathematically impossible to fulfil some combinations of fairness definitions at the same time [23],[25],[28],[29]. A prominent example of such ambiguity is COMPAS. The software used by US courts assessed an offender's risk of recidivism. To assess whether this AI tool is fair, different institutions used different fairness definitions and fairness metrics (see Section 2.2.5) and thus arrived at different results: On the one hand, the AI system COMPAS should be racist towards black defendants, on the other hand, there should be no differences in the recidivism probabilities between whites and black defendants [22]. Some fairness metrics are explained in more detail in Section 2.2.5.

2.2.2 Why is fairness in AI systems important?

Fairness in computer science/AI in general: Alongside the increased use and development progress of AI systems, the relevance, importance, and prominence of AI systems has also been growing for several years. AI systems are no longer only responsible for low impact applications such as personalised advertising in social media or content recommendation [8], but now have far more responsible decisions to make. In many cases, AI systems are now replacing human decisions with high consequences that directly affect individuals and society [26]: credit rating decisions [9], autonomous driving, hiring decisions [10], recidivism probabilities of offenders and medical diagnoses are just a few examples of usages that can already be taken over by AI systems today [30].

Legal aspect: As described in Section 2.1, adherence to fairness and non-discrimination is rooted in international human rights, EU treaties and the EU Charter of Fundamental Rights and is thus legally binding.

Acceptance and reputation: For companies who are developing AI systems, however, in addition to the legal obligations, it is of major importance to design fair AI systems, as public pressure from society is growing. Without transparently addressing fair AI solutions, the company will not be trusted by society and will not receive acceptance. The company's reputation, and therefore its success, depends strongly on its acceptance in society [31].

Fairness AI domain specific (Healthcare): Trustworthy, responsible and as such fair AI systems are of great importance in healthcare (the sector of the use case to be analysed). Because even without the integration and use of AI systems, medicine and the associated decisions and predictions are vulnerable to bias and prejudice. It has been shown in some studies that even without the (potentially biased) algorithmic influence, healthcare presents both financial and medical biases and prejudices [32]. For example, in 1992, a publication revealed the grievance that much of the medical literature considers the white male as the norm [33]. Hence, medical decisions, research and predictions are subject to bias anyway, and the use of AI systems increases the risk of unfairness in the healthcare industry [32].

2.2.3 Why is complete fairness difficult to achieve?

Although fairness is a truly desirable fundamental value for society, the real world is undoubtedly not free of discrimination and unfairness [34]. But it is often assumed and presupposed that an AI system makes decisions objectively and free of bias [26]. However, due to the process of developing and training AI systems using data, both the training data and algorithms, as well as the humans involved in the development, can be biased and thus not enable the AI system to decide objectively [35]. Self-learning AI systems, for example, find patterns in real life data sets that implicitly reflect society's biases. Thus, biases and prejudices creep into the AI systems [36]. This fact inevitably leads to concerns about fairness and non-discrimination of AI systems [35], because an AI system that calculates unfair decisions, predictions and probabilities can harm the lives of individuals and society [34].

So what can fundamentally lead to unfairness in AI systems are the real-world biases that are mirrored in the training data, algorithms and use of the AI system [34],[37]. In the following, possible reasons for bias are listed using the categories bias in data, bias in algorithm and bias in user

interaction and then related with concrete examples to the healthcare sector (domain of the given use case).

Biases in data: Biases exist in the real world and therefore also in the data sets which are used to train AI systems. There are a number of types of biases in data, some popular ones and partially applicable to the given use case are listed and briefly described below and then related to healthcare.

Historical bias [38]: Historical bias is a distortion brought about by the current or past worldview. It is a bias of the real world and the current worldview with certain values and norms.
Healthcare [39]: The conditions for lung cancer screening used to be regular smoking and an age between 55 and 80 years. Although there are differences in smoking habits and thus health outcomes between different population groups, African Americans, for example, were underrepresented.

Representation Bias [38]: The bias occurs when a certain group is underrepresented.
Healthcare [38]: There is medical data for a certain disease from patients who are classified as sick. This neglects other patients who have the disease under different circumstances.

Measurement bias [38]: This bias is a bias in the measurement and collection of data. The bias exists when the processes, selection of characteristics and quality of measurements differ across the groups being studied.
Healthcare [40]: In predicting the condition of patients' care payments, differences can occur when different characteristics are used in different groups for analysis, for example, healthcare costs are used as a measure for one group and diseases for the other.

Aggregation bias [38]: This bias occurs when there are differences between groups being studied and conclusions are drawn from one group to the other.
Healthcare [38]: For patients with diabetes, there are differences in the Hba1C value (= provides information about the blood glucose concentration of a diabetes patient of the previous weeks) value between ethnic groups and gender. For example, the Hba1C value of a woman cannot be compared with that of a man.

Temporal bias [41]: Temporal bias exists when characteristics, behaviours or other attributes of a group or individuals change over time. Most data have such biases because observations are generalized over time.
Healthcare [42]: In the 1980s, evidence for Lyme disease (= a tick-borne bacterial infection) were arthritic symptoms, today it is known that the much more prominent sign of borreliosis is the bite itself and the ring-shaped rash around it.

Biases in the algorithm: Biases can also lie in the algorithm on its own. These biases do not arise from biased training data, but from the algorithm itself. As well as by the natural (and possibly implicit) biases of the developers and by the heterogeneous definitions of fairness. The matter of when an algorithm is fair - even with completely unbiased data - is not well defined. There are a set of biases in algorithm, three of which are popular in the literature and partly applicable to the given use case, will be briefly explained here [37]:

Evaluation bias [38]: This bias is about model evaluation. Such bias occurs as a result of the way the evaluation is done. The decision for a certain ratio or metric (see Section 2.2.4, 2.2.5)

by itself is a bias, because other metrics are neglected in this way. The AI system COMPAS described above, which was evaluated as fair or unfair according to different metrics, is a prominent example.

Healthcare: Especially in the healthcare industry, the individual choice of evaluation ratios and metrics is important: Whether an AI system that determines the need for care is evaluated according to cost savings rather than the number of patients makes a difference and represents a bias.

Emergent bias [43]: This bias arises from using the system with real users in a real usage environment after development. For example, new social knowledge may emerge that is not yet integrated into AI design.

Healthcare [43]: New medical knowledge (e.g., new findings on symptoms of AIDS patients) may have an impact on the treatment of certain diseases (e.g., AIDS), but this must be regularly and continuously integrated into the algorithm, to ensure that patients can also be identified with the new symptoms.

Sampling Bias [34]: This bias in the algorithm is caused by the sample. An AI system that is developed for a specific sample/population may not be transferable to another population.

Healthcare [44]: If an algorithm that is used to predict a certain disease (For example, cardiac arrest, in which women and men have different symptoms), is only developed for men, it cannot easily be used for women.

User interaction biases: Such biases arise from the interaction of the AI system with the user. In some AI systems, there is a user interface that plays a key role in what the user can see. Moreover, the user interface is the interaction between the AI system and the user. The user interacts with his or her own biases on the user interface. Three of those biases are briefly explained here:

Presentation bias [34],[37]: The user interacting with the AI system may already be biased by the user interface of the system. Subordinate presentation biases are the following biases: Position bias and content production bias.

Position bias [37]: This bias derives out of the design of the system's user interface. For example, western world people read from up to down and from left hand to right, which is why people pay special attention to the upper left part of the user interface [37].

Healthcare: Western medical professionals who are assisted by an AI system that provides alerts about a patient's health status will notice and read the alert better if it is displayed in the upper right corner [37].

Content production bias [41]: This bias arises from the structural, lexical, semantic, and syntactic presentation of information. For example, there are differences in expressions for the same content in different cultures.

Healthcare: An AI system's warning to healthcare professionals about a patient's condition may be understood differently/misunderstood by healthcare staff from different cultures.

2.2.4 How to measure fairness or unfairness?

In order to achieve fairness (see Section 2.2.5), it must first be possible to measure the fairness or unfairness of the system and for this purpose it must first be made measurable. To measure the accuracy and quality and thus the fairness or unfairness of an AI system, computer science often

uses the confusion matrix and its values (see Table 2), and then calculates ratios and afterwards fairness metrics (see Section 2.2.5) from these values [26],[45],[46].

The rows of the confusions matrix describe the values for the predicted results and the columns the actual results. Both values can be positive and negative [8]. Applied to healthcare (domain of the given use case) and the prediction and diagnosis of diseases, the True Positive case means that the AI system predicts a disease, and the patient really has this disease. False Positive is the case where the AI system diagnoses a disease even though the patient does not have that disease. True Negative means that the AI system does not diagnose a disease in case the patient does not actually have that disease. If the AI system does not recognise a disease in a patient, but the patient still has it, this is the False Negative case. This is the case where the most harm can be done [8], [24].

		Actual	
		Positive	**Negative**
Predicted	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 2: Confusion matrix [45],[46].

The following ratios can be calculated from the values of the different cases in the confusion matrix. In this Section, the ratios Sensitivity, Specificity, Positive predictive value, and Negative predictive value, which can be derived from the values of the confusion matrix, are described by way of example and applied to the healthcare and in specific to the prediction and diagnosis of diseases. These ratios are compared with each other in the course of the given use case (see Section 3.4). In general, it can be said that for these ratios a higher value is better and for the opposite ratios (not further explained here) False Negative Rate (FNR), False Positive Rate (FPR), False Detection Rate (FDR), False Omission Rate (FOR), a lower value is desirable. In the following, $\hat{Y}$ represents the predicted value and Y the actual value:

Sensitivity/True positive rate (TPR): The sensitivity describes the probability that the actual positive result is also the predicted positive result. In terms of disease prediction, sensitivity means the probability that a system predicts a disease when it actually exists. If the sensitivity is, for example, 70%, the system will correctly detect 70% of all patients with the disease (TP), but 30% of patients with the disease are not identified (FN). Sensitivity is calculated by TP / (TP+FN). Mathematically expressed $P(\hat{Y}=1|Y=1)$ represents the TPR.

Specificity/True negative rate (TNR): Specificity describes the probability that the actual negative result is also the predicted negative result. In terms of disease prediction, specificity means the probability that a system predicts freedom from disease when this actually exists. For example, with a specificity of 70%, the system correctly identifies 70% of patients without disease as disease-free (TN). However, the system then incorrectly identifies 30% of the patients without the disease being tested as sick (FP). Specificity is calculated by TN / (TN+FP). Mathematically expressed $P(\hat{Y}=0|Y=0)$ represents the TNR.

In principle, it can be said that a low sensitivity causes greater harm, because the failure to detect a disease that is actually present is more harmful than identifying a disease that is not actually present. Since sensitivity and specificity are oppositely proportional to each other and a system cannot achieve high specificity and high sensitivity at the same time, it must be weighed up, depending on the case, which key ratio should be particularly emphasised [27],[47].

In contrast to specificity and sensitivity, positive predictive value and negative predictive value are not based on the amount of actual disease, but on the predicted diseases:

Positive predictive value (PPV): The PPV describes the probability in the case that the predicted positive result is also the actual positive result. In healthcare, this means the probability for a patient actually has the disease if the test is positive. The PPV is calculated by TP / (TP + FP). Mathematically expressed $P(Y=1|\hat{Y}=1)$ represents the PPV.

Negative predictive value (NPV): The NPV describes the probability in the case that the predicted negative result is also the actual negative result. In healthcare, this means the probability for a patient actually not having the disease if the test is negative. The NPV is calculated by TN / (TN + FN). Mathematically expressed $P(Y=0|\hat{Y}=0)$ represents the NPV [26],[48].

2.2.5 How to achieve more fairness?

To achieve fairness in algorithms, various tools can be used, such as Google's What-If Tool [49], Fairlearn [50] or IBM's AI Fairness 360 [51]. These tools are designed for developers and data scientists to examine the ML models for fairness and even mitigate possible biases. They use various fairness metrics and mitigation algorithms, which are described in this chapter. However, as the tools are not used in this master's thesis, only the various metrics and mitigation algorithms are described in the Section as an example based on the AI Fairness 360 process, which includes the following steps [51]:

Step 1: Use metrics to check fairness and identify biases.
Step 2: Use algorithms to mitigate biases.
Step 3: Recalculate metrics and compare results.
(*Step 4* if bias still exists: Apply another algorithm.)

Step 1: In order to assess whether an AI system is fair and if there may be biases in the system, one can use different metrics which are based on the values and ratios of the confusion matrix. These serve to first identify possible biases in the system. The various metrics differ in terms of the target group of the metrics in literature [34]. There is a differentiation between individual and group fairness, where the objective is that different groups should be treated equally on average. Individual fairness on the other hand, aims to treat similar individuals similarly [21],[34].

Most fairness metrics make comparisons between the privileged and the unprivileged group. This can be the difference of the male and female group, the difference of young vs. old people or the difference between different ethnic groups. The ideal and therewith a fair algorithm is when the values of the two groups do not differ. However, there are ranges of differences that are still considered fair. The ranges are determined individually depending on the metric. The AI Fairness 360 tool, for example, rates all differences between -0.1 and 0.1 as fair for most metrics. Greater than

0.1 means an advantage for the unprivileged group and less than -0.1 means an advantage for the privileged group. In these cases, there is a bias, which should be mitigated in step 2 [51]. There are a number of fairness metrics in the literature that will not be discussed in detail in the course of this master's thesis, as they do not apply in the evaluation of the use case at hand, and it would go beyond the scope here. However, in order to give an insight into the various metrics, some metrics are listed below and briefly described in Table 3. The metrics Fairness through unawareness and Equal opportunity are explained mathematically by way of example and possible problems are shown.

Individual vs. group fairness	**Fairness metric**	**Requirement**
Individual fairness	Fairness through awareness	Requirement for a fair algorithm to deal with similar individuals in a similar way [21],[26],[34].
	Counterfactual fairness	Requirement for a fair algorithm to make the same decision for an individual regardless of whether they belong to the real or counterfactual demographic group [52].
	Fairness through unawareness	Requirement for a fair algorithm to make decisions without consideration of sensitive attributes [26], [52].
Group fairness	Demographic/Statistical parity	Requirement for a fair algorithm that without taking into account the group, the probability of obtaining a positive result is the same [26], [34],[52].
	Conditional statistical parity	Requirement for a fair algorithm that both the privileged and the unprivileged group have the same probability of being assigned to the positive result, taking into account certain factors [26],[34],[53].
	Equal opportunity	Requirement for a fair algorithm that both groups have the same TPR [34],[52],[54].
	Equal odds	Requirement for a fair algorithm that both groups have the same FPR and FNR [26],[34].

Table 3: Fairness metrics sorted by group and individual fairness, modified by [34].

In mathematical terms, the metric of Fairness through unawareness is expressed as follows: Sensitive attributes A, for example age or gender, are not taken into any consideration here [26],[52].

$$\hat{Y}: X \rightarrow Y, \text{ without consideration of A.}$$

Equal opportunity is expressed mathematically as follows, where A represents the sensitive attribute that separates the two groups under investigation [34],[52],[54]:

$$P(\hat{Y} = 1|\mathrm{A} = 0, \mathrm{Y} = 1) \approx \mathrm{P}(\hat{Y} = 1|\mathrm{A} = 1, \mathrm{Y} = 1).$$

With this fairness metrics (and others), however, biases can also occur due to proxy attributes: A proxy attribute is an attribute that is not declared as sensitive, but from which one can either infer such sensitive attributes or via which one can obtain other information (e.g., the ethnicity is not taken into account, but the place of residence instead) [29],[55].

It is not possible to satisfy multiple metrics simultaneously under normal circumstances, as [27] discusses and proves in detail. Using different metrics as a basis for assessing a fair algorithm is a

question of probability and not a question of science of the underlying domain. For example, it is not a question of the medical diagnosis of a particular disease which metrics are used, but a question of probability under different conditions. Therefore, it is of great importance to decide individually, depending on context, application, and domain, which is the most appropriate fairness metric. Furthermore, the decision should be continuously critically questioned and, if necessary, adapted to new requirements and findings [27].

Step 2: Using the metrics calculated in step 1, possible biases of the algorithm are found out, which can be reduced in this step by mitigation algorithms. According to [34], there are three ways to use mitigation algorithms: In pre-processing [34],[51],[56],[57],[58],[59] the training data is modified to eliminate biases and discrimination (see Section 2.2.3: Bias in Data). In-processing [51], [34],[60],[61] means that the learning algorithm is modified and changed if necessary (see Section 2.2.3: Bias in algorithm), and post-processing [34],[51],[54],[62] describes the possibility of mitigating the biases after training the model by mitigating the biases in the prediction or the decisions of the system, which is not integrated in the process of learning. This method is used especially for black box models.

The most common mitigation algorithms are briefly presented in Table 4:

Processing-category	**Mitigation Algorithm**	**Procedure**
Pre-Processing (Mitigation in Data)	Optimized pre-processing	The data is cleared of discrimination and biases in advance [51],[57].
	Reweighing	Adjustment of the weighting of the sample [58].
	Data pre-processing	Data pre-processing processes, e.g., an algorithm that suppresses sensitive attributes [59].
	Word Embedding	E.g., neutralisation of biased words [34],[56],[61].
In-Processing (Mitigation in Classifier)	Adversarial Debiasing	Uses a classifier that minimises the possibility of finding out the sensitive attribute [51].
	Prejudice Remover	Adding a discriminatory regulator [51].
	Meta Fair Classifier	Algorithms acting as input and giving back an optimised classifier [51].
Post-Processing (Mitigation in Predictions)	Reject Option Classification	Delivers better results for the unprivileged group and worse results for the privileged group [51].
	Calibrated Equalized Odds Post-processing	Optimisation through classifier score output to generate fair output [51].

Table 4: Mitigation algorithms sorted by processing-category, modified by [34].

Step 3: After applying the mitigation algorithms, the metrics are checked again, and the results should be compared.

Step 4: If the results do not differ significantly and the bias to be mitigated still exists, a different mitigation algorithm should be used [51].

2.3 Explainability

The second focus of this work is on the ethical principle of explicability. As already described in Section 2.1, the AI HLEG [12] particularly assigns the key requirement of *transparency* to this principle. In this context, the requirement of transparency includes traceability, explainability, and communication. Since explainability also acts as a link between traceability and communication, this Section focuses on the explainability approach. The extent to which Explainable AI (XAI) and the underlying conditions are important for building trust and how these can be achieved and ensured is explained as follows. In order to create a uniform understanding of the term explainability, it is defined at the beginning.

2.3.1 How is explainability defined?

Basic theories: The reasons and logics for requesting and giving explanations have already been discussed in a wide variety of theories and models from the fields of philosophy and psychology. For example, Aristotle's Doctrine of the four causes from the books Physics and Metaphysics describe four reasons for an explanation. According to Aristotle, knowledge about a thing cannot be acquired until its causes are explored. In this process, the exploration for the reason is done by answering questions about why; answers to these questions are (usually) explanations [63]. Accordingly, explainability enables answering why questions and based on that, building knowledge.

The legal framework: The general legal necessity and responsibility of explainability is based, among other things, on Art. 47 of the Charter of Fundamental Rights of the European Union [64]. All individuals whose fundamental rights have been violated have the right to accuse the violation. This right requires explainable processes by the defendant, which led to the respective decisions. In addition to the Charter of Fundamental Rights, the GDPR [18] also imposes legal conditions on explainability. According to Art. 12-15 GDPR, organizations are subject to both the obligation to inform and communicate and the obligation to provide information upon request. For example, data processing purposes, the storage duration or the inclusion of automated decisions must be communicated to the data subject in a precise, comprehensible, and easily accessible manner. These obligations require transparent and explainable processes and documentation.

XAI in general: As mentioned earlier, explainability is also defined by the AI HLEG [12] as a prerequisite for transparency and thus for trustworthy AI systems. Explainability requires traceability and enables communication. To ensure that the outcomes of AI systems are traceable, data, processes for their acquisition and processing (incl. algorithms) as well as results and decisions should be documented. In this way, traceability, verifiability (e.g., in the case of wrong decisions) and explainability can be ensured. Both technical and human processes and decisions when using an AI system should be explainable. In order to enable this, technical understanding and the comprehensibility of technical processes are essential. In addition, it should also be possible to justify decisions for the use of the AI system, its design, and areas of application. But not only the use and interaction of and with an AI system should be justified to users, but also openly communicated. This includes functions, benefits, and limitations. The user should also have the possibility to reject the interaction with the AI system [12]. In addition, Doshi-Velez and Kim [65] emphasize interpretability in the context of explainability of AI systems and define it as the "[...] ability to explain or to present in understandable terms to a human" [65]. Thus, to make AI systems explainable, data, processes, decisions, and results must be both traceable and interpretable.

XAI Domain-specific (Healthcare): Especially in the healthcare sector (the sector of the use case to be analysed), it is essential to explain the logic involved, since it can have a direct impact on human health [66]. According to Amann et al. [67] a multidisciplinary perspective of XAI in the healthcare sector is therefore of great importance. In addition to the already mentioned legal, ethical, technical and social aspects of explainability, medical facts must also be explained. These in turn interact with other perspectives (e.g., according to Art. 9 GDPR [18] special legal regulations must be observed when processing personal health data). On the one hand, explanations can be addressed to medical staff to help understand AI decisions and to make diagnoses based on them and under the influence of their own experiences [67]. On the other hand, potentials, and risks of using the AI system can be communicated to patients and decisions can be justified [68]. Therefore, for the medical staff it is important to have a basic understanding of all perspectives as well as the ability to communicate explanations in an understandable and appropriate manner [67].

Traceability and interpretability are therefore particularly important to ensure an XAI. Depending on the field of application, it should be possible to explain domain-specific issues in addition to legal, ethical, technical, and social perspectives. In Sections 4 and 5, we refer to explicability as an overall principle for trustworthy AI systems according to the AI HLEG. The different types of explanations are described in the following Section.

2.3.2 What types of explanations exist?

Two types of explainability are to be distinguished: Global explainability and local explainability. Doshi-Velez and Kim [65] define global interpretability or explainability as the knowledge and understanding of the ML model as a whole, while local interpretability or explainability refers to a specific input and the resulting decisions of the ML model. In addition to distinguishing between global and local explainability, explanations should be adapted to other criteria and to the respective recipient of the explanation. For example, the scope of the declaration could depend on the recipient's available time and previous knowledge or domain. While laymen might expect a compact explanation in easy language, domain experts might expect a detailed explanation for which they are willing to spend a lot of time. Depending on the domain of the recipient, the explanation could also focus on, for example, ethical aspects or technical details [65]. Therefore, the explanations and the information contained differ depending on various influencing factors. In addition, they can also differ according to the type of explanation. According to Lipton [69], two categories of explanations exist: Explanations that describe the ML model based on a given transparency and so-called post-hoc explanations that provide additional information. Explanations based on transparency can be divided into three levels according to their components: Simulatability (entire model), decomposability (components and parameters), and algorithmic transparency (training algorithm) [69].

Simulatability: People should be able to understand and view the ML model as a whole. All inputs and calculations leading to a prediction should therefore be documented, accessible, traceable, and repeatable [69].

Decomposability: Components and parameters can be designed (e.g., by naming them) in such a way that they can be directly interpreted and explained [69].

Algorithmic transparency: The error surface should be understandable, and it should be possible to justify that the training model provides an unambiguous prediction, even with new data [69].

However, explanations based on transparency are not always possible, especially with black box models. While white box models, such as linear models or decision trees, are usually interpretable from the ground up [70], black box models require further explanatory methods (see Section 2.3.5). Deep learning models are often too complex and opaque, so that comprehensibility is impeded and behaviours cannot be explained. In order to still be able to interpret the model, so-called Surrogate Models are added. These are interpretable models like mentioned before (e.g., decision trees) applied to black box models to make results and behaviours explainable in retrospect [71]. In addition to explanations based on transparency, post-hoc explanations can thus serve as additional information for end users after the model has been trained. These can be designed in different forms: Text explanations, visualizations, local explanations, explanations by example [69].

Text explanations: Decisions and results could be provided as text explanations. For example, while the actual ML model makes the predictions, a second model could evaluate it and generate corresponding explanations in text form [69] (e.g., Surrogate Models).

Visualizations: Explanations can also be generated in the form of visualization. Predictions can be mapped visually by changing the input values and the change can be evaluated qualitatively [69].

Local explanations: Since it is not always possible to make global explanations comprehensible and clear, local explanations can be used to justify a specific prediction as a section of the whole system [69].

Explanations by example: Explanations can also be made by describing similar models. The determination of the most comparable model is done via methods, such as k-nearest neighbour and their computation of distance metrics. Users can assess the match via their own expertise and apply explanations to the initial model [69],[72].

Explanations should thus be adapted in terms of language and scope to the respective stakeholder and intended purpose. In addition, different forms and types need to be analysed according to the intention of the explanation. The following Section describes why explanations are of great importance in the first place.

2.3.3 Why is explainability in AI systems important and what are the risks?

"Truly trustworthy AI requires explainable AI [...]." [73]. This quote from the Select Committee on Artificial Intelligence of the US National Science and Technology Council [73] highlights the importance of explainability for the trustworthiness of AI systems. In addition to compliance with legal regulations, as described in Section 2.3.1, trust building is one of the most important objectives that can be achieved by explainability and is according to Lipton [69] one of five main objectives of interpretability and explainability of ML models: *Trust*, *causality*, *transferability*, *informativeness*, and *fair and ethical decision making*.

Trust: According to the AI HLEG [12], explicability with the sub-condition transparency and in turn explainability forms one of the four ethical principles (see Section 2.1), compliance with which

enables a trustworthy AI system. Here, Bhatt et al. [74] define trustworthiness in the context of an AI system as the extent to which the outputs of the ML model can be trusted. According to Lipton [69], this trust is based, among other things, on the user's understanding of the ML model. Accordingly, people give decisions to the system in their own interest if they are understood and consistent with their own needs and courses of action. Furthermore, trust in the system grows, as long as the AI system always makes correct predictions when the human also makes similar assessments. If the system makes incorrect decisions when the human's correct predictions deviate, intervention or monitoring by the human is necessary [69]. Especially in the healthcare domain, where physical and mental integrity is a central goal to aim for, both patients and medical staff need to be able to trust ML decisions. Medical professionals often want to focus on human needs rather than technological aspects or lack the appropriate prior technological knowledge [75]. The extent to which system and user decisions agree can be verified with explanations.

Causality: According to Lipton [69], another goal that people seek with the help of explainability is to understand causal relationships. However, due to unobserved causes, this is not always possible. Nevertheless, interpretability and explainability help to formulate hypotheses that can subsequently be further tested [69]. In the healthcare industry, this takes effect due to the treatment of patients. Because of the fact that the human body can be seen as a black box and causes and diagnoses cannot always be fully justified, finding causal relationships is usually not possible [75]. The more important it is to be able to understand and question the underlying reasoning behind ML decisions.

Transferability: Usually, the training and test data for an ML model comes from the same distribution, which is why there is often a high generalization[1] error when new data from a different distribution is entered [69]. In contrast, according to Lipton [69], humans can apply acquired knowledge better in new environments. Possible reasons for poor performance can therefore be analysed via a given explainability [77]. Feature allocations (see Section 2.3.5) and their deviations from training allocations can be observed and errors can be eliminated [74]. However, this can also lead to attacks with the help of Adversarial Examples, especially in DNN, where malicious, manipulated input is intended to force wrong decisions of the ML model [78]. In healthcare, a medical review of the AI system is therefore a precondition for its use. Due to the heterogeneous clinical environment, the performance, accuracy of predictions, and frequency of errors must be medically validated before the system can be used. Even after such validation, collaboration between the system and the medical professional is important. ML decisions can be used, e.g., through explanations, as an assistance for diagnoses by the medical staff [67].

Informativeness: Explanations serve two ways of providing information. On the one hand, explanations can support people in the decision-making process [69]. As explained in Section 2.3.2, information differs according to context, stakeholder, and scope, for example with regard to local or global explainability. According to Bhatt et al. [74], explainability also increases external transparency. Decisions and results of the AI system can thus be communicated to different stakeholders and their questions and complaints can be answered [74]. On the other hand, explainability should be given to create regulatory compliance. It must be possible to check compliance with

[1] The generalization error, i.e., the performance difference between training and test data, describes how well the model performs on new data [69],[76].

legal requirements when using the AI system in the course of internal audits. For this purpose, it must be possible to explain ML processes and predictions to the corresponding risk and legal departments [74],[77]. Building on the legal provisions of the GDPR [18], which define obligations to provide information (see Section 2.3.1), end users must also be given the opportunity to contest decisions made by the AI system about the information and explanations provided. In this regard, Ploug and Holm [79] define four dimensions of contestability specifically for AI-assisted diagnoses in the healthcare domain, which require explanations of the information incorporated into the diagnostic process: Data, bias, diagnostic performance and decision. Patients should be provided with information about personal health data, such as its different uses and data sources, about possible biases in the ML model or data, and about model testing. In addition, information about the performance of the model, the variables used, and alternative diagnoses and underlying changed conditions should be given to the patients. They should have the opportunity to be informed about the division of labor between the AI system and the medical staff, and their responsibilities in the diagnostic process. In the event of patients contesting diagnoses, organizations should be able to provide all of this information and ensure explainability.

Fair and ethical decision making: However, the analysis of ML models and data with regard to biases and discriminations should not only be possible retrospectively from the outside, but should also take place proactively by the organizations themselves. It must be ensured that the AI system delivers fair decisions. The extent to which these are subject to discriminatory influencing factors can often not be identified with the classic ML metrics, such as accuracy. Thus, in order to verify fairness, explainability is usually also necessary [69],[77]. This is especially important in the healthcare sector. According to Ploug and Holm [79], it should be possible to clarify the fairness of AI-assisted diagnoses to the patient. It should be analysed whether deviating diagnoses are actually due to relevant differences of patients or whether they are based on errors or biases.

Interpretability and explainability may bring risks or undesirable consequences as mentioned before. Physicians should, even if not always desired, acquire a basic technical understanding in order to be able to understand technical explanations. The expectation of an explanation to show causal relationships should furthermore be avoided. In addition, explanations also entail the risk of Adversarial Attacks. However, they serve a precondition for building and maintaining user trust, recognizing relationships, and making analyses based on them, providing stakeholders with relevant information, and ensuring fair decisions and results. Ploug and Holm [79] highlight here that explanations are not always and not fully requested. However, companies and organizations need to be prepared and trained to respond to questions, objections, or audits. The steps for preparing and achieving explainability are outlined in the following Sections.

2.3.4 How to achieve explainability?

In the following, the actions to achieve explainability are described first using a general process, then including explainability techniques and metrics, and finally including evaluation tools (see Section 2.3.5).

Explainability should already be taken into account in the design and development process. Bhatt et al. [74] thereby particularly emphasizes the stakeholders and their expectations of the explanation. They recommend three steps for achieving explainability:

Step 1: Identify the target group.
Step 2: Analyse the needs of the target group.
Step 3: Distinguish the benefit of the explanations for the target group.

Step 1: Initially, the respective target group of the explanations has to be identified, i.e. the respective groups of people who expect or demand an explanation of decisions and functions of the ML model [74]. In the healthcare sector, for example, this could be the patients or even the clinicians.

Step 2: The next step is to analyse the various needs of the stakeholders. Here, the type and form of the explanation as well as the desired content and scope (see Section 2.3.2) must be particularly questioned [74]. While clinicians might use explanations to verify their own diagnoses or to derive diagnoses based on the explanation (high level of detail, medical expertise required), patients might expect a summary explanation that is easy to understand.

Step 3: After the target groups and their expectations and requirements have been analysed, the final step is to distinguish the benefit of the explanations for the respective target group. Explanations can either serve as a one-time review or justification, or they can be used in feedback rounds to discuss improvements and adjustments with stakeholders on an ongoing basis [74]. The importance of a multi-stakeholder development process in Clinical Decision Support is also emphasized by Sendak et al. [75] in terms of explainability. Clinicians should not only receive one-time explanations, but should be involved in every step of the process and be allowed, for example, to co-determine parameters, data categories or requirements for the evaluation and user interface design. They provide relevant information and aspects about the diagnostic process in general, which should be understood by the developers and implemented in the ML model. With the help of regular feedback rounds, issues can be discussed and further information can be exchanged in this way. This makes it possible for all those involved to be able to explain the process and, building on this, to communicate functions and limits externally [75]. In addition, these feedback rounds can be used to prepare explanations, e.g., to find answers in advance to the questions mentioned in Section 2.3.3 regarding data, biases, diagnostic performance, and decision in the event of a challenge [79].

Accordingly, through early requirements analysis and involvement of all stakeholders in the design and development process, both traceability and interpretability can be created, enabling explainability as described in Section 2.3.1. Next, different kinds of techniques and metrics are described to generate explanations.

2.3.5 How to generate explanations with the help of techniques and metrics?

The respective type of explanations (step 2), can be created with the addition of various explainability techniques and metrics. In this Section, selected *feature-based* and *example/sample-based* techniques as well as evaluation metrics are presented.

Feature-based techniques: To test the explainability of an ML model, the influence of the features on the predictions of the ML model is compared and thus their importance is analysed. The category of feature-based techniques includes for example Partial Dependence Plots (PDPs), Individual

Conditional Expectation (ICE), Accumulated Local Effects (ALE), Global Surrogate Models (see step 2), Local Interpretable Model Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) [80], some of which (like in Section 5.2.1 Surrogate Models and SHAP) can be combined [81]. Among these, the two most commonly mentioned techniques are LIME and SHAP. While LIME provides local explanations by manipulating the input data of the training set, comparing the change in the output, and allowing conclusions about the meaning of the features, SHAP provides both local and global explanations. Here, the SHAP value of a feature describes the change in the prediction of the ML model should the feature be considered and thus the contribution to the deviation of the predicted value from the base rate. The base rate is calculated as the mean of the predictions of the ML model. In this way, positive or negative influences of features on the prediction can be analysed [80],[81]. Since we apply SHAP values in the data analysis (see Section 4.2.2 and 5.2.1), this technique is now described with an auxiliary example. More precisely, a local explanation using SHAP values is now explained exemplarily with the addition of Figure 3 and [82]. The base rate of 0.1 symbolizes the total (mean) expectation $E[f(X)]$ of the model output for the predicted feature (such as the risk for a disease) and the selected feature subset. The model output for the local prediction for the shown feature values $f(x)$ is instead 0.4. Starting at the base rate $E[f(X)] = 0.1$, each time a new feature with the corresponding feature value is added (from bottom to top), the specific impact of that feature on the model output the local prediction is shown. For example, adding the feature "BMI" with a value of 40 changes the model output of the local prediction to $f(x) = 0.2\ (+0.1)$. Adding the feature "BP" with a value of 180 then changes the model output of the local prediction to $f(x) = 0.3\ (+0.1)$. Adding the feature "Sex" with a value of "F" decreases the model output of the local prediction by 0.3 to $f(x) = 0$. Finally, the feature "Age" with a value of 65 changes the model output of the local prediction to $f(x) = 0.4\ (+0.4)$, which is also the final model output of the local prediction for this subset of features. For a global explanation using SHAP values, see Section 5.2.1.

Figure 3: Exemplary explanation with SHAP values for each feature, image detail from [83].

Example-based techniques: The explainability of the ML model can also be established by analyzing instances and their influence on the output of the ML model. Example-based techniques include Anchors, Counterfactuals, Counterfactuals guided by Prototypes, Contrastive Explanation Method (CEM), Kernel SHAP, Tree SHAP and Integrated Gradients. For example, Counterfactual Explanations show the impact of changes in feature values on prediction and can explain black box models. This can show what would need to be changed, i.e., what value a variable would need to take in order to achieve an intended prediction. Another example is CEM. The Contrastive Explanation Method provides local explanations of black box classification models. It analyses relevant positive features of an instance, which must be present/fulfilled, and relevant negative features of this instance, which must not be present/fulfilled, in order to remain assigned to the predicted class [80].

Evaluation metrics: Metrics can be used to evaluate the explanations provided by the techniques with respect to the quality of the explanation, but also with respect to the usefulness for the user. According to Bhatt et al. [74], the most commonly used metrics to evaluate the quality are Faithfulness, Completeness, Sensitivity, I2 norm and normalized I1 norm. Faithfulness examines e.g., the feature importances generated with the help of an explainability technique for their correctness (are the features really as relevant as stated) [84]. For this purpose, the correlations between feature importances and the prediction probabilities when features change are computed, allowing the correctness of the feature importances to be assessed [84]. In order to be able to assess explanations with respect to their usefulness for users, Holzinger et al. [85] recommend the System Causability Scale. Users can answer, for example, whether the explanation was understandable, whether additional help was needed, whether the level of detail could be adapted, or whether the explanation was aligned with the user's existing knowledge. After answering, it can thus be analysed to what extent the explanation is sufficient and comprehensible for the user.

These techniques and metrics can help make the AI system traceable and explainable, and consequently enable open and user-centric communication. Software tools and toolkits can be used to apply these techniques. The ML model and data can become more transparent with the help of algorithms and metrics, and the requirement of transparency for explicability (see Section 2.3.1) can thus be fulfilled. In the following Section, two selected popular toolkits for analyzing the explainability of AI systems, *AI Explainability 360* and *Interpret ML*, are presented.

AI Explainability 360: The Python-based open-source toolkit developed by IBM includes algorithms and metrics to help understand ML model predictions using influencing features [86]. In doing so, data and ML models can be analysed in terms of their interpretability and explainability [87]. The explainability algorithms supported by AI Explainability 360 (see Table 5) target five types of explanations with different forms (tabular, image, text): Explanation of the data, local directly interpretable explanations, local post-hoc explanations, global directly interpretable explanations, and global post-hoc explanations [88],[89],[90].

Type of explanation	**Explainability algorithms**	**Form of explanation**
Data Explanation	ProtoDash [91] Disentangled Inferred Prior VAE [92],[93]	Tabular, image, text Image
Local direct interpretable explanation	Teaching AI to Explain its Decision [94]	Tabular, image, text
Local post-hoc explanation	ProtoDash [91] Contrastive Explanations Method [95] Contrastive Explanations Method with Monotonic Attribute Functions [96] LIME [97] SHAP [98],[99],[100]	Tabular, image, text Tabular, image Tabular, image Tabular, image, text Tabular, image, text
Global direct interpretable explanation	Boolean Decision Rules via Column Generation (Light Edition) [101] Generalized Linear Rule Models [102]	Tabular Tabular
Global post-hoc explanation	ProfWeight [103]	Tabular, image, text

Table 5: Types of explanations and supported algorithms by AI Explainability 360 [88],[89], [90],[104].

While local and global explanations, as explained in Section 2.3.2, differ with respect to a single outcome or the model as a whole, models can also, as described earlier, either be directly understandable (e.g. directly interpretable through model transparency) or require subsequent explanations (post-hoc) for understanding [89]. In addition to the previously mentioned algorithms, two metrics are supported to measure the quality of explanations [86]. Faithfulness examines, as already explained, feature importances for their correctness [84]. Monotonicity, on the other hand, measures the monotonic relationship between features and predictions. When including new significant features (sorted in ascending order of their computed importance), the probability of correct predictions should also increase [105],[106].

Interpret ML: The second explainability toolkit presented in this master thesis is the open-source Python package Interpet ML developed by Microsoft. It supports nine algorithms for interpretability and explainability of ML models and distinguishes them into white box and black box algorithms (see Section 2.3.2) respectively, as shown in Table 6. The explanations of the different algorithms can be compared directly in the visualization platform or extracted in the form of data [77],[107].

Type of ML Model	**Explainability algorithms**
White box model	Explainable Boosting [108],[109],[110],[112],[113],[114],[115],[119] Decision Tree [116],[117],[118],[119] Decision Rule List [116],[119] Linear/Logistic Regression [116],[117]
Black box model	SHAP Kernel Explainer [98],[99],[100] SHAP Tree Explainer [98],[99],[100] LIME [97] Morris Sensitivity Analysis [120],[121] Partial Dependence [122]

Table 6: Supported explainability algorithms per model type by InterpretML [77].

Explainable Boosting Machines can be used, for example, to visualize the influence of individual features in the prediction. Either global explanations about the ML model itself or local explanations about specific predictions can be generated [77],[107].

Explainability techniques, algorithms and metrics can be used to achieve, check and monitor explainability. In addition to the tools mentioned above, theoretical evaluation methods and frameworks can also be applied to evaluate ethical aspects of an ML model. The evaluation of the use case takes place with the help of such methods and frameworks and is complemented with data analysis. The focus will thereby be on Z-Inspection®, CAE and ALTAI. These are first explained in the following Section and applied to the use case in terms of fairness and explainability analysis in Section 4.2.2.

2.4 Assessment Methods

There are different methods to evaluate an AI for the different aspects of trustworthiness holistically (not only the algorithm and the system itself, like the tools already mentioned in Section 2.2.5 and 2.3.5). For the evaluation of the given use case, Z-Inspection® (see Section 2.4.1) is applied as the overarching method. Furthermore, the CAE framework (see Section 2.4.2) and ALTAI (see Section 2.4.3) are used, which are presented in the following Sections.

2.4.1 Z-Inspection® as a (comprehensive) method

The Z-Inspection® process developed by Prof. Zicari et al. [17] represents an analytical process to evaluate AI systems holistically on the background of ethical implications. Z-Inspection® is not domain-specific and can hence be applied in different areas to assess AI systems from, for example, business, the public sector, or healthcare. It is the first scientific approach to assess the trustworthiness of real-world AI systems. In 2020, Z-Inspection® was used to analyse and evaluate the ethical and societal consequences of using AI to predict cardiovascular heart disease. Further evaluations using Z-Inspection® are in progress, such as the use case "Machine Learning as a supportive Tool to recognise Cardiac Arrest in emergency calls", which forms the basis of this master's thesis and is introduced in Section 3.2 and partly explained following the Z-Inspection® process described in [17].

In the following, the Z-Inspection® analysis process is briefly described on the basis of Figure 4: Basically, the Z-Inspection® analysis and evaluation process is structured in three phases: The Set-Up-Phase, the Access-Phase and the Resolve-Phase. A protocol is kept throughout the entire process.

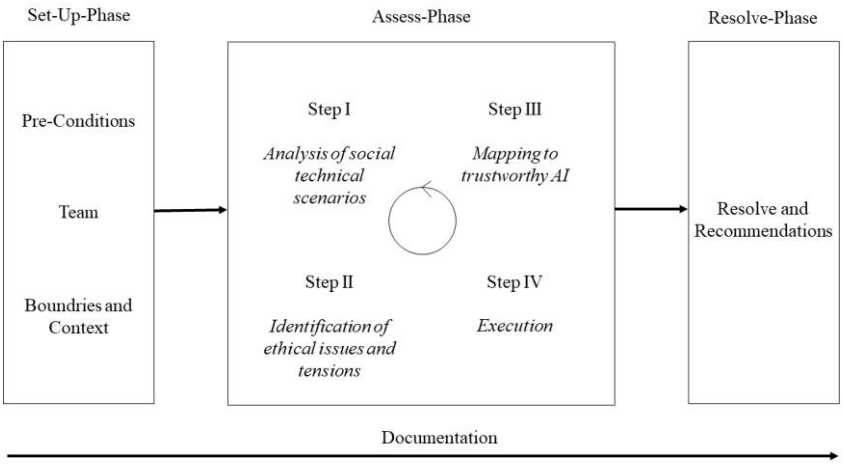


Figure 4: Z-Inspection® analysis process (modified from [17]).

Set-Up-Phase: First, some pre-conditions need to be checked. Eight introductory questions make it easier to identify the pre-conditions. For example, they ask who is requesting the inspection, why, and for whom it is relevant. Further, it should be clarified in advance whether and to what extent there are conflicts of interest. It should be ensured that there is neither a conflict of interest between the evaluators and the organisation or between the evaluators and the development. It is also of particular importance that the team carrying out the assessment and analysis is free of bias and as heterogeneous, dynamic, and multidisciplinary as possible. Furthermore, the context and its boundaries are defined in the Set-Up-Phase. In this context, the question is asked, in which ecosystem the AI system is to be implemented. After all, an AI system does not exist without interaction with society. It is important to consider human rights, concentrations of power and the social system.

Assess-Phase: The Assess-Phase is the central part of Z-Inspection®. Here the assessment happens and consists of four steps. The Assess-Phase is an iterative process that may be repeated until full and holistic feedback on the content of this phase can be achieved (see Figure 4).

Step I: Analysis of social technical scenarios:
In this step, socio-technical scenarios are created. This means describing the goal of the AI system, listing the stakeholders and their expectations. In addition, the process, the domain, and the context in which the AI is/will be embedded are described. For example, some domains or states have different guidelines, regulations, and laws than others. The same applies to different contexts in which different values and rights may exist. In the USA, for example, the approval of a medical device is based on the product's efficacy, whereas in Europe the requirement for approval is that it fulfils the planned effect [123]. Further, this step establishes a body of evidence that supports the claims of the development and other stakeholders. Most importantly, this step helps to improve understanding of the AI system and its context. Awareness of the different actors and their interdependencies is created and topics which should not be included in the assessment are defined. In Section 3.3, this step is used for the description of the domain and context of the use case.

A possible framework for the creation and description of social-technical-scenarios are, for example, discussion workshops in which the expert team works together. Here, the team of experts can use checklists, toolkits, or domain-specific frameworks. A framework that was originally developed to investigate security and trustworthiness is the CAE framework. It is described in Section 2.4.2 and then examined in Section 4.1.1 to see whether it is suitable for use in this step, i.e., for the description of social technical scenarios and in particular the description of the evidence situation.

Step II: Identification of ethical issues and tensions:
In this step, possible ethical issues and flags that may arise in the use of AI systems are identified. Flags represent ethical issues that require further investigation to validate if they are in fact ethical issues, e.g., because technical details need to be clarified. Some ethical issues and flags raise tensions where values are in conflict. In the Z-Inspection® process, these are called ethical tensions. These should be identified and presented in this step. For example, there might be a tension between the competitive values of efficiency and privacy, where the use of data is to improve efficiency, but the privacy of the data provider suffers. Or the tension between accuracy and fairness, because an algorithm that is very accurate on a mean can possibly neglect a minority [124]. In this step, the interdisciplinary composition of the expert team is again particularly important in order to consolidate as many different perspectives from different interest groups as possible.

Step III: *Mapping to trustworthy AI:*
After the creation, presentation and classification of the ethical issues, flags, and ethical tensions, the third step of the Access-Phase follows. Here, Z-Inspection® uses the four ethical principles and the seven key requirements for trustworthy Artificial Intelligence of the High-Level Expert Group on AI (see Section 2.1 In the course of this step, the ethical issues, flags and ethical tensions formulated in step II are mapped to the four principles (Respect for human autonomy, prevention of harm, fairness und explainability) and the seven key requirements (Human agency and oversight, technical robustness and safety, privacy and data governance, diversity, non-discrimination and fairness, societal and environmental wellbeing, accountability, transparency*).*

In Section 4.1.2 and 4.1.3, step II and step III are applied in combination with the help of the CAE framework. These steps are used to identify ethical issues, flags and tensions related to the use case and to map them to the four ethical principles and seven key requirements of the AI HLEG.

Step IV: Execution:
Step IV is the last step in the Assess-Phase. In this step, the team of experts decides on an overall strategy and an implementation plan for the assessment and analysis based on the ethical issues and flags that have been identified. The aim here is to decide at which level of abstraction an assessment and analysis of the AI system is necessary and meaningful. The result of this step is the delivery of evidence on the given ethical issues and flags. In Section 4.2 and 5 this step is applied with a focus on fairness and explicability. Ethical issues and flags related to the principles of respect for human autonomy and prevention of harm are already provided with evidence in Section 4.1.3.

Iterate:
With the fourth step, the Assess-Phase is completed. However, the Access-Phase can be an iterative process and the individual steps can be repeated depending on the use case. This could be necessary if, for example, flags cannot be conclusively examined, and further information is needed or if circumstances change.

Resolve-Phase: In this phase, the ethical issues and tensions are presented, with the indication that they can be resolved if a possible solution has been identified in the Assess-Phase. Recommendations are made to the various stakeholders. The protocol documented throughout the process can also be shared with the stakeholders. Sections 5 and 6 describe possible recommendations in the context of the use case.

2.4.2 Claims, Arguments and Evidence Framework

Another method to assess the security and trustworthiness of systems is the Claims, Arguments and Evidence (CAE) framework. Developed for safety and assurance cases, i.e., independent of AI systems, it provides a framework to question and verify properties of a complex system and build a basic understanding [125]. This is especially a prerequisite for the ethical principle of explicability that should be fulfilled and which in turn forms the basis for a trustworthy system. As shown in Figure 5, the framework consists of the three components claim/subclaim, argument and evidence[126].

Claim: A claim describes a property of the system or object in the form of a true-false statement. This statement should convince of the correctness of the claim and must either be subdivided into further sub-claims (1) or, if the claim cannot be further specified (2), strengthened with the corresponding evidence. A single claim should be formulated directly and focused on a specific system or object. It should be brief and concise, omitting words such as and/or to allow for clear verification of the true-false statement and proper attribution of the evidence. A claim should follow the form: System X has property Y [127].

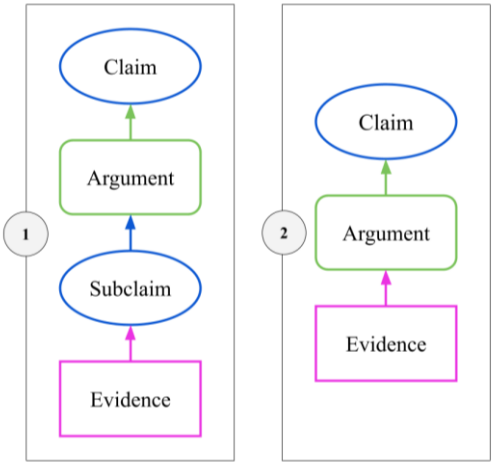


Figure 5: CAE concepts and notation (modified from [126]).

Argument: An argument connects claims with their respective sub-claims or evidence. It is formulated as a rule and justifies the division of the claim into sub-claims or its evidence [127]. Extensions can be used to describe arguments in more detail. There are five so-called CAE building blocks (*decomposition, substitution, concretion, calculation, and evidence incorporation*) that explain the validity of claims or the way claims are derived into sub-claims [128].

- *Decomposition*: Subdivision of the claim into sub-aspects.
- *Substitution*: Replacing the claim with a similar claim of an equivalent system/object.
- *Concretion*: More detailed description of the claim.
- *Calculation*: Calculation of the value of the claim.
- *Evidence incorporation*: Evidence to directly support the claim [128].

Evidence: Evidence concludes a claim, argument, (subclaim) and confirms or disproves the (sub)claim made. It presents facts that create confidence in the claim or in the sub-claim [126]. Therefore, the trustworthiness of the evidence itself is of great importance. It should be real evidence that can be verified at any time and comes from a reliable source. If an evidence supports or disproves only a part of the claim or argument, further sub-claims and arguments are necessary [127].

By establishing claims, arguments and evidence in this way, the functions and properties of a system can be examined with regard to desired requirements and explained and questioned with regard to security and trustworthiness. Section 4.1.1 examines to what extent this framework can be used to create social technical scenarios and thus to identify issues, flags and tensions in Step II: Identification of ethical issues and tensions and Step III: Mapping to trustworthy AI of the Assess-Phase of the Z-Inspection® process (see Section 2.4.1) for the use case.

2.4.3 Assessment List for Trustworthy Artificial Intelligence

The third evaluation method for trustworthy AI systems presented in this thesis is the Assessment List for Trustworthy Artificial Intelligence (ALTAI) developed by the AI HLEG [129]. Building on the requirements for a trustworthy AI system described in Section 2.1, ALTAI includes questions, that

can help companies to assess the trustworthiness of the AI system already during the development phase of the AI system. The assessment list[2] is also available online via the Prototype Web-based tool[3] developed by the AI HLEG and team at the Insight Centre for Data Analytics at University College Cork. It dynamically guides developers, lawyers, and other stakeholders through the list of questions using single-choice, multiple-choice, or text responses [131],[132]. As shown in Figure 6, there is a detailed questionnaire for each of the seven key requirements.

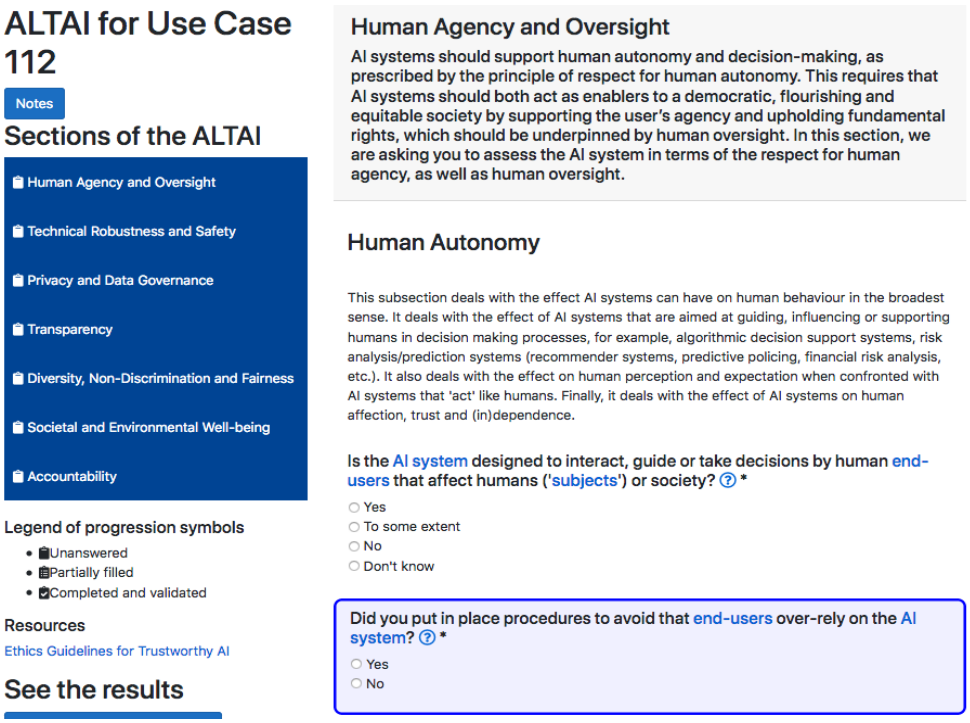


Figure 6: ALTAI Example: Human Agency and Oversight [131].

The questions to be answered are marked with three colours to indicate their respective meaning. Questions with a white background symbolize the reference to properties of the AI system; answers to questions outlined in blue are included in the generation of recommendations; questions outlined in red are intended to help users self-assess the company's or organization's compliance with specific requirements. For a better understanding, each submenu briefly describes the respective requirement and provides a glossary and examples for each question (see Figure 6). After answering all questions, the evaluation results can be viewed in two Sections. While the self-assessment is visualized as a network diagram with the achieved scores in all seven requirements, recommendations are displayed and listed for each submenu/requirement [131],[132]. A multidisciplinary team with competencies in the seven key requirements (see Section 2.1) should preferably answer the questions, whether by checklist or online tool, in order to be able to assess and evaluate all questions from different perspectives. Team members can be assigned with different organizational areas, such as design, development, compliance or management [132].

The analysis of the given use case takes place in Sections 3.2, 4, and 5 with the help of the three methods described above (Z-Inspection®, CAE Framework, and ALTAI). Thereby, different elements of the methods are focused and combined. The questions of the ALTAI assessment list serve

[2] Available online at [130].
[3] Available online at [131].

as a guide for the division of levels in the CAE Framework and thus as an aid in analysing the trustworthiness of the AI system under investigation. The application of the CAE framework is again assigned to the two steps Step II: Identification of ethical issues and tensions and Step III: Mapping to trustworthy AI in the Z-Inspection® process. Before analysing the use case in Section 4, it is described in the next Section.

3 Use case: ML supported OHCA recognition

In the following Section, we provide an overview of the use case available to us. It is an AI system that will be used to assist in the detection of OHCA and consequently in the healthcare domain. It starts in Section 3.1 by describing AI systems in general in the healthcare domain and their importance. Then, in Section 3.2, the AI system is classified according to application and methods in order to be able to analyse the use case in terms of socio-technical scenarios (see Section 2.4.1) in Section 3.3. Finally, in Section 3.4, an overview of the previous studies of the use case is given.

3.1 AI systems in healthcare

One domain in which AI systems are gaining more and more importance is medicine. Numerous examples show that AI systems can be very effective and accurate: A ML model from Japan predicts cancerous tumors with 94% accuracy [133]. With an accuracy of 98.5%, the Guardian Connect ML model can identify changes in the blood glucose level of diabetics one hour in advance and warn the patient of hypoglycaemia [11],[134]. More examples show that AI systems are being applied to a wide range of medical problems: AI systems that can substitute a doctor's diagnosis of retinopathy for a retinal scan of diabetic patients was approved by the US Food and Drug Administration in 2018 [135]. Medical follow-up after surgery, medical image analysis, tuberculosis diagnosis and optical coherence tomography (OCT) diagnosis are now also supported by AI systems [11]. As it can be seen, the application of AI systems in the healthcare industry covers a wide range of medical areas, such as surgery, assistance to healthcare professionals, medical consultations, administration and processes, diagnosis, and medication management. Due to the increasingly prominent shortage of healthcare professionals and ever-rising costs in the healthcare sector, the need for supporting or even replacing systems has long been there and as described above, is already being used frequently. AI systems can improve processes and diagnoses, save time, bring quality growth to medicine, control abuse and help with decision-making through their use in a wide range of applications in the healthcare industry [11].

The benefit of AI systems is measured by the extent to which the AI system improves the results of prognoses, diagnoses, and interpretations, reduces the workload of medical professionals and, of course, whether and how the costs of healthcare can be reduced through the use of AI. Among other things, this predicted savings potential shows that the AI market in the healthcare industry has great market potential. It has an average global growth rate of 28% [11]. Information about the quality of an AI system in healthcare can also be provided by the ratios, which were defined in Section 2.2.4.

The application and functioning of AI systems in the health sector can be divided into different categories. On the one hand, the distinction as to whether the AI systems interact with humans/healthcare professionals or perform tasks independently, and on the other hand, the distinction as to whether the systems are specific or adaptive. According to [11], four different application categories emerge:

- *Assisted Intelligence* are specific AI systems that support medical staff in decisions or actions.
- *Automated Intelligence* are specific AI systems that perform routine tasks autonomously and automatically.

- *Augmented Intelligence* are adaptive AI systems that help medical staff make decisions and continue to learn from the context.
- *Autonomous Intelligence* is an adaptive AI system that can autonomously adapt to different contexts without human interaction. Table 7 shows a visualization of these categories.

	Human interaction	**No human interaction**
Specific systems	Assisted Intelligence	Automated Intelligence
Adaptive systems	Augmented Intelligence	Autonomous Intelligence

Table 7: Categories of AI application domains in the healthcare industry [11].

A further distinction can be made with regard to the AI methods used in the healthcare sector: ML model (neural networks and deep learning) are most frequently used in medicine, but due to the variety of application areas in the healthcare sector and thus the different requirements for different systems, machine vision/computer vision, rule-based expert systems, physical robots and robotic process automation methods are also used and developed [11],[16].

3.2 Description of use case's AI system

This Section gives an introduction to the AI system of the use case "Machine Learning as a supportive Tool to recognise Cardiac Arrest in emergency calls". It briefly explains and delineates what the system is used for. Subsequently, the technology of the AI system is described in rudimentary form and applications, and methods from Section 3.1 are classified. The base is the information publishes in a retrospective study [136] and RCT [20]. We also make use of expert opinions from the working group that evaluates the use case using the Z-Inspection® process [17],[137].

The AI system supports medical dispatchers in the detection of OHCAs. Such an OHCA is a life-threatening status where the patient is not in a medical facility and quick detection can save lives. It is important to mention here that this use case is only about the detection of OHCA. Among non-professionals, there is often no distinction made between cardiac arrest and heart attack. The beginning of a heart attack is usually announced a few hours beforehand by pain in the upper part of the body, while cardiac arrest is a sudden and unexpected stop of the heartbeat caused by an electrical malfunction of the heart. This prevents oxygen-rich blood from reaching the brain, lungs, or other organs. This sudden appearance without warning is the reason why patients almost never make the emergency call themselves in the case of an OHCA. Patients with cardiac arrest often breathe agonally. Agonal breathing is a short, laboured, and gasping type of breathing that occurs due to the lack of oxygen supply to the brain. Without treatment, the patient loses consciousness and pulse within seconds and death occurs within minutes [138]. In both cases, a quick call to the medical emergency centre is of major importance. However, while in the case of a heart attack the caller can no longer provide medical assistance after making the emergency call, in the case of a cardiac arrest the bystander can initiate cardiopulmonary resuscitation (CPR) and even use an automated external defibrillator (AED) if available. These techniques would increase the patient's chance of survival. Therefore, the rapid recognition of OHCA by the bystander or a dispatcher is even more important so that CPR and possibly even an AED can be used before the arrival of the rescue service. For every minute that passes without resuscitative measures, the probability of

survival is reduced by app. 10%. Therefore, it is important that a trained medical dispatcher recognise a cardiac arrest within the shortest possible time. Since the bystander is often a non-professional, the assessment of the dispatcher is essential to be able to give instructions for CPR. The aim of the AI system is to assist dispatchers in recognising OHCAs over the phone [136].

Technology: In developing the AI system that helps dispatchers identify OHCA more quickly, special emphasis was placed on high sensitivity. This means that the ML model has a high probability of identifying an OHCA if the patient actually has the OHCA. With lower probability, the ML system will not identify a patient who actually has OHCA. In return, one accepts more false positive cases. The potential harm is therefore less than with high specificity. It causes less harm to identify non-OHCAs as one than not to identify actual OHCAs.

As described in Section 3.1, there are many different methods that are used in the field of AI in healthcare. For the use case described here, an individual ML model was developed by the company Corti [139], which is based on a deep neural network and the AI system's task is NLP (see Section 1). A off the shelf system was not an option for this use case because the quality and language of the calls are sometimes poor. Emergency calls are not only made from calm environments, but also from traffic areas. In addition, the language of the callers is very different. It has to be considered that the person who makes the emergency call is in an extreme situation and often speaks unclearly and emotionally. Since the ML model has been developed and tested in Denmark, it is Model which is trained on Danish vocabulary. First, the audio track and the spoken words of the dispatcher and the bystander are translated into text and then analysed by a classifier. This then identifies an OHCA or not [137]. Furthermore, as a DNN, the ML model is considered a black box model, so no direct statements can be made about the way the prediction is made [137].

According the knowledge available to us, the ML model for the use case at hand is an Assisted Intelligence AI system (see Section 3.1). It only serves to support the medical dispatcher in diagnosing an OHCA. The ML model was developed specifically for this application and cannot initially fulfil any other tasks [137].

3.3 Social-technical analysis of the use case

The AI system is evaluated by an interdisciplinary and international team of experts under the direction of Prof. Zicari with the help of the Z-Inspection® process regarding ethical implications and trustworthiness. In the course of this master's thesis, we had the opportunity to participate in some workshops of the expert team. The evaluation of the use case in the context of this master's thesis focuses in particular on the ethical principles of fairness and explicability (with focus on explainability) of the ML model and is partly (not entirely) based on the Z-Inspection® process described in Section 2.4.1.

In the following, the aim of the AI system is explained in more detail and the stakeholders as well as their interactions are listed. It also describes the domain, context, and process in which the AI system is embedded. This description represents the Step I: Analysis of social technical scenarios of the Assess-Phase of the Z-Inspection® process described in Section 2.4.1.

Assess-Phase, Step I: Analysis of social technical scenarios
Aim: Every year, more than 600000 people in the USA and Europe are victims of OHCA. However, this accounts for only a small proportion of total emergency calls and thus presents a challenge in recognising these OHCA. Around a quarter of OHCA emergency calls are not correctly assessed by the dispatchers during the telephone call [20]. The reason for this is on the one hand the rarity of the cases and thus the lack of routine on the part of the dispatchers. Statistics show that the reason for an emergency call is made only 1% of the time because of an OHCA. This means that a dispatcher is only confronted with about 10-20 OHCAs per year. On the other hand, the person who makes the emergency call is in all cases not the patient himself (as mentioned in Section 3.2), but a bystander (e.g., the husband or a passer-by). A bystander often has a limited attention span due to the extreme situation. There are also language barriers to overcome. The idea is that there are patterns in the emergency calls that a ML model can classify and thus facilitate and improve the recognition of cardiac arrests. Most importantly, the ML model should be able to perform OHCA prediction faster than a medical dispatcher and thus support the human dispatcher in detecting OHCAs.

Context and domain: The AI system is used in the healthcare domain. As soon as an emergency call is received by the Copenhagen emergency call centre (through the emergency number 112), the AI system listens to the conversation between the medical dispatcher and the caller and as soon as the AI system suspects a cardiac arrest, the dispatcher at the emergency call centre will be informed about it. The dispatcher then has the option to either ignore the alarm or accept it and then initiate CPR or AED instructions accordingly [137].

Actors: The parties directly involved in the AI system are the patient, the bystander, and the medical dispatcher. Further involved is Corti, the company that developed the ML system discussed here. Corti is a company focused and specialised in the development of support systems in the healthcare industry [139]. Stig Nikolaj Bloomberg and his team provide the medical requirements for the system. They also deliver the data with which the ML model is trained. Other stakeholders in the long run are hospital owners, governments, and the whole population/society.

Actor's interactions with the AI system: A patient suffers an OHCA, the patient's heartbeat suddenly stops. A bystander makes an emergency call and gets through to a medical dispatcher. The dispatcher asks the caller certain questions to determine the patient's health status. During the conversation, the AI system listens, but does not have the opportunity to interact with the patient itself. The AI system is a silent observer. However, as soon as it suspects an OHCA, it initiates an alarm for the dispatcher. The dispatcher is free to react to the alarm or to ignore it. If the ML model's prediction is accepted or if the dispatcher independently detects an OHCA, the dispatcher instructs the caller to take life-saving measures. He also sends an ambulance with a doctor to the patient. Until the ambulance arrives, the dispatcher is in contact with the bystander to provide auditory support for resuscitation measures.

3.4 Use case studies

The AI system described in Section 3.2 and 3.3 for the use case has already been tested in Denmark at different times and in different ways. Both studies, the retrospective [136] on the one hand and the randomized clinical trial [20] on the other, use the same ML model without changes in the

architecture or in other structures [137]. Both test procedures are shortly described in the following.

In conducting the first test in 2014, the ML model has been tested using recorded real calls received and recorded at the Emergency Medical Dispatch Centre in Copenhagen from 1/1/2014 to 31/12/2014. The experiment has been authorised by the Danish Health and Data Protection Authority on the one hand and the regional ethics committee on the other. The aim of this retrospective study was to compare the performance of the ML model with that of a medical dispatcher: Does the ML model detect an OHCA fundamentally better than a medical dispatcher? Are there temporal differences in the detection of an OHCA between the medical dispatcher and the ML model [136]? In this study, the ML model was fed with a complete recorded call and then let the ML model make a prediction. In the process, the AI system does not interact in any way with either the dispatcher or the caller [137]. A total of 108607 received emergency calls could be used for the analysis, excluding calls where, for example, the audio file was corrupted, or the patient had already died. Of the 108607 emergency calls, 918 patients suffered from an OHCA. The danish cardiac arrest registry identified and confirmed this [136]. From the results published in the paper "Machine learning as a supportive tool to recognise cardiac arrest in emergency calls", we were able to calculate the following confusion matrix (see Table 8, and Table 9) with the values which were defined in Section 2.2.4 for both ML model recognition and medical dispatcher prediction. The ratios from Section 2.2.4 could be derived from the confusion matrix. In some cases, the ratios are published in [136], as shown in brackets after the values we have calculated. The calculation of the ratios yielded slightly different results for us, but this can be attributed to rounding differences (see Table 10).

	Actual positive	**Actual negative**	
Predicted positive	772	2908	3680
Predicted negative	146 (!)	104781	104927
	918	107689	108607

Table 8: Confusion matrix with calculated absolute values of the ML model's OHCA recognition from the retrospective study [136][4].

	Actual positive	**Actual negative**	
Predicted positive	665	1292	1957
Predicted negative	253 (!)	106397	106650
	918	107689	108607

Table 9: Confusion matrix with calculated absolute values of the dispatcher's OHCA recognition from the retrospective study [136][5].

The sensitivity of the ML model is significantly higher: In 84.09% of the cases, the ML model correctly identifies an actual OHCA, whereas the medical dispatchers correctly identify an actual OHCA only in 72.4% of the cases. This means that the ML model fails to identify an actual OHCA patient in app. 15% (FNR), whereas medical dispatchers fail to identify an actual OHCA patient in 27.56% (FNR). The specificity is slightly lower with the ML model: In 97.3% of the cases, the ML model did not diagnose OHCA in a patient not suffering from OHCA, whereas 98.8% of the patients

[4] (!): Harmful.
[5] (!): Harmful.

not suffering from OHCA were also not diagnosed with OHCA by the medical dispatchers. Furthermore, a significantly lower PPV can be said for the ML model. This is because the false positive value is higher in the case of ML model recognition and thus the specificity is slightly lower than with dispatcher recognition. This has a strong influence on the PPV given the low number of OHCA: 20.9% of the cases predicted as OHCA by the ML model actually suffer from OHCA. Among the OHCAs predicted by the medical dispatchers, 33.0% actually suffer an OHCA. The NPV hardly differs between the ML model and the medical dispatchers (99.9% vs. 99.8%): In almost all of the non-OHCAs predicted by both the ML model and the medical dispatchers, the patients did not actually suffer an OHCA. However, it is important to mention here that the large number of non-OHCAs and the naturally unequal distribution of positive and negative cases distort the NPV statement.

Ratio	ML model	Dispatcher
Sensitivity	84.1% (84.1%)	72.4% (72.4%)
Specificity	97.3% (97.3%)	98.8% (98.8%)
PPV	20.98% (20.9%)	33.98% (33.0%)
NPV	99.86% (99.9%)	99.76% (99.8%)

Table 10: Calculated ratios of ML model and dispatcher. Number in brackets from [136].

In addition to the general detection of OHCA by the ML model compared to the medical dispatchers, the time to detection was also investigated in this study. The time to detection is defined by the time interval from the answer of the emergency call until the dispatcher pronounces an OHCA or formulates the need for CPR or AED. The ML model registers a significantly shorter recognition time of a median of 44 seconds compared to the medical dispatchers of a median of 54 seconds [136]. Precisely because speed of detection of OHCA is a life-saving component, 10 seconds difference matters a lot.

Since this study only worked with recorded audio files, a second examination has been carried out under real-time conditions. In the second live study [20], the AI system presented in Section 3.2 and 3.3, makes the prediction during the emergency call and alerts the medical dispatcher when an OHCA is detected. The purpose of this RCT is to demonstrate the potential of the ML model to assist in the detection of OHCAs and if there is a reduction in detection time compared to medical dispatchers. Furthermore, it is to be determined whether and to what extent the ML model influences the clinical practice of medical dispatchers. The data on which this work is based comes from this study (see Section 4.2). This study was registered with the Danish Data Protection Agency and approved by the Danish Patient Safety Authority. In the context of the RCT, all emergency calls from 9/1/2018 to 12/31/2019 received by the emergency number 112 from the medical dispatcher in Copenhagen were considered. The ML model analyses - in parallel with the medical dispatcher - the ongoing live calls. Once the ML model identifies an OHCA, it randomly sorts the call into one of two groups: The intervention group and the control group. The dispatchers in the intervention group, as soon as the OHCA is detected by the ML model, receive a notification of the OHCA detection. If the call was randomly placed in the control group, nothing changes for the dispatcher and they do not receive notification of the detection of the OHCA by the AI system. The medical dispatchers have not been informed about the groups and therefore did not know

whether the ML model is in operation for the current call received. If the dispatcher receives the notification of the OHCA detection, he/she is free to decide whether to rely on the ML model and take the necessary steps or to ignore the alert.

During the study, the ML model was tested on a total of 169049 emergency calls. Of these 169940 emergency calls, 605 calls were excluded. From the remaining 168444 emergency calls, the ML model identified 5242 of the calls as OHCA, of which 959 patients actually suffered from OHCA. From the results published in [20], we were able to calculate the following confusion matrix (see Table 11). For comparison, here is again the confusion matrix from the retrospective study. The results of the ratios in this study are very similar to those in the retrospective study (see Table 8), which is why the ratio results are not shown here.

	Actual positive	**Actual negative**	
Predicted positive	959	4283	5242
Predicted negative	169 (!)	163033	163202
	1128	167316	168444

Table 11: Confusion matrix with calculated absolute values of the ML model's OHCA recognition from the RCT [20][6].

	Actual positive	**Actual negative**	
Predicted positive	772	2908	3680
Predicted negative	146 (!)	104781	104927
	918	107689	108607

Table 8: Confusion matrix with calculated absolute values of the ML model's OHCA recognition from the retrospective study [136][7].

Of the 959 cases correctly identified as OHCA (TP), 305 cases have been screened out due to the presence of an ambulance, broken alarm chain or because CPR was already started before the call. This leaves 654 calls that have been randomly divided into the intervention and control groups. This results in the following matrix in Table 12.

	Interventions group (with ML)	**Control group (without ML)**	
Dispatcher recognises OHCA	296	304	600
Dispatcher does not recognises OHCA	22	32	54
	318	**336**	654

Table 12: Division of recognised calls between control group and intervention group [20].

This study mainly compared the control group with the intervention group to find out to what extent the ML model influenced the dispatcher. The focus of this master's thesis is to compare the OHCA recognition time of the ML model and the medical dispatcher in order to investigate possible biases (fairness) and explainability (see Section 4.2). The assessment of the use case follows in the next Section.

[6] (!): Harmful.
[7] (!): Harmful.

4 Use case assessment

In this Section, the previously described use case is ethically assessed. The theoretical evaluation methods described in Section 2.4 are combined and then supplemented with a data analysis of the true positives. In the first step in Section 4.1, the evaluation is carried out by identifying ethical issues, flags and tensions. This is followed by the analysis and evaluation of the data with the help of selected features in Section 4.2.

4.1 Identification and evaluation of ethical issues and flags

The CAE framework (see Section 2.4.2) is examined in the following Section 4.1.1 for its suitability and then, in Section 4.1.2, applied in combination with the questions of the Assessment List for Trustworthy AI (see Section 2.4.3). The resulting ethical issues and flags are then described (Section 4.1.3) and tensions between them are derived.

4.1.1 Transferability test of the CAE framework

As already explained in Section 2.1, the AI HLEG defines the trustworthiness of an AI system via the fulfilment of the four ethical principles and their seven key requirements. These can be specified across several levels, making a detailed analysis of requirements for a trustworthy AI system possible and necessary in order to evaluate the AI system and uncover possible weaknesses. Since the CAE framework can be used to map and concretise different levels via the respective arguments, it is worth discussing and investigating whether the CAE framework is suitable for analysis and evaluation as part of the Z-Inspection® process (see Section 2.4.1). The Z-Inspection® process represents a uniform model for the evaluation of AI systems with regard to trustworthy AI systems. Since the evaluation of an AI system for trustworthiness in the context of a use case is an extensive project, the categorisation, organisation and structuring of the individual levels and problems is of high importance. The CAE framework provides a standardised framework for this structuring. In addition, the CAE framework is designed to deal with problems of security and trustworthiness. This is in line with the issues addressed in the Z-Inspection® process. Although the CAE framework is not calibrated for AI systems, it can be applied to issues of AI systems. At the end of a path in the CAE framework, after subdividing the main claim by arguments into sub-claims, there is evidence. In the classical CAE framework, there is only one notation for an evidence. It is not directly visible whether it is evidence, counter-evidence or the need for further investigation. In Section 4.1.3, we therefore add three notations to the claims. In the context of the Z-inspection process, there are three cases that can be at the end of such a path of the CAE framework: The case where the sub-claim can be verified and there is evidence for the claim. Such verified claims are referred to as $VC_{1...n}$ in the following. If there is no evidence for the end of a path (sub-claim) but there is counter-evidence, this represents an ethical issue within the framework of the Z-inspection process. In the further course, this is noted as $I_{1..n}$. If there is neither evidence nor counter-evidence for the sub-claim, this case is called according to the Z-inspection flag. In the following here represented by $F_{1..n}$. Therefore, if the CAE framework is applied in the context of the Z-Inspection® procedure, a supplement for these cases is required.

In conclusion, it can be said that the CAE framework can be used in the crystallisation of ethical issues and flags in the course of the Z-Inspection® process (see Step II of the Assess-Phase, Section 2.4.1) through the standardised division and concretisation of the main claim *The AI system is*

trustworthy. By splitting the claims in a structured way, which can be done using the four ethical principles and seven key requirements, even Step III can be performed using the CAE framework. The subdivision of the claims is partly based on the structure of ALTAI (see Section 2.4.3). However, extensions are needed to represent the case of an issue or a flag. In Section 4.1.3, according to Z-Inspection®, the case of an ethical issue is described with $I_{1...n}$ and the case of flags with $F_{1...n}$. In a next step, tensions can be derived from the issues, flags and VCs that have been identified. The issues and flags require further consideration and analysis, which is then carried out using the data and information available to us (see Section 4.2).

4.1.2 Application of the CAE framework to the use case

After Step I of the Assess-Phase has been applied to the use case in Section 3.3, the Z-Inspection® process then provides for the creation of an evidence base for the stakeholders' assertions, the description of social technical scenarios and the derivation of issues, flags and tensions and the assignment of these to the four ethical principles and the seven key requirements (see Step II and III). In this step, we use the CAE framework for the reasons mentioned above (see Section 4.1.1). The structuring of the questions from the Assessment List for Trustworthy AI (see Section 2.4.3) is partly used for the concretisation and division of the claims. The CAE framework applied to the use case is described below.

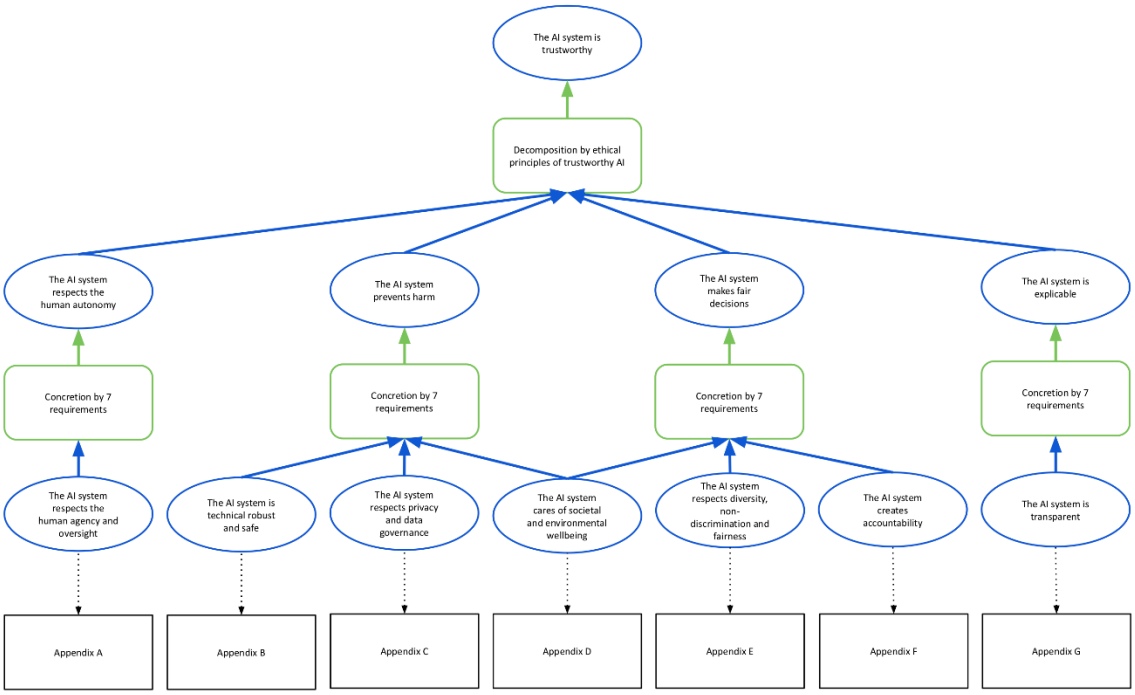


Figure 7: CAE decomposition by ethical principles of trustworthy AI.

The main problem to be investigated and addressed by the international and interdisciplinary team of experts is the question of whether and to what extent the AI system of the use case is trustworthy or ethically justifiable. The main claim to be investigated is therefore *The AI system is trustworthy*. Since both Z-Inspection® and the ALTAI use the four ethical principles of the AI HLEG to evaluate AI systems, the first argument is a decomposition of trustworthy AI into the four ethical principles, as shown in Figure 7. Subsequently, the argument of a concretion by the seven

requirements assigns the seven requirements to the four principles (see Section 2.1 and Figure 7). The further division of the subclaims into other subclaims with the help of various arguments until the end of the path are shown in Appendix A-Appendix G, depending on the requirements.

The end of each path represents the most concrete subclaims after division and concretisation. From these, ethical issues, flag and tension are derived in Section 4.1.3, as intended by the Z-inspection® process.

4.1.3 Mapping of ethical issues, flags and derivation of tensions

In Section 4.1.2 the CAE framework was applied to the present use case. At the end of a path, a concrete claim is made, representing either verified ethical claims, abbreviated as VC in the following course, ethical issues I (unverified claims) or flags F (claims for further investigation). Below, these are listed in the form *VC/I/$F_{1,\dots,n}$(Ethical principle; Key requirement)* according to ethical principles and key requirements (see Section 2.1) and are described using the CAE notation (Claim, Evidence). This represents Step III of the Assess-Phase of the Z-Inspection® process. The evidence is thereby extended by a content-related justification:

Verified Claims $VC_{1,\dots,n}$: The use case fulfils the following ethical claims.

VC_1(Respect for human autonomy; Human agency and oversight):

Claim: The dispatcher is allowed to ignore the AI decision.

Evidence: The dispatcher is not forced to accept the ML alert. He/she can make his/her own prediction despite the ML decision and thus act autonomously [20],[137]. This is also demonstrated in the next Section of the data analysis (see Section 4.2.2). It can be seen that only a small part of the alerts is accepted (8.82%) and consequently the dispatcher meets his own prediction.

VC_2(Respect for human autonomy; Human agency and oversight):

Claim: Communication to patients only by dispatcher.

Evidence: The patient communicates only with the dispatcher and not with the system. Appropriate responses and supervised decisions can thus be ensured [20],[137].

$VC_{3\text{-}5}$(Prevention of harm; Privacy and data governance):

Claims: The AI system respects the GDPR standards (use of data, storage duration, ...), respects the ethical data protection standards, and respects the data protection in healthcare.

Evidence: The AI system is developed and deployed within the EU. It therefore respects the right to privacy and complies with the applicable data protection regulations, for example on the purposes for which the data is used or how long it is stored [20],[136]. The retrospective study [136] was authorised by the regional ethics committee. The regional ethics committee has decided not to re-approve the live study[20]. Therefore, it can be assumed that the standards are respected.

VC_6(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system does not force the dispatcher to adopt recommendations.

Evidence: Dispatchers can decide for themselves whether to trust the ML decision or to disagree if necessary. The acceptance of the AI system is present due to the non-coercion [20],[136], however, the trust in the decisions of the AI system does not

seem to be there (yet), because in the use case presented to us only 8.82% of the medical dispatchers accept the ML alter (see Section 4.2.2).

VC_7(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system makes faster predictions.

Evidence: The ML model has a lower mean OHCA recognition time compared to that of the dispatcher [20],[136]. The faster OHCA recognition time of the ML model by about 36s in mean is also shown in Section 4.2.2. Thus, the ML model can detect an OHCA earlier and actions could be taken sooner if the alert is accepted.

VC_8(Fairness; Diversity, non-discrimination and fairness):

Claim: The ML model is developed with clinicians and dispatcher's participation.

Evidence: Medical professionals and their expertise were included in the development process. A consideration of diversity, non-discrimination and fairness is therefore given from multiple perspectives [137].

VC_9(Fairness; Accountability):

Claim: The AI system gets audited by internal and external organizations.

Evidence: See Evidence of VC_4.

Ethical Issues $I_{1,...,n}$: The use case does not or not completely fulfil the following claims. The evidence therefore serves as counter-evidence. Issues that relate to the ethical principles respect for human autonomy and prevention of harm are briefly provided with counter-evidence at this point using the knowledge available to us [20],[136],[137]. Issues related to the ethical principles fairness and explicability are discussed in detail in Section 5.1.1 and 5.2.1.

I_1(Respect for human autonomy; Human agency and oversight):

Claim: The patient knows about the AI system.

Evidence: The patient does not receive any information about the use and functioning of the AI system during the call, and thus his autonomy is compromised by the asymmetric information [137].

I_2(Respect for human autonomy; Human agency and oversight):

Claim: The patient knows about the division of labour.

Evidence: The patient does not receive any information about the division of labor between the dispatcher and the AI system during the call. This also leads to an asymmetric distribution of information and consequently impairs the patient's autonomy [137].

I_3(Respect for human autonomy; Human agency and oversight):

Claim: The dispatcher is trained to use the system.

Evidence: The dispatcher does not receive full training on the AI system and therefore cannot supervise it due to the lack of knowledge [20]. Training and additional explanations of functions could help to strengthen the autonomy of the dispatcher and enable supervision of the system.

I_4(Prevention of harm; Technical robustness and safety):

Claim: The AI system has a low rate of technical faults, outages, attacks and misuse.

Evidence: The AI system is subject to server downtimes, which leads to failures of the AI system. Consequently, it is not technically protected against failures and is therefore not technically robust [20].

I_5*(Prevention of harm; Technical robustness and safety)*:

Claim: A backup server is installed.

Evidence: The AI system is executed via a server, which is not mirrored. Consequently, there is no backup server and reliability cannot be guaranteed [20].

I_6*(Prevention of harm; Technical robustness and safety)*:

Claim: The AI system is tested in different contexts (e.g. different hospitals, languages).

Evidence: The AI system has so far been tested in a single real-world context and in one language (Danish), so reproducibility cannot be guaranteed. However, according to Corti, the NLP model is currently being extended to other languages [137].

I_7*(Prevention of harm; Privacy and data governance)*:

Claim: The AI system allows the selection of privacy settings and the approval of ML support.

Evidence: The patient/caller does not have the opportunity to consent to privacy settings or the general use of the AI system during the call. He therefore does not have a complete right to privacy and data protection (assumption based on [140]). It must also be examined here to what extent the legal bases of the GDPR apply.

I_8*(Prevention of harm; Privacy and data governance)*:

Claim: The AI system records all information correctly and makes adequate predictions due to this information.

Evidence: The AI system does not evaluate all calls due to faults like system downtimes or broken call chains, for example, and also makes misclassifications. System integrity is therefore not given [20],[136].

I_9*(Prevention of harm/Fairness; Environmental and societal well-being)*:

Claim: The dispatcher is trained on the AI system.

Evidence: The dispatcher is not sufficiently trained in dealing with the AI system [20]. The extent to which this negatively influences the dispatcher's trust and acceptance of the AI system is explained in more detail in Section 5.1.1.

I_{10}*(Fairness; Diversity, non-discrimination and fairness)*:

Claim: The audio files are not biased.

Evidence: Only calls from Denmark in Danish are used for training and testing [137]. Whether and which biases are present will be evaluated in Section 5.1.1.

I_{11}*(Fairness; Diversity, non-discrimination and fairness)*:

Claim: The AI system offers alternatives for disabled dispatchers (e.g. auditive usage for blind dispatchers).

Evidence: The AI system offers only one form of use via a static user interface [140]. In Section 5.1.1, the requirement of the possibility of full accessibility is examined.

I_{12}*(Fairness; Diversity, non-discrimination and fairness)*:

Claim: The AI system is developed with patients and population participation.

Evidence: The AI system was developed without patient influence [20]. An assessment of diversity, non-discrimination, and fairness including all perspectives follows in Section 5.1.1.

I_{13}(Explicability; Transparency):

Claim: The decisions are explainable by the ML engineers.

Evidence: The ML model is a black box model, so decisions are not apparent without further analysis [137]. The possibilities for such analysis are explained in Section 5.2.1.

I_{14}(Explicability; Transparency):

Claim: The decisions are explainable by the dispatcher.

Evidence: Due to human diversity in terms of complaints and the incomplete familiarization of the dispatcher with the AI system [137], it is assumed that not all decisions can be explained. To what extent decisions of the AI system can be explained follows in Section 5.2.1.

I_{15}(Explicability; Transparency):

Claim: The decisions are explainable by the clinicians.

Evidence: Like mentioned before it is assumed that not all decisions can be causally justified due to human characteristics and the black box model [137]. Therefore, transparency of the decision-making process cannot exist from the ground up, but could be built up as described in Section 5.2.1.

I_{16}(Explicability; Transparency):

Claim: The patient and the bystander are informed about the usage of the AI system.

Evidence: The patient does not receive any information about the use of the AI system during the call [137],[140]. How the limited communication affects the transparency of the AI system is presented in Section 5.2.1.

Flags $F_{1,\dots,n}$: The fulfilment of the following claims should be further investigated. As evidence is not available at the time of analysis, this should be provided by further investigation. As with the ethical issues, flags that are related to the first two ethical principles (respect for human autonomy and prevention of harm) are briefly explained in this step. Fairness and explicability flags are discussed in detail in Sections 5.1.2 and 5.2.2.

F_1(Respect for human autonomy; Human agency and oversight):

Claim: The dispatcher does not rely on the system's decision/uses his own mind.

Evidence (further investigation): In order to preserve the autonomy of the dispatcher in the future, it should be investigated to what extent the dispatcher relies on the decisions of the AI system without questioning them himself. Such dependence can lead to damage in cases of unavailability or incorrect decisions by the system. This could happen, for example, through false manipulated alerts or calls.

F_2(Respect for human autonomy; Human agency and oversight):

Claim: The dispatcher is able to manipulate the system (stop, back, help button).

Evidence (further investigation): The dispatcher should be able to supervise the AI system and track decisions. Additional functions, such as jumping back to the last instruction or calling up explanations, could be helpful here but also affects the speed (T_1). Perhaps a feasibility study could be conducted here. The published video does not show any such possibility for the dispatcher [140].

F_3(Prevention of harm; Technical robustness and safety):

Claim: Risks are measured and assessed and possible consequences are identified.

Evidence (further investigation): A risk analysis should be performed before the AI system is deployed. Risks and their effects should be evaluated and later monitored to ensure technical robustness and to avoid damage. Whether such an analysis has been carried out must be examined in more detail.

F_4(Prevention of harm; Technical robustness and safety):

Claim: The AI system is compliant with security standards.

Evidence (further investigation): The AI system should meet current security standards and be regularly updated. It should be possible to exclude failures and harmful attacks in order to guarantee technical robustness and damage prevention.

F_5(Prevention of harm; Technical robustness and safety):

Claim: The AI system is continuously adapted to the latest OHCA research.

Evidence (further investigation): The AI system should be continuously adapted to the latest research in OHCA in order to always keep the accuracy of predictions at the highest level and avoid possible damage (Temporal bias, see Section 2.2.3).

F_6(Prevention of harm; Technical robustness and safety):

Claim: Accuracy measures are monitored and communicated.

Evidence (further investigation): The accuracy of the ML model should be monitored throughout. In case of a deviation, these key figures can be discussed e.g. in feedback rounds to ensure a safe system without damage.

F_7(Prevention of harm; Technical robustness and safety):

Claim: Processes exist for testing and verification of reliability and reproducibility.

Evidence (further investigation): The reliability and repeatability of the system should be tested and verified at regular intervals to maintain a safe and robust system.

F_8(Prevention of harm; Privacy and data governance):

Claim: The AI system is trained on anonymised data.

Evidence (further investigation): In order to be able to verify compliance with legal data protection regulations also when training the ML model, the training data should be examined for the handling of personal health data (Art. 9 DSGVO [18]).

F_9(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system improves the quality of predictions.

Evidence (further investigation): As described in Section 3.4, the results of the two studies available to us [20],[136] are similar and show a significantly higher sensitivity in the intervention group. This is an indication for the better quality of the OHCA prediction of the ML model, but a consideration of other metrics is needed, which is done in Section 5.1.2.

F_{10}(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system does not replace professionals.

Evidence (further investigation): It should be openly communicated to what extent the AI system will take over tasks from the dispatcher. In the use case, it is not clearly defined how the division of labor between the dispatcher and the AI system is to function in the future. At the moment, OHCA account for only a small percentage of calls. Reference is made to this in Section 5.1.2.

F_{11}(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system has no impact on dispatchers training.

Evidence (further investigation): The current state of knowledge does not include any additional training of dispatchers in the use of the AI system. Section 5.1.2 examines whether and to what extent this should be taken into account in future training.

F_{12}(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system gets promoted in terms of acceptance and transparency.

Evidence (further investigation): The functions and limitations of the AI system should also be communicated externally to the population in order to establish their acceptance and trust together with a general social well-being. Section 5.1.2 discusses what has been communicated so far and what possibilities there may be to improve transparency.

F_{13}(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system does not harm the environment (Green IT, ecological balance).

Evidence (further investigation): The extent to which the AI system affects the ecological environment should also be reviewed. Possibilities for analysis and evaluation are listed in Section 5.1.2.

F_{14}(Fairness; Diversity, non-discrimination and fairness):

Claim: The developer is not biased.

Evidence (further investigation): The impartiality of the developers should be confirmed at the beginning to avoid possible influences in the algorithm and to ensure the fairness of the AI system. Section 5.1.2 discusses impartiality of human developers.

F_{15}(Fairness; Diversity, non-discrimination and fairness):

Claim: The client is not biased.

Evidence (further investigation): In addition to the developers, the other stakeholders, such as the clients, should also be checked for their impartiality. Whether they could influence the development of the algorithm is evaluated in Section 5.1.2.

F_{16}(Fairness; Diversity, non-discrimination and fairness):

Claim: The algorithm is not biased.

Evidence (further investigation): It should be ensured that the algorithm is (as far as possible) free of bias. Whether and which biases are present in the algorithm is given in Section 5.1.2.

F_{17}(Fairness; Diversity, non-discrimination and fairness):

Claim: The training data is not biased.

Evidence: It should be ensured that the training data is (as far as possible) free of bias. Whether biases are present in the training data is examined in Section 5.1.2.

F_{18}(Fairness; Diversity, non-discrimination and fairness):

Claim: Dispatchers and developers are trained to be aware of possible bias.

Evidence (further investigation): Dispatchers and developers should be provided with information on biases and discrimination to be trained in potential influences and impacts to enable a fair AI system. A short assessment to this effect is provided in Section 5.1.2.

F_{19}(Fairness; Diversity, non-discrimination and fairness):

Claim: A fairness definition has been determined.

Evidence (further investigation): Fairness, as described in Section 2.2, should be defined at the beginning of development and communicated to all stakeholders in order to create an equal understanding of a fair AI system. The extent to which a uniform definition is feasible and useful is examined in Section 5.1.2.

F_{20}(Fairness; Diversity, non-discrimination and fairness):

Claim: Easy use of the AI system for every dispatcher regardless of their abilities (easy UI design).

Evidence (further investigation): Every dispatcher and user should be able to operate the AI system equally and independently of possible limitations to enable fair operability. Section 5.1.2 briefly mentions the extent to which this equal treatment takes place.

F_{21}(Fairness; Diversity, non-discrimination and fairness):

Claim: The AI system allows by predicting OHCA deviating process alternatives (different disease progressions).

Evidence (further investigation): Due to the diversity of human complaints and courses in the case of OHCA, the system should offer alternative process steps and instructions to ensure non-discrimination of patients. Whether this is respected in the use case is discussed in Section 5.1.2.

F_{22}(Fairness; Diversity, non-discrimination and fairness):

Claim: Developers are aware of patients and dispatchers needs.

Evidence (further investigation): Requirements for the user interface and its functions and access should be discussed with all stakeholders already in the design and development phase. Possibilities for including different perspectives and thus providing a fair system are explained in Section 5.1.2.

F_{23}(Fairness; Accountability):

Claim: Standards of GDPR are accepted by the audit organization.

Evidence (further investigation): The applicable legal basis, in this case the GDPR, should be accepted by the auditor performing the audit and be fully known to him.

F_{24}(Fairness; Accountability):

Claim: Processes and definitions are protocolled.

Evidence (further investigation): In order to be able to meet accountability in the legal sense as well, the basic data protection regulation applicable to the European Union should be complied with. A summary appraisal for flags $F_{23\text{-}29}$ is given in Section 5.1.2.

F_{25}(Fairness; Accountability):

Claim: Responsibilities and accountabilities are clearly defined.

Evidence (further investigation): All responsible and accountable parties, depending on the risk area, should be identified in the course of risk management and documented to ensure the accountability of the AI system.

F_{26}(Fairness; Accountability):

Claim: Processes and communication outside the organization are defined.

Evidence (further investigation): Furthermore, it should be determined within the risk management which processes and communication methods are to be initiated in case of a contestation of decisions regarding the exchange with patients.

F_{27}(Fairness; Accountability):

Claim: Processes and communication inside the organization are defined.

Evidence (further investigation): Internal processes, contact persons and communication channels should also be clearly defined and documented internally in order to be able to react quickly in the event of a contestation and to ensure accountability.

F_{28}(Fairness; Accountability):

Claim: Risk measures are defined and monitored.

Evidence (further investigation): In order to counteract a contestation in advance, risk indicators (e.g. in feedback rounds) should be defined and continuously monitored. The legal obligation requires, among other things, such risk management.

F_{29}(Fairness; Accountability):

Claim: Stakeholders know about risks, contact people and consequences (e.g. training).

Evidence (further investigation): If risk management takes place without involving all stakeholders, risks, contact persons and possible effects of risks should not only be documented, but also communicated to stakeholders to ensure equal knowledge of accountability.

F_{30}(Explicability; Transparency):

Claim: Documentation of predictions and involved actors.

Evidence (further investigation): All decisions in the diagnosis and first aid process, both those of the ML model and those of all human participants, should be documented to enable traceability and transparency. The degree to which this should be pursued further is assessed in Section 5.2.2.

F_{31}(Explicability; Transparency):

Claim: Documentation of data (quality), recordings and their storage duration.

Evidence (further investigation): In addition, all recorded audio files and the data generated by the call as well as their quality and storage duration should be documented. How this makes processes and predictions traceable and may increase transparency is explained in Section 5.2.2.

F_{32}(Explicability; Transparency):

Claim: Possibilities and boundaries of the AI system are analysed and documented.

Evidence (further investigation): Functions and limitations of the AI system should be analysed, documented and openly communicated to all stakeholders. Possible requirements and implications are outlined in Section 5.2.2.

F_{33}(Explicability; Transparency):

Claim: Stakeholders are aware of possibilities and boundaries of the AI system.

Evidence (further investigation): However, stakeholders should not only be informed of these possibilities and limitations through the use of the AI system, but they should also be explained in an understandable and comprehensible manner. Reference is made to this in Section 5.2.2.

The four ethical principles and their seven key requirements are often interdependent or influence each other. The previously mentioned verified claims, issues and flags therefore also exhibit so-called tensions. Tensions, as explained in Section 2.4.1, are tensions between principles and requirements where underlying values are in conflict. These conflicts should be weighed and, if possible, mitigated. The following tensions, abbreviated as T in the rest of this Section, can be observed in the present use case and are listed and described in the form $T_{1,\dots,n}$(VC/I/$F_{1,\dots,n}$;VC/I/$F_{1,\dots,n}$):

T_1(VC7; F_2): The AI system makes faster predictions vs. the dispatcher is able to manipulate the system:

Adding more buttons that allow going back to previous instructions or invoking help or explanations of the current instruction can strengthen the dispatcher's autonomy and oversight. At the same time, however, this measure may also affect the speed of predictions using AI support in that going back or invoking explanations represent additional process steps that take additional time. Predictions can thus be traced and checked, for example, but instructions that depend on them can consequently only take place later, which can lead to damage, especially in the case of an OHCA.

T_2(VC3; I_6): The AI system respects the GDPR standards vs. the AI system is tested in different contexts:

The system should also be tested in other contexts. However, since the system was developed and tested in the European Union, it is assumed that it is aligned with European legal regulations. For example, it implements the specific legal requirements of the GDPR for data protection. Should it be used in a new context with a different legal provision, it may be necessary to adapt the AI system to legal standards. Compliance with applicable laws in the development of the AI system therefore also limits the reproducibility to this specific development context (without adaptations).

T_3(F_{20}; F_{21}): Easy use of the AI system for every dispatcher regardless of their abilities vs. the AI system allows by predicting OHCA deviating process alternatives:

While implementing alternative process options may best respect and accommodate different human complaints and progressions in the event of an OHCA, it may also affect the simplicity of operation and user interfaces. For example, the dispatcher could be offered choices that lead to different further decisions and instructions. However, this makes the system more complex and the operation of the user interface more difficult.

T_4(I_4; F_{13}): The AI system has a low rate of technical faults, outages, attacks and misuse vs. the AI system does not harm the environment:

A low rate of system failures and consequently a high uptime of the servers is of great importance to be able to guarantee the technical robustness of the AI system. This is particularly necessary in the healthcare sector to prevent damage at any time. However, the continuous availability of the servers and the system also results in higher energy resource consumption. Accordingly, ensuring technical robustness has a negative impact on the ecological environment and its well-being.

T_5(VC5; F_9): The AI system respects the data protection in healthcare vs. the AI system improves the quality of predictions:

The AI system is intended to support and improve the OHCA process by making improved predictions. As explained earlier, human symptoms and courses can vary widely, making a causal explanation difficult even for human physicians. One way to improve the accuracy of the ML model is

to use more data in the training process. However, this is not always possible due to privacy regulations or is limited by the right to privacy. Thus, the availability of more data would improve the accuracy of the AI system, but also affects patient privacy and depends on the applicable privacy regulations.

$T_6(VC_1; F_9)$: The dispatcher is allowed to ignore the AI decision vs. the AI system improves the quality of predictions:

The dispatcher should be able to act autonomously and make their own decisions at any time. Human autonomy should be preserved in accordance with the four principles for a trustworthy AI system when using the system. Since the AI system should thus only serve as a support and the dispatcher is not forced to accept the alert, the accuracy of the ML model only has a limited impact on OHCA prediction. The more emphasis placed on the accuracy of the diagnostic process (i.e., the more frequently the ML decision should be relied upon), the less the dispatcher can act autonomously. Accordingly, the focus on accuracy may affect the autonomy of the dispatcher.

$T_7(F_9; F_{10})$: The AI system improves the quality of predictions vs. the AI system does not replace professionals:

This accuracy of the AI system, as previously described, may lead to dispatchers being expected to rely more on the decisions of the AI system and make fewer decisions of their own. As a result, there is a risk of increasing substitution of dispatchers by the AI system. This can negatively affect both the acceptance of dispatchers per se and the general acceptance of society toward the AI system. A strong focus on the use of the system could be seen as leading to a poor reputation, e.g., due to the loss of jobs. However, since OHCAs currently account for only a small percentage of calls, this is to be observed in the long-term view.

$T_8(VC_3; F_{31})$: The AI system respects the GDPR standards vs. documentation of data (quality), recordings and their storage duration:

In order to enable transparency and traceability (and consequently explainability), it is important to document and keep records of data processing processes, purposes for which the data is used, and how long it is stored, as well as the data itself. The more precise and detailed these records are, the higher the degree of transparency and the better decisions can be traced, interpreted and explained. However, the scope and level of detail also depends on the legal data protection regulations to be complied with. Health data in particular falls under the so-called special category of personal data (Art. 9 GDPR [18]) and requires sensitive handling and compliance with stricter regulations. Transparency and explainability are therefore influenced and in part restricted by the respective data protection law basis.

$T_9(F_8; F_{17})$: The AI system is trained on anonymised data vs. the training data is not biased:

These stricter regulations with regard to the collection, processing and storage of health data may lead to an anonymization of the data basis. Data categories can thus either be omitted entirely or replaced with average values. As a result, individual data points can no longer be assigned to specific categories of analysis, making it difficult to assess fairness and discrimination in the data. For example, data could not be analysed with respect to favoured groups or minorities because the respective data categories may not be collected or processed due to the legal provision. Accordingly, compliance with data protection regulations can influence and affect an assessment of fairness aspects.

$T_{10}(VC_{7;}\, I_{16})$: The AI system makes faster predictions vs. the patient and the bystander are informed about the usage of the AI system:

To ensure the transparency of the AI system, it is important to make patients and callers aware that the AI system is active in the background during the call. Since the conversation takes place exclusively between the dispatcher and the caller, the caller has no knowledge of the use of the system without such information. However, this would take additional time, which in the case of an OHCA would increase the detection time and could have serious consequences. Moreover, the information requirement should not only be weighed against the speed of the AI system, but should also be clarified in terms of legal GDPR principles (this would furthermore include VC_3: The AI system respects the GDPR standards.).

Since the focus of this work is on the two fundamental ethical principles of fairness and explicability (with focus on explainability), the respective issues and flags explained earlier will serve in Section 5 to evaluate the use case with respect to the two principles. In addition to these claims, the following data analysis of true positives will support the evaluation.

4.2 Data Analysis

In addition to the analysis of the use case with the help of CAE and ALTAI, the data basis is examined and evaluated in the following Section with regard to ethical issues and flags. The ML model is a black box model, which is not accessible for this evaluation. Furthermore, since the database only contains true positives (positive cases correctly classified by the ML model, see Section 2.2.4), no comparison between correctly and incorrectly predicted cases of the AI system and the underlying features is possible. Therefore, the evaluation of the requirements is performed via a Python-based data analysis focusing on general feature importances. It starts by describing the data basis and its preparation, followed by an explanation of the analysis.

4.2.1 Data processing

The database available to us includes all True Positive cases of the RCT (959, see Section 3.4), which is why only calls that were correctly classified as OHCA by the ML model are included. It does not include cases in which there was no OHCA (correctly or incorrectly predicted by the ML model), nor cases that were incorrectly predicted as OHCA by the ML model. The database contains dimensions such as the particular group of dispatchers (control group or intervention group), the gender of the patient or caller, their relationship to each other (e.g. relative or healthcare professional), the age of the patient, the environment from which the call was made or symptoms that occurred, as well as metrics such as the duration of the call, the time to recognition of OHCA (differentiated by ML model and dispatcher), the time at which the alert was displayed (exclusively for intervention group), or the time to start CPR instructions. In order to be able to analyse the data (see Section 4.2.2), data cleaning and transformation is performed at the beginning. Both the processing of the data and the analysis take place with the help of the programming language Python and the web-based application Jupyter Notebook (see Appendix H). The following Section explains the data processing steps and provides an overview of the database used for the analysis.

Data cleaning: Due to missing column descriptions in terms of content and column names (abbreviations) that were not comprehensible, assumptions were made at the beginning and noted in the notebook to enable comprehensibility of the database. Furthermore, various lexical errors

were present, which were also corrected and commented by us. Columns about which we could not make assumptions, or which were not relevant to the subsequent analysis were deleted. Similarly, data entries that would bias the analysis were deleted identical to the RCT. The deleted entries, as shown in Figure 8, include calls with a broken chain (21 entries), calls for which an OHCA was witnessed by EMS (34 entries), and calls for which a CPR had already been performed (250 entries).

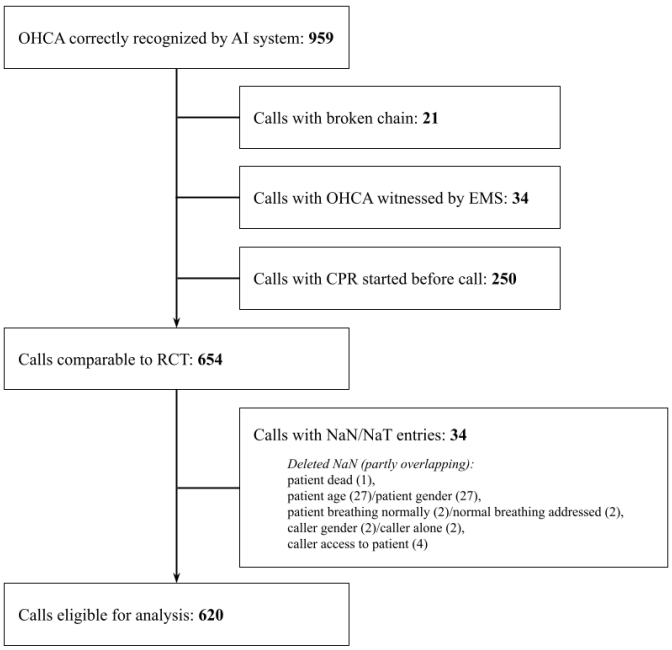


Figure 8: RCT data cleaning for further evaluation.

Furthermore, all calls containing NaN or NaT in dimensions or metrics were either deleted (see Figure 8) or, if possible, filled with the entries False, None, not defined or 0. This cleaning of the data resulted in a total of 620 calls relevant for the analysis, as shown in Figure 8.

Data transformation: After cleaning the data, the transformation of column names and data types followed, as well as the change of the order of the column arrangement. This improved the understanding and facilitated later comparisons of dimensions and metrics. Furthermore, two additional columns *Recognition_dispatcher* and *age_group* were added.

Recognition_dispatcher: This column compares the length of the call with the time required for the dispatcher to recognise the OHCA. It defines the OHCA as not recognised by the dispatcher in case of a matching duration.

age_group: In order to be able to evaluate the age of the patients, calls were assigned to different age groups via the respective age of the patient in the column age_group. Four groups were formed in accordance with [141], with age groups under 35 years combined due to the small sample. The distribution of calls among the different age groups can be seen in Table 13:

	Number of calls		
Age group	**Control group (N=314)**	**Intervention group (N=306)**	**Total (N=620)**
<= 35	13	8	21
36-64	89	85	174
65-79	133	111	244
>= 80	79	102	181

Table 13: Distribution of calls (cleaned data) per age group.

The data set was then split into two subsets depending on the group: The control group subset now contained 314 calls, while the intervention group subset contained 306 calls. This was necessary to compare the predictions of the two groups and the influence of the AI system. In the subset of data for the intervention group, the column *accept_MLalert* was also added, which tests whether the dispatcher accepts the ML alert or not. It is assumed that the dispatcher takes up to five seconds of response time to process the ML alert. Thus, the dispatcher takes the same amount of time or up to five seconds longer than the AI system to detect the OHCA, the dispatcher is assumed to accept the alert. Otherwise, the alert is considered ignored.

Overview of the database: The cleaned and transformed dataset now contains 620 calls with corresponding dimensions and metrics in a total of 60 different columns. The modified data set leads to a difference in the comparative values used in the RCT, which compares the control and intervention groups. The database used for the analysis that starts in the next Section generally has the following relations (for more details see Appendix I), which are consistent with the RCT's [20] relations despite different values:

a. The data set includes mostly patients of older age in both groups (control group: 69.1 mean, 16.3 sd [RCT: 69.1 mean, 16.1 sd]; intervention group: 71.4 mean, 15.6 sd [RCT: 71.4 mean, 15.6 sd]).
b. The data set includes more male than female patients in both groups (416 vs. 204 [RCT: 419 vs. 235]) and more female than male callers (427 vs. 193 [RCT: 447 vs. 207]).
c. The data set includes more calls from relatives and family members of the patient in both groups (340, 54.84% [RCT: 345, 52.75%]).
d. The data set includes more calls from bystanders in both groups, which are close to the patient (561, 90.48% [RCT: 584, 89.30%]).
e. In both groups, the data set includes more calls originating from the patient's private environment (524, 84.52% [RCT: 539, 82.42%]).

Due to the change in the data basis (deletion of NaN and NaT), the confusion matrix also differs compared to Section 3.4 as follows (see Table 14): The intervention group includes 306 calls, of which 281 cases of OHCA are detected. In the control group, 282 OHCA cases are detected out of a total of 314 calls. Accordingly, most of the calls are detected as OHCA by the dispatchers regardless of the support provided by the AI system.

	Control group (alert not shown)	Intervention group (alert shown)	
Dispatcher recognises OHCA	282	281	**563**
Dispatcher fails to recognise OHCA	32	25	**57**
	314	**306**	**620**

Table 14: Division of recognised calls (cleaned data) by group of dispatcher.

The analysis of the data that follows in the next Section is based on this edited database with 620 entries. It should be emphasized that the representativeness of the results must be questioned in cases of unequal distribution of the dimensions. In particular, subgroups with few data points can be observed for various dimensions like accept_MLalert (True), age_group (<= 35) or area_call (Natural area).

4.2.2 Data evaluation

The analysis and final evaluation of the data represents Step IV: Execution of the Assess-Phase of the Z-Inspection® process. The analyses carried out are described in this Section. In the course of the analysis, the patient and call(er) characteristics are dealt with in more detail, which are further explained in Appendix J.

Cardiac arrest is life-threatening and rapid detection - even before the ambulance arrives - and thus rapid initiation of life support measures is essential for the patient's survival [136]. Because the data we have only contains the true positives (TP), there is no other way than to examine the model in terms of the speed of OHCA recognition time, because with the True Positive data it is not possible to find out under which circumstances the ML model classifies OHCA correctly or incorrectly. This means that the metrics described in Section 2.2.5 cannot be calculated and no judgement can be made about possible biases and discriminations in this form. In this work and with the data available, we have decided to compare the OHCA recognition time of the ML model and that of the medical dispatchers and compute the feature importances (see Section 2.3.5) according to the following categories in order to find possible explanations, biases and counterevidence for the ethical issues and flags from Section 4.1.3:

- *Comparison of the ML model's OHCA recognition time in principle with that of the medical dispatchers.*
- *Comparison of the ML model's OHCA recognition time regarding patient and call(er) characteristics.*
- *Further comparisons of the ML model's OHCA recognition time of two patient and call(er) characteristics combined.*

In the following, the comparisons of the three categories are presented, possible explanations, biases and counter-evidence are elaborated, which are examined and evaluated in Section 5.

Comparison of the ML model's OHCA recognition time in principle with that of the medical dispatchers: The ML model listens in on all calls, regardless of whether the call is randomly assigned to the control group (N=314) or the intervention group (N=306). Therefore, an OHCA recognition-time-comparison of the entire true positive sample (after cleaning, N=620) and the control group

(N=314) is useful to analyse the performance in the next steps. In the control group, an influence of the ML model on the decision and OHCA recognition time can be ruled out, which is why the control group is suitable for the comparison. The comparison of the whole sample (N=620) with the control group (N=314) shows that the ML model detects an OHCA in mean about 36 seconds faster than a medical dispatcher (80.10s mean, 83.38s sd vs. 116.31s mean, 108.85s sd). Comparing the time to OHCA recognition between the ML model and the medical dispatchers only within the control group, there is a time difference of app. 38 seconds (78.24s mean, 87.46s sd vs. 116.31s mean, 108.85s sd). The difference is therefore two seconds larger. It should be noted that the OHCA recognition times compared last are both from the control group (same sample size), but there is more data available to analyse the recognition time of the ML model, which is why the focus is on the entire sample.

Nevertheless, out of 306 calls of the intervention group, the alert is only accepted 27 times (8.82%) by the medical dispatchers. For this reason, and because the focus of this master's thesis is on the ethical principles of fairness and explicability, the following analysis does not concentrate on the extent to which the AI system influences the dispatchers (question of the RCT [20]), but mainly examines possible biases and their explanation of the ML model.

The fact that the ML model identifies an OHCA faster runs through all tested features, with the only exception of the area characteristic. Here, the medical dispatchers - if the call comes from a traffic area - needed a lower OHCA recognition time than the ML model (see Table 15, mean time, 101.25s mean, 93.81s sd vs. 97.18s mean, 80.29s sd). The reason for this may be a degraded speech quality due to traffic-related background noise, but the absolute sample size for both groups is small. Otherwise, there are no noteworthy differences in the comparison of the OHCA recognition time of the medical dispatcher and the ML model, which is why the mean values of the OHCA recognition time of the dispatcher are included for the sake of completeness in the further category comparisons with regard to the patient and call(er) characteristics but are no longer taken into account in the comparison. In the comparison between the times of the dispatchers for different characteristics of the calls, biases of the dispatchers can also be determined, but the focus here is on possible biases and explanations of the ML model.

Comparison of the ML model's OHCA recognition time regarding patient and call(er) characteristics: Table 15 compares the various patient and call(er) characteristics considered in the course of the master's thesis of both the ML model and the medical dispatcher. As mentioned before, only the comparisons within the features of the ML model are discussed in the description.

In order to find out whether and which patient and call(er) characteristics have a strong influence on the faster OHCA recognition time of the ML model, the SHAP[8] Value (see Figure 9) - a method of explainability (see Section 2.3.5, and 5.2.1) - was used. This analysis showed that the features *dispatcher_assertive_passive, area_call, age_group, patient_conscious, patient_gender and caller_relation* have an increased influence (> 5) on the OHCA recognition time (see Section 5.2.1 for a more detailed description of the feature importance analysis). Furthermore, our own examination of the data revealed that the symptoms *Asystoli* (=stop of heart interval action), *VF* (assumption: VF=Ventricular Fibrillation), *PEA* (assumption: PEA=Pulseless Electrical Activity) occur

[8] Package: shap, version: 0.39.0.

most frequently in the data set. Therefore, these characteristics are also worth looking at and are discussed here. For reasons of overview, the other characteristics of the data are not considered further.

	ML model	**Control group**
	mean, sd in seconds	**mean, sd in seconds**
Unrecognition time		
total	80.10, 83.38 (N=620)	116.31, 108.85 (N=314)
patient_gender		
female	82.30, 89.02 (N=204)	126.71, 106.50 (N=100)
male	**79.03, 80.56 (N=416)**	111.45, 109.85 (N=214)
age_group		
<=35	74.94, 50.01 (N=21)	76.87, 44.34 (N=13)
36-64	83.03, 84.33 (N=174)	114.18, 104.27 (N=89)
65-79	**79.98, 89.70 (N=244)**	119.00, 117.55 (N=133)
>=80	**78.06, 76.91 (N=181)**	120.67, 106.30 (N=79)
caller_relation		
Healthcare professional	**75.27, 79.01 (N=162)**	120.90, 122.63 (N=73)
Other	89.22, 77.30 (N=118)	113.21, 96.57 (N=61)
Relative	79.25, 87.32 (N=340)	115.50, 107.39 (N=180)
area_call		
Other area	93.13, 83.00 (N=46)	133.33, 124.59 (N=30)
Private area	**77.19, 82.73 (N=524)**	115.64, 109.91 (N=257)
Natural area	88.42, 66.27 (N=10)	132.60, 55.70 (N=5)
Traffic area	101.25, 93.87 (N=40)	97.18, 80.29 (N=22)
patient_breathing_normally		
No	**78.27, 79.93 (N=596)**	114.16, 106.33 (N=303)
Yes	125.67, 139.95 (N=24)	175.51, 159.52 (N=11)
patient_conscious		
No	**76.64, 76.31 (N=597)**	112.10, 100.24 (N=304)
Yes	170.11, 170.45 (N=23)	244.29, 234.55 (N=10)
caller_emotional_distressed		
No	**79.69, 80.99 (N=417)**	114.92, 100.80 (N=209)
Yes	80.96, 88.28 (N=203)	119.07, 123.81 (N=105)
dispatcher_assertive_passive		
Assertive	**76.70, 86.33 (N=404)**	99.09, 104.95 (N=200)
N/A	90.47, 78.19 (N=131)	152.60, 116.95 (N=71)
Passive	80.33, 76.07 (N=85)	136.47, 96.24 (N=43)
caller_gender		
female	80.61, 84.50 (N=427)	119.20, 115.61 (N=213)
male	**78.98, 81.05 (N=193)**	110.23, 93.25 (N=101)
caller_alone		
No	**78.47, 77.95 (N=325)**	117.93, 114.75 (N=172)
Yes	81.90, 89.08 (N=295)	114.35, 101.62 (N=142)
caller_access_patient		
Patients side	80.73, 84.59 (N=561)	116.82, 110.80 (N=287)
Access	**69.99, 49.96 (N=38)**	112.03, 91.63 (N=21)
No Access	81.76, 99.97 (N=21)	106.94, 75.75 (N=6)
symptom		
Asystoli	**70.68, 79.11 (N=353)**	103.91, 99.19 (N=179)
PEA	111.23, 99.97 (N=92)	160.53, 129.11 (N=42)
VF	88.80, 87.38 (N=105)	112.77, 102.68 (N=55)
Other	73.66, 62.01 (N=70)	131.00, 125.86 (N=38)

Table 15: Comparison of patient and call(er) characteristics. The first value shows the mean and the second value the standard deviation. The values that will be examined in more detail are marked in bold.

In the following, the patient and call(er) characteristics with mean(|SHAP value|)>5 are presented with regard to the OHCA recognition time of the ML model and explained on the basis of Table 15 with regard to abnormalities and possible fairness and explainability aspects (see Sections 2.2 and 2.3).

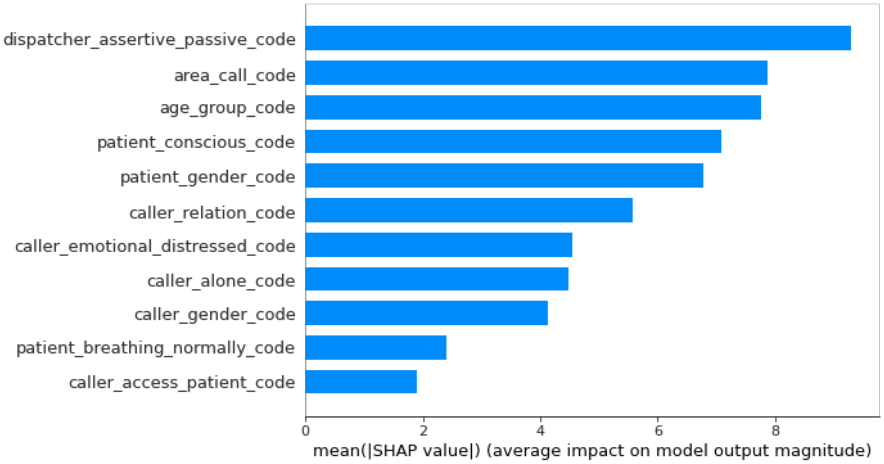


Figure 9: Most important features (SHAP values) for comparison by recognition time.

dispatcher_assertive_passive: Whether or not a dispatcher takes a passive or active/assertive role in the emergency call has an impact on the OHCA recognition time of the ML model. In case the dispatcher is active and determined, the recognition time is shorter compared to a passive dispatcher (76.70s mean, 80.99s sd vs. 80.33s mean, 76.07s sd).
Fairness*:* The shorter recognition time of the ML model with an active and determined dispatcher may represent a representation bias (see Section 2.2.3), as the majority of dispatchers take an active role. The ML model is dependent on the dispatcher in this respect, as it has no interaction of its own with the patient. However, the dispatcher's behaviour depends on many factors such as his personality. This represents a human bias, but not a bias of the ML model.
Explainability: An explanation for this could be given post-hoc in text form through interpretation (see Section 2.3.2). The fact that the dispatcher uses active formulations could eliminate additional (feedback) questions and answers. Consequently, this would also have an impact on time. Since the ML model (based on NLP [137]) evaluates the communication between dispatcher and caller, relevant information could be processed faster, which can lead to faster recognition.

area_call: It also matters from which area the call is made. If the emergency call is made from a private environment, the ML model only needs 77.19s in mean (82.73s sd) to identify an OHCA. If, on contrast, the call comes from a traffic area, the ML model requires 101.25s in mean (93.87s sd). In this special case, the average OHCA recognition time of the dispatcher is once faster than that of the ML model.
Fairness: The shorter OHCA recognition time of the ML model in the private environment may represent a representation bias, as 84.52% of the calls are from private environments.
Explainability: An explanation could take place through algorithmic transparency (see Section 2.3.2). Since the ML model is based on NLP [137], poor audio quality could lead to speech not being processed correctly. This occurs especially in traffic and natural areas, where noise may be present. This would justify a longer prediction time of the ML model.

patient_gender, age_group: Comparing the ML model's time to OHCA recognition by gender, it is noticeable that the ML model identifies OHCA faster in male patients than in female patients (79.03s mean, 80.56s sd vs. 82.30s, 89.02s sd). In order to compare OHCA recognition times based on age, the sample was clustered into age groups as described in Section 4.2.1. The ML model

identifies OHCA within 74.94s in mean (50.01s sd) for patients up to 35 years of age. Except for this age group, the OHCA recognition time of the ML model decreases with increasing patient age.
Fairness: The increased number of patients with an increase in age and male patients and the resulting faster OHCA recognition times suggest a representation bias.
Explainability: The result of a faster OHCA recognition time in men and patients of older age can be justified post-hoc via explanations by example (see Section 2.3.2) and possible causal human relationships. According to [141], the risk of cardiac arrest is higher in men and increases with age from 35 years. This fact is supported by both the mean times and the absolute distributions of the data. 67.1% of the patients in our study were male. Furthermore, the number of patients increases with the age of the patients, see Table 15.

patient_conscious: A positive influence on OHCA recognition is when the caller reports the patient's unconsciousness to the dispatcher. In this case, the ML model identifies OHCA about 164s in mean faster than when the caller reports that the patient is conscious.
Fairness: This may also be a representation bias, as 96.29% of patients are addressed as not conscious. However, it may also be due to the fact that unconsciousness is a strong indication of OHCA, which is why the NLP classifies according to it. But the fact that the caller may give false information about the consciousness status results in a disadvantage for the patient who is falsely declared to be conscious. This may represent a sampling bias (see Section 2.2.3), as NLP may be trained on unconsciousness and not easily transferable to patients who are reported as conscious.
Explainability: Ones the caller expresses that the patient is unconscious, the ML model recognises the OHCA more quickly. Therefore, algorithmic transparency (see Section 2.3.2) could be included in the explanation. It is possible that the ML model could have been trained on this feature, leading to faster detection.

caller_relation: The ML model is faster than the mean total (80.10s mean, 83.38s sd) if the caller is a healthcare professional (75.27s mean, 79.01s sd) or a related person (79.25s mean, 87.32s sd).
Fairness: The examination of the absolute distribution shows that the ML model identifies the OHCA fastest for a healthcare professional even though the sample is smaller than for a relative. Therefore, a representation bias can be ruled out. However, it is possible that the NLP model was developed on the vocabulary of a healthcare professional and classified according to certain medical terms. Then this can also represent a sampling bias.
Explainability: An explanation based on algorithmic transparency (see Section 2.3.2) could be given here. If the ML model has been trained on medical vocabulary, the system could filter out technical terms, symptoms and other indications that point to OHCA and thus predict OHCA more quickly. In addition, another interpretation would be that medical professionals are trained in dealing with OHCA and can already make their own assessments, which the ML model processes.

Symptom: Analysis of the data for specific symptoms associated with OHCA revealed that the symptoms Asystoli, VF and PEA occur most frequently. Asystoli is detected by the ML model within 70.68s in mean (79.11s sd). The OHCA recognition time by the ML model for a PEA, on the other hand, takes about 40s (mean) longer.
Fairness: The most common symptom is Asystoli. In 56.93% of cases, patients suffer from Asystoli. The faster recognition in this case indicates a representation bias. A sampling bias that the algorithm of the NLP was developed on the symptom Asystoli cannot be excluded. Despite the -

smaller but not negligible - sample of VF (N=105) and PEA (N=92), OHCA recognition time is longer for these symptoms. This also suggests a sampling bias.
Explainabiliy: Whether and to what extent certain symptoms Asystoli, are associated with OHCA could be explained with post-hoc textual explanation (see Section 2.3.2) about causal human relationships. This requires further medical assessment. If Asystoli frequently occurs before an OHCA, this could account for the faster detection time of the ML model, which may have been trained to detect these signs. This in turn could be explained by algorithmic transparency (see Section 2.3.2).

Whether the biases are confirmed when two patient and call(er) characteristics are considered in combination, such as the relationship between a healthcare professional in a particular age group, will be explained in the next step.

Further comparisons of the ML model's OHCA recognition time of two patient and call(er) characteristics combined. The individual patient and call(er) characteristics have been examined and the OHCA recognition times of the ML model have been compared based on various characteristics. Notable combinations of features are shown below. For reasons of overview, not all analyses are explained here. The OHCA recognition times of the dispatchers are not considered here as well. The following tables show the mean of the OHCA recognition time of the ML model and the standard deviation in seconds.

caller_relation x dispatcher_assertive_passive: Table 16 shows the combined OHCA recognition time of the ML model of the features *caller_relation* and *dispatcher_assertive_passive*. It can be seen that the combination of the two fastest features (assertive and healthcare professional) leads to an even faster OHCA recognition time (69.50s mean, 75.73s sd). Although the sample here is also rather large in comparison with 92 calls, there are more calls answered by an assertive dispatcher from a relative (N=226).

caller_relation dispatcher_assertive_passive	Healthcare professional	Other	Relative
Assertive	**69.50, 75.73 (N=92)**	88.26, 79.97 (N=86)	75.22, 92.41 (N=226)
N/A	75.50, 67.33 (N=49)	109.41, 86.23 (N=14)	97.36, 82.97 (N=68)
Passive	99.99, 111.77 (N=21)	78.06,54.75 (N=18)	72.25, 62.19 (N=46)

Table 16: Feature comparison caller_relation x dispatcher_assertive_passive (mean, sd in seconds). The value that will be examined in more detail is marked in bold.

Fairness: The two conditions healthcare professional and an assertive dispatcher have a positive influence on the OHCA recognition time of the ML model. Certainly, the relatively high sample (representation bias) has an influence on this, but the sample is higher for relatives. Therefore, as described for the feature *caller_relation*, it can be assumed that the ML model is trained for medical terms and classifies them accordingly. Furthermore, the dependence of the ML model on the dispatcher should be mentioned here.
Explainability: By combining the exemplary explanations of the two features given earlier, an interpreted connected explanation could be the following: Healthcare professionals already have medical expertise. Instructions can thus be given more quickly and accurately from both sides

(dispatcher and caller). Additional questioning might so not be necessary. In addition, the instructions and answers could correspond to the trained vocabulary.

caller_relation x age_group: The OHCA recognition time with the feature *caller_relation* and the *age_group* is shown combined in Table 17. The faster OHCA recognition time if the caller is a healthcare professional and the decreasing OHCA recognition time with increasing age (except for the small sample of under 36 years old) reinforce each other. Both the absolute sample and the OHCA recognition time with increasing age increase and decrease respectively when the caller is a healthcare professional. For a patient over 80 years of age, for whom the emergency call is placed by a healthcare professional, the average OHCA recognition time is 68.95s (51.24s sd). This is shorter than the recognition time of the healthcare professionals (75.27s mean, 79.01s sd) and the over 80 years old (78.96s mean, 76.91s sd).

caller_relation age_group	Healthcare professional	Other	Relative
<=35	34.05, 33.00 (N=2)	131.39, 41.85 (N=7)	48.83, 20.41 (N=12)
36-64	91.79, 109.04 (N=18)	86.00, 67.34 (N=43)	80.50, 86.36 (N=113)
65-79	79.12, 94.99 (N=67)	99.43, 96.22 (N=47)	73.40, 84.01 (N=130)
>=80	**68.95, 51.24 (N=75)**	58.89, 41.97 (N=21)	90.82, 98.05 (N=85)

Table 17: Feature comparison caller_relation x age_group (mean, sd in seconds). The value that will be examined in more detail is marked in bold.

Fairness: Here, too, it is confirmed that the algorithm of the ML model was trained on the healthcare professional condition and classifies according to medical terms. This reinforces the assumption of a sampling bias.
Explainability: In this case, an explanation of the shorter detection time among older age groups and healthcare professionals could be given by combining the explanations of the individual features and interpreting them. For example, it could be assumed that with increasing age it becomes more likely to have regular visits from caregivers. At the same time, older age makes it more likely to have an OHCA. Caregivers could also be trained in OHCA and thus use medical vocabulary when calling.

area-call x age_group: OHCA recognition time is faster both at increasing age and in a private environment than at younger patient age and in outdoor environments. For calls that meet both conditions (older age and private environment), OHCA recognition time is again faster than when looking at the single feature.

Fairness: The amplification of the faster OHCA recognition time of the ML model of the single observation (see Table 18) and the observation of the absolute numbers, support the assumption of the representation bias with regard to older patients.
Explainability: The combination of the individual feature explanations here leads to the assumption that the ML model works particularly quickly and also accurately - a large proportion of the true positives are patients of a higher age in a private environment - on calls without background noise and with older patients.

area_call age_group	Other area	Natural area	Private area	Traffic area
<=35	135.70, 32.58 (N=3)	141.38, NaN (N=1)	58.18, 42.54 (N=16)	94.35, NaN (N=1)
36-64	108.47, 95.49 (N=16)	53.51, 35.69 (N=4)	83.82, 86.63 (N=140)	54.49, 40.56 (N=14)
65-79	85.26, 93.58 (N=17)	125.46, 79.13 (N=4)	**74.72, 88.85 (N=205)**	124.81, 89.27 (N=18)
>=80	69.17, 40.79 (N=10)	27.36, NaN (N=1)	**76.46, 73.92 (N=163)**	135.18, 151.76 (N=7)

Table 18: Feature comparison area-call x age_group (mean, sd in seconds). The values that will be examined in more detail are marked in bold.

patient_conscious x symptom: As shown in Table 19, an unconscious patient has the symptom Asystoli in 57.45% of the cases. Excluding the very small and negligible sample of two calls that were not unconscious and had the symptom VF and excluding the characteristic Other, the ML model is faster in identifying the OHCA associated with unconsciousness and Asystoli with 68.87s in mean when looking at the individual features (unconsciousness and Asystoli).

symptom patient_conscious	Asystoli	PEA	VF	Other
No	**68.87, 74.09 (N=343)**	100.81, 81.04 (N=88)	89.67, 88.00 (N=103)	63.86, 46.16 (N=63)
Yes	133.00, 178.19 (N=10)	340.43, 198.67 (N=4)	43.89, 10.66 (N=2)	161.86, 110.12 (N=7)

Table 19: Feature comparison patient_conscious x symptom (mean, sd in seconds). The value that will be examined in more detail is marked in bold.

Fairness: Looking at the absolute numbers, even after considering the combination of these two characteristics, we can conclude that there is a representation bias in the data. The assumption of a sampling bias in the algorithm is also reinforced.
Explainability: If both unconsciousness and Asystoli are signs of OHCA, an explanation could be provided by interpreting medical interrelations. Possibly, the two features themselves could also show a correlation. Such a correlation could also be found in the training model, which could be explained by an algorithmic transparency (see Section 2.3.2).

symptom x age_group: Considering the symptoms in the context of the age group, it can be seen (Table 20) that the symptom Asystoli in particular is more likely to occur in the two older age groups. In general, when an Asystoli occurs, the OHCA recognition of the ML model is comparatively fast. For patients with an age over 79 years, the ML model needs 59.78s in mean (49.72s sd) to identify an OHCA, taking into account a relatively large sample.

Fairness: Due to the large sample of patients suffering from Asystoli and in combination with older patients, a representation bias can still be assumed here. A look at the OHCA recognition times and absolute numbers for the symptoms PEA and VF suggests a sampling bias in the algorithm. Presumably, the ML model classifies according to the symptom Asystoli.
Explainability: Here, too, an explanation could be given via the combination of possible causal relationships of the individual characteristics. In this case, not only is the recognition time faster

with the presence of Asystoli and higher age, but the ML model could generally work more accurately with these two feature expressions (largest proportion of true positives). To confirm the latter assumption, an examination of all data and their classifications is needed.

symptom age_group	Asystoli	PEA	VF	Other
<=35	63.48, 33.31 (N=15)	123.45, 103.38 (N=2)	108.89, 82.73 (N=2)	78.41, 89.06 (N=2)
36-64	78.43, 86.46 (N=84)	85.63, 68.62 (N=15)	92.26, 93.87 (N=53)	92.26, 93.87 (N=22)
65-79	74.93, 93.77 (N=147)	107.62, 98.10 (N=37)	74.31, 64.72 (N=33)	76.56, 77.87 (N=27)
>=80	**59.78, 49.72 (N=107)**	124.21, 112.64 (N=38)	103.75, 106.89 (N=17)	65.69, 33.15 (N=19)

Table 20: Feature comparison symptom x age_group (mean, sd in seconds). The value that will be examined in more detail is marked in bold.

Overall, it can be said that the faster OHCA recognition time of individual features in combination with another faster feature results in the mean OHCA recognition time being lower than for the single features. This shows that the faster features strengthen together. The fairness and explanation aspects mentioned above are complemented below, taking into account the issues and flags mentioned in the Section 4.1.3.

5 Findings and recommendations

In order to finally evaluate the ML model of the use case with regard to fairness and explicability, the ethical issues and flags presented in Section 4.1.3 are used as a guideline. The verified claims are not considered here, because the evidence for these claims is already documented in Section 4.1.3. The same applies to the ethical issues and flags that deal with the ethical principles of respect for human autonomy and prevention of harm. The evaluation makes use of the analyses from Section 4.2.2 and further the available information from workshops of the international research group and the published papers [20],[136],[137].

5.1 Fairness

As the available data only covers the true-positive cases, it is not possible to apply the fairness metrics and mitigation algorithms described in Section 2.2.5. In addition, the data set only includes 620 data points. Therefore, although statements are made about possible biases, they always take into account the small data set.

5.1.1 Ethical issues related to the ethical principle of fairness

I_9(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The dispatcher is trained on the AI system.

Evidence: The dispatcher is not sufficiently trained in dealing with the AI system. This negatively affects the dispatcher's trust and acceptance of the AI system [20]. Of the 306 analysed calls belonging to the intervention group, only in 27 cases the medical dispatcher accepted the alert (see Section 4.2.2). This suggests that there is insufficient acceptance of the AI system and that the dispatchers have not been adequately trained and informed. Furthermore, the OHCA recognition times of the ML model show that the model relies on the dispatcher taking an active role in the emergency call. For an assertive dispatcher, the ML model takes 76.70s in mean (86.33s sd) for OHCA detection, whereas for a passive dispatcher it takes 80.33s in mean (76.07s sd). The dispatcher should be made aware of this dependency to increase acceptance.

I_{10}(Fairness; Diversity, non-discrimination and fairness):

Claim: The audio files are not biased.

Evidence: Since only calls from Denmark in Danish are used for training and testing, the AI is consequently only trained on the Danish language [137]. The sampling bias defined in 3.2.1 may apply here. In this case, the algorithm is trained on the population in Denmark. Both the language and possible characteristics of the OHCA are trained in Denmark and Copenhagen respectively. However, according to Corti, research is being conducted to be able to apply the AI system to other languages as well. Nevertheless, the sampling bias with regard to possibly existing different characteristics of OHCAs from different populations must also be considered here.

I_{11}(Fairness; Diversity, non-discrimination and fairness):

Claim: The AI system offers alternatives for disabled dispatchers (e.g., auditive usage for blind dispatchers).

Evidence: The AI system only offers a form of use via a static user interface. It therefore does not offer the possibility of full accessibility [140]. To train as a medical dispatcher, special training in first aid is required. There is no explicit requirement that the medical dispatcher should be completely healthy, but it can be assumed that blind people, for example, do not have the possibility to be trained as a medical dispatcher.

I_{12}(Fairness; Diversity, non-discrimination and fairness):

Claim: The AI system is developed with patients and population participation.

Evidence: The AI system was developed without the influence of patients. Both the acceptance of patients and the population as well as the inclusion of all perspectives can thus be ruled out [20]. This inclusion would be important not only for the sake of fairness, but also to increase acceptance of the ML model.

5.1.2 Flags related to the ethical principle of fairness

F_9(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system improves the quality of predictions.

Evidence (further investigation): The results of the intervention group show a significantly higher sensitivity in both studies. The sensitivity is a key ratio to measure the quality of an AI system (see Section 2.2.4). Accordingly, the prediction of the OHCA is better in the AI system support scenario. However, the evaluation bias defined in Section 2.2.3 applies here: Which indicator is used as the basis for evaluation is a bias in itself. For when looking at the indicators of specificity and PPV, there is no significant difference between the intervention and control group (see Section 3.4). Therefore, different aspects of the quality of the system should be considered as far as possible and it should be weighed up according to which key ratio such an evaluation will take place.

F_{10}(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system does not replace professionals.

Evidence (further investigation): In the short term, it is unlikely that the AI system will replace the human dispatcher. As explained in 4.2, the AI system is classified as assisted intelligence and is intended to support medical dispatchers. This division of labour should be openly communicated. It should be made clear that it is only a support to the diagnostic process in order to maintain the acceptance of the dispatchers and consequently create a social well-being.

F_{11}(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system has no impact on dispatchers training.

Evidence (further investigation): At the moment there is no consideration to fundamentally change the training of dispatchers. However, this should be done if the ML model is used continuously. As mentioned in I_9, the ML model is dependent on the active role of the dispatcher (see Section 4.2.2). Training on how to encourage an active role in the emergency call would be useful.

F_{12}(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system gets promoted in terms of acceptance and transparency.

Evidence (further investigation): So far, there are published scientific papers in journals on the two studies that describe the use of the AI system so far, but these are not intended for the mainstream population and are not easy to understand. Furthermore, there is a video that explains the process of the AI system [138]. At the latest before the permanent use of the AI system in practice, the functions and limits of the AI system should be communicated more intensively to the population in order to establish their acceptance and trust as well as a general social well-being.

F_{13}(Prevention of harm/Fairness; Environmental and societal well-being):

Claim: The AI system does not harm the environment (Green IT, ecological balance).

Evidence (further investigation): The information available to us does not provide any findings on this flag. For example, drawing up a life cycle assessment could help to monitor ecological goals, measures and environmental sustainability. Further, green electricity (e.g. through solar panels [142]) should be considered.

F_{14}(Fairness; Diversity, non-discrimination and fairness):

Claim: The developer is not biased.

Evidence (further investigation): No information is available on whether and to what extent the developers are biased. Nevertheless, no human being is exempt from certain biases and it is natural not to practise complete neutrality. Biases of the developers can show up in the algorithm itself as well as in the choice of evaluation methods and algorithms (Evaluation bias, see Section 2.2.3). However, these should of course be as insignificant as possible and possibly be reduced by the cooperation of several developers.

F_{15}(Fairness; Diversity, non-discrimination and fairness):

Claim: The client (e.g. customers, hospitals commisioners) is not biased.

Evidence (further investigation): In addition to the developers, the other stakeholders, such as the clients, should also be checked for their objectivity. They too could influence the development of the algorithm. No concrete information is available on this. It can be assumed, for example, that hospital commissioners, who want to keep the costs of ambulance missions low, are interested in sending as rarely as possible high-priority ambulances (including emergency physicians). Such interests may influence the algorithm. In the study [20], more high-priority ambulances were dispatched, which may prompt stakeholders to change the algorithm to reflect it.

F_{16}(Fairness; Diversity, non-discrimination and fairness):

Claim: The algorithm is not biased.

Evidence (further investigation): As described in Section 3.2, the ML model is a black box model. For this reason, it is fundamentally difficult to explain the decisions (see Section 2.3.2). The biases in algorithms described in Section 2.2.3 are applied to the ML model at this point:

Evaluation bias: As already described in F_9 and F_{14}, the algorithm can be biased, on the one hand by the developer but on the other hand also by the choice of methods and evaluation algorithms used. According to the information provided by Corti, the ML model is tested against biases and these are then mitigated. According to Corti, sampling strategies, augmentations and the addition of data points were

used for this purpose [137]. According to [136], the development team aimed to achieve the highest possible sensitivity. There is a bias at this place.

Emergent Bias: Even this model is not immune to emergent bias once it is permanently implemented in practice. It is important that new scientific knowledge regarding OHCA recognition is continuously incorporated into the AI system.

Sampling Bias: As described in I_{10}, the sampling bias also applies in this case. The algorithm is trained on the population in Denmark/Copenhagen. Furthermore, a sampling bias can be inferred for the patient and call(er) characteristics patient_conscious, symptom, since the NPL is classified according to unconsciousness, and Asystoli and thus the algorithm was developed accordingly.

F_{17}(Fairness; Diversity, non-discrimination and fairness):

Claim: The training data is not biased.

Evidence: In Section 4.2.2, the available true positive data were analysed for possible biases (see Section 2.2.3). As already mentioned, the only basis for analysis is the comparison of OHCA recognition times for different patient and call(er) characteristics and the consideration of the sample size for the different features. However, it is important to mention again here that the data set available with 620 cases that can be analysed does not necessarily allow to make representative statements.

Historical Bias: In the given use case, the sample is biased in that the ML model only analyses calls from Copenhagen/Denmark (sampling bias, see I_{10}, F_{16}). However, in our view, the sample for the population in Denmark is random and without active selection of a specific sample, since all incoming calls [20] are evaluated and none are excluded (except due to technical errors).

Representation Bias: When looking at the sample size and OHCA recognition time, the characteristics patient_gender (underrepresentation of women), area_call (most calls are from a private environment), patient_breathing_normally (96.12% cases with abnormal breathing), patient_consious (96.29% cases with unconscious patients), dispatcher_assertive_passive (underrepresentation of passive dispatchers), age_group (increasing sample with increasing age), caller_emotional_distressed (67.25% of callers are not stressed) and for the feature symptom (56.94% of cases with the symptom Asystoli), it was found that the OHCA recognition time of the ML model is faster for the feature expression for which the sample is largest. This suggests a representation bias, defined in Section 2.2.3.

Measurement Bias: According to the information available, the measurement bias described in Section 2.2.3 is not present in the use case discussed here. It is a fact that the process of collecting the data and, above all, the selection of the characteristics represent a bias: The database we have contains some characteristics for analysis, which are listed in Figure 9. For example, there is no information on the ethnic origin of the patients in the database. But measurement bias means using different characteristics for the groups to be analysed. In this use case, the same characteristics are analysed for all calls.

Aggregation Bias: For example, men and women suffer from different symptoms if they experience an OHCA. Based on the available information, it cannot be conclusively

assessed that the ML model is aware of these differences and incorporates them into the identification [141].

Temporal Bias: This bias exists in the use case at hand and the accompanying database. So far, there are two studies [20],[136] where the AI system has been applied. It will be some time before the actual and practical implementation and until then, further studies should be done, and the processes and science should be continuously reviewed.

F_{18}(Fairness; Diversity, non-discrimination and fairness):

Claim: Dispatchers and developers are trained to be aware of possible bias.

Evidence (further investigation): Information on whether the dispatchers have received such an introduction is not available. According to Corti [137], however, the algorithm is continuously tested for biases, which suggests that the developers and those responsible are aware of possible biases.

F_{19}(Fairness; Diversity, non-discrimination and fairness):

Claim: A fairness definition has been determined.

Evidence (further investigation): A common understanding of what fairness means and which fairness metric (see Section 2.2) should be used as a basis for the assessment is of fundamental importance. Depending on the fairness metric, the system can be assessed as fair or unfair. As there is no information on which metric the AI system should be fair, this flag cannot be evaluated conclusively.

F_{20}(Fairness; Diversity, non-discrimination and fairness):

Claim: Easy use of the AI system for every dispatcher regardless of their abilities (easy UI design).

Evidence (further investigation): Similar to the ethical issue I_{11}, it can be assumed that a medical dispatcher must fulfil certain requirements in order to exercise the profession at all. It can also be discussed here that, for example, an older medical dispatcher will have more problems interacting with the AI system. In addition, the biases in user interaction presented in chapter 3.2.1 are briefly discussed.

Position bias: The user interface of the ML model is programmed - assumingly - for the representation of the western world. A dispatcher who comes from a different population group and is familiar with a different representation of writing and images may need more time to get used to the representation.

Content production bias: The same applies to this bias. There is also no information on whether the user interface is available in different languages and in which form of expression, syntax and semantic.

F_{21}(Fairness; Diversity, non-discrimination and fairness):

Claim: The AI system allows by predicting OHCA deviating process alternatives (different disease progressions).

Evidence (further investigation): Whether the different courses and complaints that can occur in the course of an OHCA are integrated in the ML model cannot be conclusively assessed in the course of F_{17} under the aggregation bias.

F_{22}(Fairness; Diversity, non-discrimination and fairness):

Claim: Developers are aware of patients and dispatchers needs.

Evidence (further investigation): Although the developers are not medical professionals, the ML model was developed with the help of medical experts. They are aware of the need.

F_{23-29}(Fairness; Accountability):

The claims on fairness and accountability are summarised in this Section.

Evidence: No information is available for final assessment regarding GDPR, protocols, responsibilities and accountabilities, communication within and outside the organisation, risk assessment and documentation. However, it can be assumed that such processes and responsibilities exist. Further, there are the authors of the published papers.

After taking a closer look at the use case and the ethical issues and flags highlighted with regard to fairness, it can be concluded that there are still open points (flags that could not be conclusively assessed) that should be taken into account by the development team. Furthermore, although possible biases can be identified on the basis of the available data, no fairness metrics could be applied because the data only contains true positives. The identified biases should therefore first be verified with the entire data (see Section 2.2.5, step 1) and then mitigated with mitigation algorithms (see Section 2.2.5, step 2). In order to compare whether the biases have been attenuated and whether there may be a need to apply further mitigation algorithms, steps 3 and 4 described in Section 2.2.5 should be followed.

5.2 Explainability

The assessment of the AI system/ML model in terms of explainability is based on the key principle explicability and the requirement for transparency mentioned in Section 2.3.1. The assessment therefore also includes the further requirements for traceability and communication linked to explainability. Functions, decisions and limits of the system should thus be traceable as well as explainable and communicable internally and externally. The respective issues and flags derived in Section 4.1.3 serve as a basis for the following assessments.

5.2.1 Ethical issues related to the ethical principle of explicability

In the following, all ethical issues outlined in Section 4.1.3 are explained in more detail and possible solutions are given as examples.

I_{13-15}(Explicability; Transparency):

Claim: The decisions are explainable by the ML engineers, the dispatcher and the clinicians.

Evidence: The Issues I_{13-15} can be listed as unverified claims due to several factors. On the one hand, human characteristics and complaints are difficult to explain by causal reasons due to the high diversity. Second, the ML model is a black box model, which is why explainability cannot be demonstrated based on model transparency (see Section 2.3.2) with the components simulatability (entire model), decomposability (components and parameters), and algorithmic transparency (training algorithm). Consequently, AI system transparency can only be established via post-hoc explanations. One possibility to create these are so-called surrogate models

(Section 2.3.5). This possibility is explained in the following with reference to the use case.

In order to generate post-hoc feature-based explanations, a surrogate model was trained on the true positives during the analysis of the data. Only features of interest for our analysis were selected. Since the dataset contains only the OHCA cases correctly classified by the black box model, as explained previously, the original dependent feature of OHCA classification (1 or 0) could not be adopted - the OHCA class in the underlying data is always 1. Since we focus on the speed of OHCA prediction of the black box model in the analysis of the data, this feature was chosen as the dependent variable (see Table 21) for the surrogate model to generate post-hoc explanations.

Surrogate Model	Random Forest Regressor
X	patient_gender, caller_gender, caller_relation, caller_access_patient, area_call, patient_breathing_normally, caller_emotional_distressed, age_group, dispatcher_assertive_passive, patient_conscious, caller_alone
y	unrecognition_time_corti
Explainability technique	SHAP[9]

Table 21: Overview of surrogate model and explainability technique for data evaluation.

Since the independent features are categorical variables and thus categorical and numerical features need to be combined, a random forest regressor [143],[144] as listed in Table 21 was implemented as a surrogate model and extended with SHAP. This allowed features to be compared in terms of their influence on the speed of OHCA detection of the black box model and significant features to be further analysed. The explainability technique SHAP provided as output the most significant features. A high influence (mean|SHAP value|>5[10]) on the speed (see Table 22) has therefore, as already explained in more detail in 5.2, the fact whether the dispatcher acts assertively or passively, the environment from which the call is made, the age group of the patient, the fact whether the patient is conscious, the gender of the patient as well as his relationship to the caller.

Feature	**Feature importance (SHAP)**
dispatcher_assertive_passive	9.306525
area_call	7.875187
age_group	7.757750
patient_conscious	7.090782
patient_gender	6.792288
caller_relation	5.577899
caller_emotional_distressed	4.537966
caller_alone	4.493633
caller_gender	4.123097
patient_breathing_normally	2.403856
caller_access_patient	1.890195

Table 22: Explanation by top feature importances on ML model's recognition time (SHAP values).

[9] Package: shap, version: 0.39.0.

[10] Since the median feature importance (see Table 22) is 5.58, we decided to consider only features with a feature importance higher than 5.00.

To further explain to what extent and in what way (positive or negative) the feature affects the speed of OHCA detection, the SHAP values can be used as explained in Section 2.3.5. In Figure 10, all features and their positive or negative influence are mapped. In addition, it can be seen to what extent the characteristics of the features (feature value) lead to a negative or positive SHAP value. In the following list, which supports the explanations mentioned in Section 4.2.2, we also restrict ourselves to the features with the highest feature importances (mean|SHAP value|>5):

dispatcher_assertive_passive [0: 'Assertive', 1: 'N/A', 2: 'Passive']: A low value of the feature influences the recognition time of the ML model on average with a negative SHAP value (and vice versa). Thus, a passive dispatcher increases the recognition time of the ML model, whereas an assertive dispatcher decreases the recognition time of the ML model.

area_call [0: 'Other area', 1: 'Natural area', 2: 'Private area', 3: 'Traffic area']: A medium value of the feature influences the recognition time of the ML model on average with a negative SHAP value. A high and low value of the feature influences the recognition time of the ML model on average with a positive SHAP value. A call from a traffic environment results in a longer recognition time, whereas a call from a private environment shortens the recognition time of the ML model.

age_group [0: '<= 35', 1: '36-64', 2: '65-79', 3: '>= 80']: A low value of the feature influences the recognition time of the ML model on average with a positive SHAP value. The higher the value of the feature, the more decreasingly the recognition time of the ML model is influenced. Consequently, higher age groups lead to a reduction of the ML recognition time.

patient_conscious [0: 'N/A', 1: 'No', 2: 'Yes']: A high value of the feature influences the recognition time of the ML model on average with a positive SHAP value (and vice versa). Thus, if the caller addresses that the patient is conscious, this will increase the detection time of the ML model.

patient_gender [0: 'female', 1: 'male']: A low value of the feature influences the detection time of the ML model on average with a positive SHAP value. Thus, the ML model detects OHCA for male patients faster than for female patients on average.

caller_relation [0: 'Healthcare professional', 1: 'Other', 2: 'Relative']: A low value of the feature influences the recognition time of the ML model on average with a negative SHAP value. Higher values of the feature lead on average to a positive SHAP value on the recognition time of the ML model. So if the caller is a healthcare professional the recognition time of the ML model will be shortened.

The previously explained possibilities of interpreting a black box model via surrogate models with the help of feature importances and SHAP values could help to establish a transparency of decisions of the ML model and to ensure an explainability at least partially via feature-based explanations. To generate further explanations for the general OHCA classification of calls by the black box model, counterfactual explanations (see Section 2.3.5) could also be used to describe the black box decisions. If all data from the ML model (i.e., including calls that the black box model has classified as not OHCA) were accessible, this would allow local example-based explanations to be provided about what values one or more features of a call would have to take in order to be classified as OHCA by the model. For example, it could be used to identify what age, gender, or symptoms a patient would need to have in order to be classified as an OHCA patient. This could help to make decision-making processes of the ML model more transparent for ML developers,

dispatchers and clinicians (but also for patients in case of questions or contestations), to justify individual cases and, if necessary, to uncover weaknesses (incorrect functioning).

Figure 10: Explanation by impact of features on ML model's recognition time (SHAP values).

I_{16}(Explainability; Transparency):

Claim: The patient and the bystander/caller are informed about the usage of the AI system.

Evidence: Neither the patient nor the caller receives any information about the use of the AI system when calling. However, such information is essential to create transparency and consequently to gain the trust and acceptance of the population. Tension T_{10} should therefore be examined and analysed. Through awareness, the detection time of the OHCA possibly increases. Whether the GDPR Art. 12, 13 and 22 [18] come into play in the use case, though, and thus whether communication is necessary, should be questioned further. According to Art. 12 GDPR paragraph (1), information must be communicated in a transparent and easily understandable form. Art. 13 GDPR [18] further provides that information about the use of automated decision making, such as its logic and effects, must be communicated at the time of data collection. Art. 22 GDPR and the possibility to disagree with automated decision making also requires closer examination due to the special category of personal data (health data).

5.2.2 Flags related to the ethical principle of explicability

This Section explains all the flags associated with the ethical principle of explicability and describes which aspects require further analysis.

F_{30}(Explicability; Transparency):

Claim: Documentation of predictions and involved actors.

Evidence (further investigation): Both ML-based processes and human participants in decision-making should be documented to ensure transparency and consequently

traceability. It should be recorded in more detail, for example, which steps are initiated by the AI system when a call is made, which dispatcher processes the call, and whether the ML alert is accepted or ignored, i.e., who makes the prediction of the OHCA. Here, it should be examined to what extent such documentation is already continuously maintained. This has an impact both on the requirements of the GDPR [18] described above, which prescribes information obligations for the processing of data and for automated decision-making, and on the transparency obligation in the event of a contestation by the patient.

F_{31}(Explicability; Transparency):

Claim: Documentation of data (quality), recordings and their storage duration.

Evidence (further investigation): Just as the decision-making process and all persons (groups) involved in it should be documented, the calls themselves should be logged. The data generated and processed by the call, such as on patients, their feature characteristics, the time and duration of the call or the respective OHCA prognosis, should be stored and their quality, processing purposes and storage duration transparently documented. Again, it should be verified that such storage and documentation is continuous in addition to the training data we have. Transparent data storage and logging may be a prerequisite for justification in case of a contestation of the decision by the patient or caller. In addition, it forms the basis for the patient's right to information as set out in the GDPR [18]. According to Art. 15 GDPR, any data subject has the possibility to request information on the purposes of processing, categories of data, storage period and other characteristics of the processing. If the information is not possible, the acceptance and trust in the AI system but also in the respective emergency call center could be weakened.

F_{32}(Explicability; Transparency):

Claim: Possibilities and boundaries of the AI system are analysed, documented and communicated.

Evidence (further investigation): Furthermore, all functions, possibilities and limits of the AI system should be documented to enable a transparent description and delimitation. This also includes a clear division of tasks. The division of labor between dispatcher and ML model should therefore be recorded and can support transparency and traceability for dispatchers, patients, callers and all other stakeholders both before and after the AI system is deployed. Furthermore, such documentation can help to check the use of the AI system for its suitability depending on the use case. Whether and to what extent this is already being done, next to the documentation of the published studies, needs to be verified in more detail.

F_{33}(Explicability; Transparency):

Claim: Stakeholders are aware of possibilities and boundaries of the AI system.

Evidence (further investigation): This distinction of possibilities and limitations through the use of the AI system should therefore be provided and communicated to all stakeholders. Stakeholders should also be able to understand and explain the AI system. Information should be explained in an understandable and comprehensible way, depending on prior knowledge (see Section 2.3.2). This could be done in text form, for example, or already in feedback rounds to enable the same level of

understanding about the AI system and, if necessary, serve as a basis for discussions and suggestions for improvement. There are already the two published studies and the promotional video, which give a first overview of functionalities and results. Whether additional information is provided and whether the same level of transparency and explainability is already being established among all stakeholders must be examined in the future. In addition, the usefulness of the explanations already given, as mentioned in Section 2.3.5, could be checked with the help of the System Causability Scale in order to analyse potential for improvement.

Overall, according to our assessment, it cannot be fully determined whether measures to ensure transparency are already being implemented. This includes documentation as well as communication to all stakeholders. The analysis also revealed that, in particular, communication and information about the use of the AI system to patients is not taking place to the best of our knowledge. Furthermore, algorithmic transparency is not given due to the ML model being too complex. Since transparency is a prerequisite for explicability (and partly also for legal regulations), the points explained above should be examined more closely and, if necessary, the recommendations mentioned should be implemented. In addition, the tensions mentioned in Section 4.1.3 must be weighed up in terms of their respective importance and resolved in the best possible way.

6 Conclusion

Elon Musk estimates that artificial intelligence will be so advanced that it will exceed human intelligence by 2025 at the latest [1], and with this statement he is stirring up a lot of discussion about its veracity. This shows that AI is increasingly in the discourse of society and is being researched to a growing extent. More and more AI systems are being developed and used in a wide variety of applications. Thereby, the implementation and training in the development environment often works very well, but the performance in reality is not yet fully acceptable, or the acceptance of the users and society is lacking, like shown in the use case.

This is also presented in Sections 3.2, 3.3, and 3.4. A complete acceptance of the dispatchers alone cannot be confirmed so far; only a small part of the ML alerts is accepted by them. The AI system is not more accurate, but it is faster in recognising an OHCA. If the acceptance of the dispatchers were present and the alert was consequently accepted more frequently, faster help would be possible [20]. In order to strengthen the acceptance of users and society, it is therefore even more important to focus on establishing the trustworthiness of the AI system throughout the entire life cycle of the AI system (from development to deployment). In order to guarantee such a trustworthy AI system, the AI HLEG advises in the European area (association of states in the present use case) among other things to ensure fairness and explicability (see Section 2.1) [12].

AI systems can now be used for very responsible and critical applications (hiring decisions, autonomous driving and credit approvals etc). Therefore, it is more important than in the past that AI systems make fair decisions and predictions. Section 2.2.3 describes that the development of such a fair AI system is not that easy due to several issues: Besides the fact that, according to social psychology, fairness is a subjective perception that first has to be made measurable, there are three main sources that can introduce biases and thus discrimination against certain individuals or groups into the AI system. Firstly, the biases in the training data with which the AI system is trained, secondly, biases in the algorithm and thirdly, biases in the user interaction with the AI system. With regard to the use case at hand, no conclusive statements can be made about such biases, as we only have a small sample of true positive data available. However, there are tendencies towards biases in the data (for example, a representation bias with regard to some characteristics. The sample of male patients, for instance, is larger and the recognition of the ML model is faster in this group), in the algorithm (e.g. a sampling bias, because the use case was developed and tested in Denmark, so the algorithm was trained for the Danish language, this represents a discrimination for e.g. emergency callers who do not speak Danish) and in the user interaction (the presentation of the alarm is, as far as we know, adapted to the western world, Dispatchers from other parts of the world may take longer to realise the alarm. This represents a position and content production bias).

As explained in Section 2.3.1, explicability comprises the key requirement of transparency with the components traceability, explainability and communication. In order to enable explainability of decisions in the decision-making process, data and processes (including algorithms) should be documented. In addition, it should also be possible to explain the reasons for the general use of the AI system and functions as well as limitations of the AI system should be made available to all stakeholders. They should also be given the opportunity to disagree with the interaction with the AI system. The explanation should be based on the particular stakeholders and their prior

knowledge. This includes both the content and the scope. Explanations can refer to the whole model (global) or individual calls (local). When generating explanations, inclusion of multidisciplinary perspectives is essential to comply with legal, ethical, technical, and medical requirements. Explainability and interpretability in particular are necessary to enable trust, causality, transferability, informativeness and fair and ethical decision making. However, there is also the risk that medical staff may want to focus on people rather than on technical aspects or may lack technical background knowledge. Furthermore, causal relationships are not always understandable from the ground up. However, explanations can be useful here for further analyses, but also bear the risk of enabling malicious attacks via malicious, manipulated input. Related to the use case at hand, this means that documentation should be available on the division of labour between ML model and dispatcher and the associated processes and influencing factors in OHCA classification decisions. This needs to be further verified and cannot be fully verified by us. To our knowledge, communication to patients, callers, and other stakeholders has occurred only through published studies. However, direct information may be necessary to comply with legal requirements, as explained in Section 5.2.2. Since the ML model is a black box model, it is not possible to explain it from the ground up. However, this can be made possible via extension using interpretable models. This showed that some caller characteristics, such as especially calls from patients of higher age groups or calls from healthcare professionals shortened the OHCA detection time of the ML model (see Section 4.2.2).

Nevertheless, since the data available to us only contains true positives, i.e. OHCA cases correctly classified by the ML model, it is not possible to fully evaluate the AI system used in the use case with regard to the ethical principles of fairness and explicability. As the ML model is a black box model, a surrogate model was used to explain the differences in recognition time per feature of the ML model and, based on this, several features were considered and analysed with regard to fairness and explainability aspects. If all data (i.e. also the OHCA cases incorrectly classified by the ML model) were available, further analyses could be added specifically on the accuracy of the predictions of the ML model. It could be discussed which characteristics tend to lead to which prediction (OHCA or no OHCA) or how fair these classifications are. Further fairness and explainability techniques and metrics (see Sections 2.2.4, 2.2.5, and 2.3.5) could be used to build on this (e.g. using tools such as Microsoft Fair Learn, Google What If, AI Explainability 360 or Interpret ML).

If we had had access to all the data, in addition to the data analysis where we found bias tendencies, we could have applied fairness metrics (such as equal opportunity, which could find out whether the probability of the correct OHCA recognition was the same regardless of whether the patient was a man or a woman) to really identify biases and then reduce them using mitigation algorithms (see Section 2.2.5).

In addition to the used explainability technique SHAP (see Section 5.2.1), counterfactual explanations or the Contrastive Explanation Method (CEM) could be added. Counterfactual Explanations would, as mentioned in Section 2.3.5, show which value a feature has to take in order to be correctly classified as OHCA by the ML Model. The CEM would also show which features must be present and which should not be present in order to remain in the positive predicted class of the ML model (the ML model classifies the respective call as OHCA). The feature importances determined by SHAP could also be supplemented with the Faithfulness metric (see Section 2.3.5) to

check their correctness. The overall explanations generated by the techniques and supplemented with human descriptions could finally be checked with the System Causability Scale (see Section 2.3.5) for their usefulness for the caller or for further stakeholders.

With regard to the fairness of the AI system, the requirements for the AI system from society's point of view should be questioned in the future. As explained in Section 2.2.1, society per se is not completely fair; every human being is fundamentally biased and makes subjective decisions. It should therefore be clarified whether the requirement for the AI system is to act as fairly as humans, or to make decisions more fairly than humans. If the AI system is supposed to act more fairly than humans, it must also be analysed who evaluates fairness, i.e. who assesses the level of fairness and consequently defines a completeness of fairness.

In general, it could also be observed in the future whether and to what extent the training focus of the dispatchers changes. If the acceptance of the dispatchers is strengthened and consequently more alerts are accepted, the training could address the understanding of the ML model and its explicability in addition to the OHCA detection itself. In this way, patients or callers could receive initial explanations and information about the functioning and limitations of the ML model directly from the dispatcher.

Furthermore, additional functions and support for the dispatcher could be implemented. Besides the possibilities already mentioned in Section 4.1.3, such as a help button with additional explanations, the ML model could also suggest specific questions to the dispatcher, which could accelerate the recognition of the OHCA by the ML model. The questions could be generated based on the model training from past call records (for example, including the factors influencing the speed of the ML model as explained in Sections 4.2.2 and 5.2.1). However, not only the training focus might change in the future. The intended purpose of the AI system could also change to other medical emergencies. For example, the ML model could be trained to deal with strokes or blood clots in order to provide timely assistance [137].

The general use of AI systems in the health sector will also increase in the future. So far, many medical tasks are already supported with the help of speech (as in this use case) and text recognition. In the future, image analysis will also be increasingly used, e.g. to support radiological or pathological examinations. In this context, not only ensuring accuracy and usefulness will be of great importance, but also increasingly the acceptance of society. Particularly in the health sector, this represents a future challenge due to the sensitivity involved [145]. Davenport et al. [145] estimate that complete acceptance of the AI system can only be achieved after the AI system has fully matured and therefore does not expect comprehensive use of AI systems in the health sector until around 2029. According to Davenport et al., the focus will not be on substitution, but rather on the support of human work by AI systems. However, the human tasks themselves will change. They will probably include tasks that can only be performed by humans. These include, for example, tasks that require empathy or other purely human skills. A final question that could be further questioned and analysed in the future could be whether it will not even become a requirement in the health sector (and other areas) to use AI systems in order to be able to guarantee the latest technical standards and thus to be able and allowed to operate in the medical market at all.

References

[1] IANS, „AI will be smarter than humans within 5 years, says Elon Musk“, Nationalheraldindia.com, 28-Juli-2020. [Online]. Available: https://www.nationalheraldindia.com/science-and-tech/ai-will-be-smarter-than-humans-within-5-years-says-elon-musk. [Accessed: 13-May-2021].

[2] New York Times, “Times topics: Artificial Intelligence” [Online]. Available: https://www.nytimes.com/topic/subject/artificial-intelligence. [Accessed: 13-May-2021].

[3] „Revenue growth of leading tech companies 2019“, Statista.com. [Online]. Available: https://www.statista.com/statistics/277917/revenue-growth-of-selected-tech-companies/. [Accessed: 13-May-2021].

[4] J. Howard, „Artificial intelligence: Implications for the future of work“, Am. J. Ind. Med., Bd. 62, Nr. 11, S. 917–926, 2019.

[5] R. Lee, Artificial intelligence in daily life, 1st ed. Singapur, Singapur: Springer, 2020.

[6] V. Kepuska and G. Bohouta, “Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home),” in 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018.

[7] J.-A. Choi und K. Lim, „Identifying machine learning techniques for classification of target advertising“, ICT Express, Bd. 6, Nr. 3, S. 175–180, 2020.

[8] L. Agner, B. Necyk, und A. Renzi, „Recommendation systems and machine learning: Mapping the user experience“, in Design, User Experience, and Usability. Design for Contemporary Interactive Environments, Cham: Springer International Publishing, 2020, S. 3–17.

[9] P. K. Donepudi, “AI and Machine Learning in Banking: A Systematic Literature Review”, Asian j. appl. sci. eng., vol. 6, no. 3, pp. 157-162, Dec. 2017.

[10] M. Vasconcelos, C. Cardonha, and B. Gonçalves, “Modeling epistemological principles for bias mitigation in AI systems: An illustration in hiring decisions,” in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018.

[11] A. Väänänen, K. Haataja, K. Vehviläinen-Julkunen, and P. Toivanen, “AI in healthcare: A narrative review,” F1000Res., vol. 10, p. 6, 2021.

[12] “High-Level Expert Group on Artificial Intelligence (European Commission), “Ethics Guidelines for Trustworthy AI,” Apr, 2019.

[13] ‘Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)’, Europa.eu. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence. [Accessed: 13-May-2021].

[14] M. Brundage et al., ‘The malicious use of artificial intelligence: Forecasting, prevention, and mitigation’, arXiv [cs.AI], 2018.

[15] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, ‘The practical implementation of artificial intelligence technologies in medicine’, Nat. Med., vol. 25, no. 1, pp. 30–36, 2019.

[16] T. Davenport and R. Kalakota, “The potential for artificial intelligence in healthcare,” Future Healthc. J., vol. 6, no. 2, pp. 94–98, 2019.

[17] R. V. Zicari u. a., „Z-inspection®: A process to assess trustworthy AI“, IEEE Trans. Technol. Soc., S. 1–1, 2021.

[18] “EUR-Lex - 02016R0679-20160504 - EN - EUR-Lex,” Europa.eu. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04. [Accessed: 02-May-2021].

[19] J. Andress, The basics of information security: Understanding the fundamentals of InfoSec in theory and practice, 2nd ed. Syngress Publishing, 2014.
[20] S. N. Blomberg et al., “Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: A randomized clinical trial: A randomized clinical trial,” JAMA Netw. Open, vol. 4, no. 1, p. e2032320, 2021.
[21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS ’12, 2012.
[22] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, “How do fairness definitions fare?: Examining public attitudes towards algorithmic definitions of fairness,” in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019.
[23] A. Chouldechova, „Fair prediction with disparate impact: A study of bias in recidivism prediction instruments“, Big Data, Bd. 5, Nr. 2, S. 153–163, 2017.
[24] J. Rawls, “Justice as Fairness,” in Contemporary Political Theory: A Reader, 1 Oliver’s Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2014, pp. 13–21.
[25] “Title III: Equality,” Europa.eu. [Online]. Available: https://fra.europa.eu/en/eu-charter/title/title-iii-equality. [Accessed: 30-Apr-2021].
[26] S. Verma and J. Rubin, “Fairness definitions explained,” in Proceedings of the International Workshop on Software Fairness, 2018.
[27] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” arXiv [cs.LG], 2016.
[28] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, und K. Q. Weinberger, „On Fairness and Calibration“, arXiv [cs.LG], 2017.
[29] D. Pessach und E. Shmueli, „Algorithmic Fairness“, arXiv [cs.CY], 2020.
[30] A. Chouldechova and A. Roth, “A snapshot of the frontiers of fairness in machine learning,” Commun. ACM, vol. 63, no. 5, pp. 82–89, 2020.
[31] P. Herbig and J. Milewicz, “The relationship of reputation and credibility to brand success,” J. Consum. Mark., vol. 12, p. 5+, 1995.
[32] M. A. Ahmad, A. Patel, C. Eckert, V. Kumar, and A. Teredesai, “Fairness in Machine Learning for Healthcare,” in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020.
[33] R. Dresser, “Wanted. Single, white male for medical research,” Hastings Cent. Rep., vol. 22, no. 1, pp. 24–29, 1992.
[34] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” arXiv [cs.LG], 2019.
[35] R. Binns, “Fairness in machine learning: Lessons from political philosophy,” arXiv [cs.CY], 2017.
[36] A. Howard and J. Borenstein, “The ugly truth about ourselves and our robot creations: The problem of bias and social inequity,” Sci. Eng. Ethics, vol. 24, no. 5, pp. 1521–1536, 2018.
[37] R. Baeza-Yates, “Bias on the web,” Commun. ACM, vol. 61, no. 6, pp. 54–61, 2018.
[38] H. Suresh and J. V. Guttag, “A framework for understanding unintended consequences of machine learning,” arXiv [cs.LG], 2019.
[39] “Evaluations of AI Applications in Healthcare,” Coursera.org. [Online]. Available: https://www.coursera.org/learn/evaluations-ai-applications-healthcare. [Accessed: 02-May-2021].

[40] Obermeyer, Powers, and E. al Vogeli, “Dissecting racial bias in an algorithm used to manage the health of populations,” Yearbook of Paediatric Endocrinology, 2020.
[41] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, “Social data: Biases, methodological pitfalls, and ethical boundaries,” Front. Big Data, vol. 2, p. 13, 2019.
[42] W. Yuan et al., “Temporal bias in case-control design: preventing reliable predictions of the future,” Nat. Commun., vol. 12, no. 1, p. 1107, 2021.
[43] B. Friedman and H. Nissenbaum, “Bias in computer systems,” ACM Trans. Inf. Syst., vol. 14, no. 3, pp. 330–347, 1996.
[44] J. Herlitz u. a., „Is there a difference between women and men in characteristics and outcome after in hospital cardiac arrest?“, Resuscitation, Bd. 49, Nr. 1, S. 15–23, 2001.
[45] S. Claude und I. Webb Geoffrey, Encyclopedia of machine learning. Springer Science+Business Media, 2010.
[46] Foster Provost and Ron Kohavi, On Applied Research in Machine Learning. In Machine learning, 1998.
[47] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, “Understanding and using sensitivity, specificity and predictive values,” Indian J. Ophthalmol., vol. 56, no. 1, pp. 45–50, 2008.
[48] A. G. Lalkhen and A. McCluskey, “Clinical tests: sensitivity and specificity,” Contin. Educ. Anaesth. Crit. Care Pain, vol. 8, no. 6, pp. 221–223, 2008.
[49] „WIT - getting started“, Github.io. [Online]. Availabel: https://pair-code.github.io/what-if-tool/get-started/. [Accessed: 08-May-2021].
[50] „Fairlearn“, Fairlearn.org. [Online]. Available: https://fairlearn.org/. [Accessed: 08-May-2021].
[51] “AI Fairness 360,” Mybluemix.net. [Online]. Available: http://aif360.mybluemix.net/. [Accessed: 02-May-2021].
[52] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, “Counterfactual Fairness,” arXiv [stat.ML], 2017.
[53] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, “Algorithmic decision making and the cost of fairness,” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’17, 2017.
[54] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” arXiv [cs.LG], 2016.
[55] S. Barocas und A. D. Selbst, „Big data’s disparate impact“, SSRN Electron. J., 2016.
[56] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, “Understanding the origins of bias in word embeddings,” arXiv [cs.LG], 2018.
[57] F. P. Calmon, D. Wei, K. N. Ramamurthy, and K. R. Varshney, “Optimized data pre-processing for discrimination prevention,” arXiv [stat.ML], 2017.
[58] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, “Adaptive sensitive reweighting to mitigate bias in fairness-aware classification,” in Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18, 2018.
[59] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” Knowl. Inf. Syst., vol. 33, no. 1, pp. 1–33, 2012.
[60] B. Fish, J. Kun, and Á. D. Lelkes, “A confidence-based approach for balancing fairness and accuracy,” in Proceedings of the 2016 SIAM International Conference on Data Mining, 2016.

[61] B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018.

[62] B. d’Alessandro, C. O’Neil, and T. LaGatta, “Conscientious classification: A data scientist’s guide to discrimination-aware classification,” Big Data, vol. 5, no. 2, pp. 120–134, 2017.

[63] A. Falcon, “Aristotle on Causality,” Metaphysics Research Lab, Stanford University, 2019.

[64] P. Aalto et al., “Article 47 – right to an effective remedy and to a fair trial,” in The EU Charter of Fundamental Rights, Nomos, 2014, pp. 1240–1319.

[65] F. Doshi-Velez and B. Kim, “Towards A rigorous science of interpretable machine learning,” arXiv [stat.ML], 2017.

[66] D. Dave, H. Naik, S. Singhal, and P. Patel, “Explainable AI meets healthcare: A study on heart disease dataset,” arXiv [cs.LG], 2020.

[67] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and Precise4Q consortium, “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective,” BMC Med. Inform. Decis. Mak., vol. 20, no. 1, p. 310, 2020.

[68] P. Ahttps://aix360.readthedocs.io/en/latest/. Noseworthy et al., “Shared decision-making in atrial fibrillation: navigating complex issues in partnership with the patient,” J. Interv. Card. Electrophysiol., vol. 56, no. 2, pp. 159–163, 2019.

[69] Z. C. Lipton, “The mythos of model interpretability,” Commun. ACM, vol. 61, no. 10, pp. 36–43, 2018.

[70] “InterpretML,” Interpret.ml. [Online]. Available: https://interpret.ml/. [Accessed: 09-May-2021].

[71] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018.

[72] R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, and D. Johnson, “Case-based explanation of non-case-based learning methods,” Proc. AMIA Symp., pp. 212–215, 1999.

[73] United States Government, “THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN: 2019 UPDATE” United States Government [Online] Available: The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update (nitrd.gov). [Accessed: 02-May-2021].

[74] U. Bhatt et al., “Explainable machine learning in deployment,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020.

[75] M. Sendak et al., “The Human Body is a Black Box: Supporting Clinical Decision-Making with Deep Learning,” in Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAT* ’20), 2020.

[76] M. Hellkvist, A. Ozcelikkale, and A. Ahlen, “Generalization error for linear regression under distributed learning,” in 2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2020.

[77] “Interpretml/interpret,” Github.com [Online]. Available: https://github.com/interpretml/interpret [Accessed: 02-May-2021].

[78] H. Xu et al., “Adversarial attacks and defenses in images, graphs and text: A review,” Int. J. Autom. Comput., vol. 17, no. 2, pp. 151–178, 2020.

[79] T. Ploug and S. Holm, “The four dimensions of contestable AI diagnostics - A patient-centric approach to explainable AI,” Artif. Intell. Med., vol. 107, no. 101901, p. 101901, 2020.

[80] R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, and D. Johnson, “Case-based explanation of non-case-based learning methods,” Proc. AMIA Symp., pp. 212–215, 1999.

[81] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” arXiv [cs.AI], 2017.

[82] “An introduction to explainable AI with Shapley values — SHAP latest documentation,” Readthedocs.io. [Online]. Available: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html. [Accessed: 09-May-2021].

[83] “Welcome to the SHAP documentation — SHAP latest documentation,” Readthedocs.io. [Online]. Available: https://shap.readthedocs.io/en/latest/. [Accessed: 02-May-2021].

[84] D. Alvarez-Melis and T. S. Jaakkola, “Towards robust interpretability with self-explaining neural networks,” arXiv [cs.LG], 2018.

[85] A. Holzinger, A. Carrington, and H. Müller, “Measuring the quality of explanations: The System Causability Scale (SCS): Comparing human and machine explanations: Comparing human and machine explanations,” KI - Künstl. Intell., vol. 34, no. 2, pp. 193–198, 2020.

[86] V. Arya et al., “AI explainability 360: Hands-on tutorial,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020.

[87] “AI Explainability 360 — aix360 0.1 documentation,” Readthedocs.io. [Online]. Available: https://aix360.readthedocs.io/en/latest/. [Accessed: 02-May-2021].

[88] “Trusted-AI/AIX360,” Github.com [Online]. Available: https://github.com/Trusted-AI/AIX360 [Accessed: 02-May-2021].

[89] V. Arya et al., “One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques,” arXiv [cs.AI], 2019.

[90] “Trusted-AI/AIX360,” Github.com [Online]. Available: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/README.md [Accessed: 02-May-2021].

[91] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, “Efficient data representation by selecting prototypes with importance weights,” in 2019 IEEE International Conference on Data Mining (ICDM), 2019.

[92] A. Kumar, P. Sattigeri, and A. Balakrishnan, “Variational inference of disentangled latent concepts from unlabeled observations,” arXiv [cs.LG], 2017.

[93] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” arXiv [stat.ML], 2013.

[94] M. Hind et al., “TED: Teaching AI to explain its decisions,” in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019.

[95] A. Dhurandhar et al., “Explanations based on the missing: Towards contrastive explanations with pertinent negatives∗,” Arxiv.org. [Online]. Available: http://arxiv.org/abs/1802.07623v2. [Accessed: 09-May-2021].

[96] R. Luss et al., “Generating contrastive explanations with monotonic attribute functions,” arXiv [cs.LG], 2019.

[97] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, 2016.

[98] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” arXiv [cs.AI], 2017.

[99] S. M. Lundberg, G. G. Erion, and S.-I. Lee, “Consistent individualized feature attribution for tree ensembles,” arXiv [cs.LG], 2018.

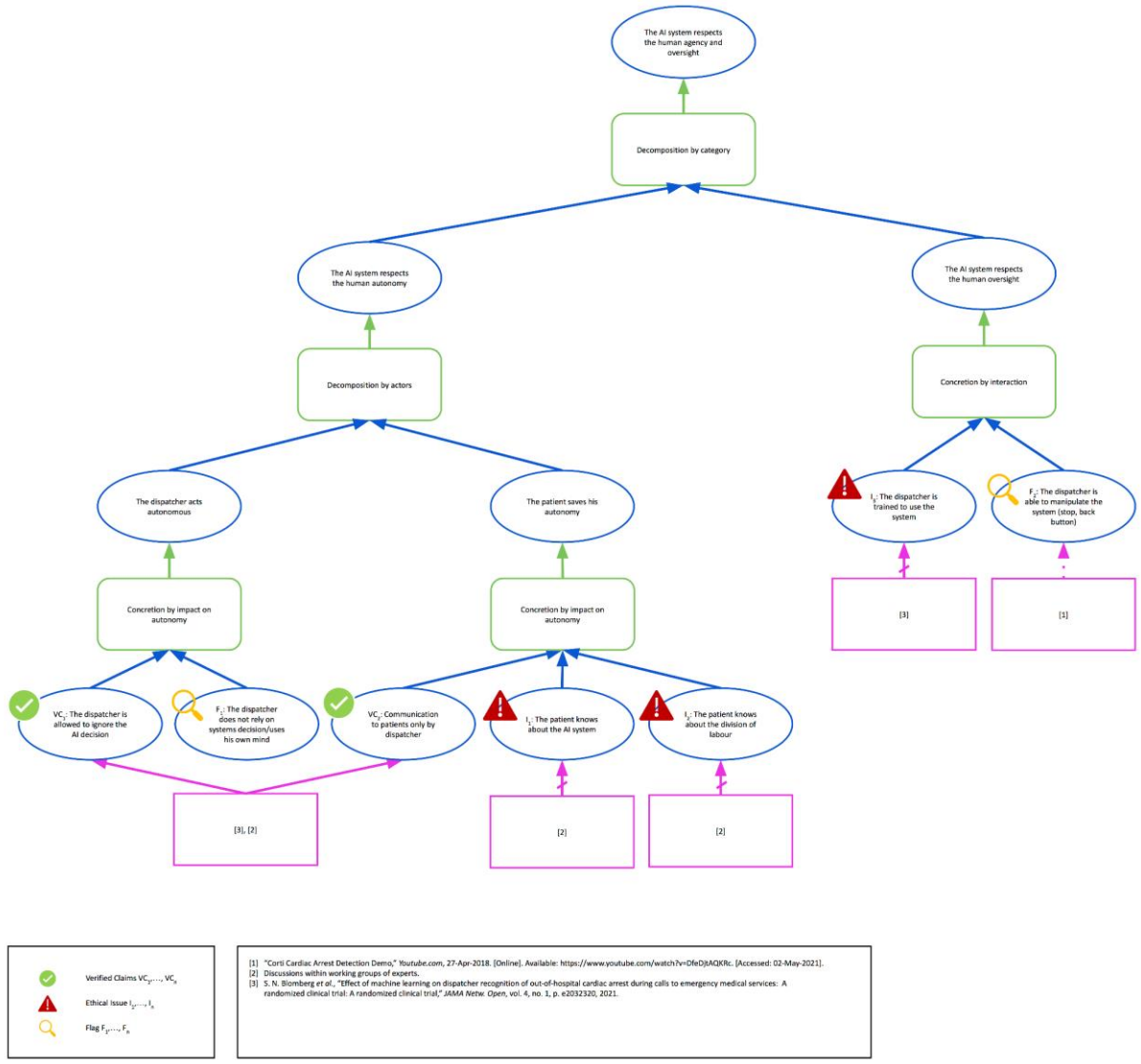
[100] S. M. Lundberg et al., “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery,” Nat. Biomed. Eng., vol. 2, no. 10, pp. 749–760, 2018.
[101] S. Dash, O. Günlük, and D. Wei, “Boolean decision rules via column generation,” arXiv [cs.AI], 2018.
[102] D. Wei, S. Dash, T. Gao, and O. Günlük, “Generalized Linear Rule Models,” arXiv [cs.LG], 2019.
[103] A. Dhurandhar, K. Shanmugam, R. Luss, and P. Olsen, “Improving simple models with confidence profiles,” arXiv [cs.LG], 2018.
[104] “Algorithms — aix360 0.1 documentation,” Readthedocs.io. [Online]. Available: https://aix360.readthedocs.io/en/latest/algorithms.html. [Accessed: 09-May-2021].
[105] A. Rai, “Explainable AI: from black box to glass box,” J. Acad. Mark. Sci., vol. 48, no. 1, pp. 137–141, 2020.
[106] R. Luss et al., “Generating contrastive explanations with monotonic attribute functions,” arXiv [cs.LG], 2019.
[107] H. Nori, S. Jenkins, P. Koch, and R. Caruana, “InterpretML: A unified framework for machine learning interpretability,” arXiv [cs.LG], 2019.
[108] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission,” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’15, 2015.
[109] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, “Accurate intelligible models with pairwise interactions,” in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013.
[110] Y. Lou, R. Caruana, and J. Gehrke, “Intelligible models for classification and regression,” in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12, 2012
[111] X. Zhang, S. Tan, P. Koch, Y. Lou, U. Chajewska, and R. Caruana, “Axiomatic Interpretability for Multiclass Additive Models,” in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019.
[112] S. Tan, R. Caruana, G. Hooker, and Y. Lou, “Distill-and-compare: Auditing black-box models using transparent model distillation,” in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018.
[113] B. Lengerich, S. Tan, C.-H. Chang, G. Hooker, and R. Caruana, “Purifying interaction effects with the Functional ANOVA: An efficient algorithm for recovering identifiable additive models,” arXiv [stat.ML], 2019.
[114] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, “Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning,” in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.
[115] C.-H. Chang, S. Tan, B. Lengerich, A. Goldenberg, and R. Caruana, “How Interpretable and Trustworthy are GAMs?,” arXiv [cs.LG], 2020.
[116] I. H. Witten, E. Frank, and M. A. Hall, Data mining: Practical machine learning tools and techniques, 3rd ed. Morgan Kaufmann, 2011.
[117] D. Esposito and F. Esposito, Introducing Machine Learning. Boston, MA: Addison Wesley, 2020.

[118] D. Cook, Practical machine learning with H2O: Powerful, scalable techniques for deep learning and AI. O'Reilly Media, 2017.
[119] J. Bell, Machine learning: Hands-on for developers and technical professionals, 2nd ed. Nashville, TN: John Wiley & Sons, 2020.
[120] J. Herman and W. Usher, "SALib: An open-source Python library for Sensitivity Analysis," J. Open Source Softw., vol. 2, no. 9, p. 97, 2017.
[121] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," Technometrics, vol. 33, no. 2, pp. 161–174, 1991.
[122] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.
[123] T. G. Maak und J. D. Wylie, „Medical device regulation: A comparison of the United States and the European union", J. Am. Acad. Orthop. Surg., Bd. 24, Nr. 8, S. 537–543, 2016.
[124] J. N. Whittlestone, R. Alexandrova, A. Dihal, K. Cave, und S, Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Nuffield Foundation, 2019.
[125] "CAE FRAMEWORK," Claimsargumentsevidence.org. [Online]. Available: https://claimsargumentsevidence.org/. [Accessed: 02-May-2021].
[126] "CAE Concepts," Claimsargumentsevidence.org. [Online]. Available: https://claimsargumentsevidence.org/notations/claims-arguments-evidence-cae/. [Accessed: 02-May-2021].
[127] "CAE Concise Guidance," Claimsargumentsevidence.org. [Online]. Available: https://claimsargumentsevidence.org/notations/concise-guidance/. [Accessed: 02-May-2021].
[128] "CAE Building Blocks," Claimsargumentsevidence.org. [Online]. Available: https://claimsargumentsevidence.org/notations/cae-building-blocks/. [Accessed: 02-May-2021].
[129] B. Futurium, Ed., Futurium: House of Futures. ALTAI- The Assessment List on Trustworthy Artificial Intelligence [Online]. Available: https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence. [Accessed:02-May-2021].
[130] Europa.eu. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342. [Accessed: 02-May-2021].
[131] "Home page - ALTAI," Insight. [Online]. Available: https://altai.insight-centre.org/. [Accessed: 29-Apr-2021].
[132] "How to complete ALTAI - ALTAI," Insight. [Online]. Available: https://altai.insight-centre.org/Home/HowToComplete. [Accessed: 29-Apr-2021].
[133] F. Taher, "Early detection of lung cancer based on artificial intelligence techniques," in 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), 2017.
[134] M. P. Christiansen et al., "Accuracy of a fourth-generation subcutaneous continuous glucose sensor," Diabetes Technol. Ther., vol. 19, no. 8, pp. 446–456, 2017.
[135] S. Dalton-Brown, "The ethics of medical AI and the physician-patient relationship," Camb. Q. Healthc. Ethics, vol. 29, no. 1, pp. 115–121, 2020.
[136] S. N. Blomberg et al., "Machine learning as a supportive tool to recognise cardiac arrest in emergency calls," Resuscitation, vol. 138, pp. 322–329, 2019.
[137] Discussions within working groups of experts.

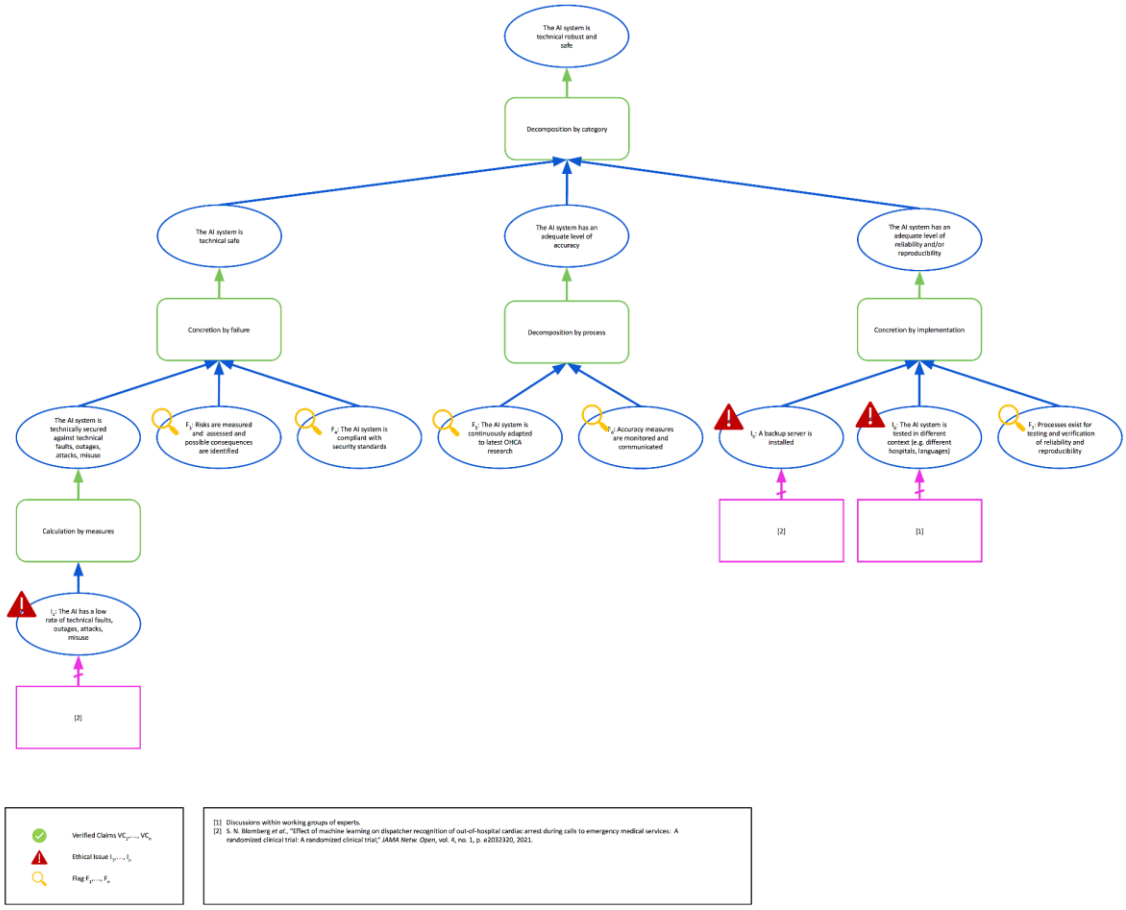
[138] R. N. Fogoros, “The important difference between a heart attack and a cardiac arrest,” Verywellhealth.com. [Online]. Available: https://www.verywellhealth.com/heart-attack-vs-cardiac-arrest-1746273. [Accessed: 02-May-2021].
[139] Corti.ai. [Online]. Available: https://www.corti.ai. [Accessed: 02-May-2021].
[140] “Corti Cardiac Arrest Detection Demo,” Youtube.com, 27-Apr-2018. [Online]. Available: https://www.youtube.com/watch?v=DfeDjtAQKRc. [Accessed: 02-May-2021].
[141] J. Herlitz et al., “Characteristics of cardiac arrest and resuscitation by age group: an analysis from the Swedish Cardiac Arrest Registry,” Am. J. Emerg. Med., vol. 25, no. 9, pp. 1025–1031, 2007.
[142] “Green hospitals,” Healthcaredenmark.dk. [Online]. Available: https://www.healthcaredenmark.dk/the-case-of-denmark/sustainable-hospitals/green-hospitals/. [Accessed: 02-May-2021].
[143] D. Esposito and F. Esposito, Introducing Machine Learning. Boston, MA: Addison Wesley, 2020.
[144] D. Cook, Practical machine learning with H2O: Powerful, scalable techniques for deep learning and AI. O’Reilly Media, 2017.
[145] T. Davenport and R. Kalakota, “The potential for artificial intelligence in healthcare,” Future Healthc. J., vol. 6, no. 2, pp. 94–98, 2019.
[146] Supplementary to RCT Data.

Appendix

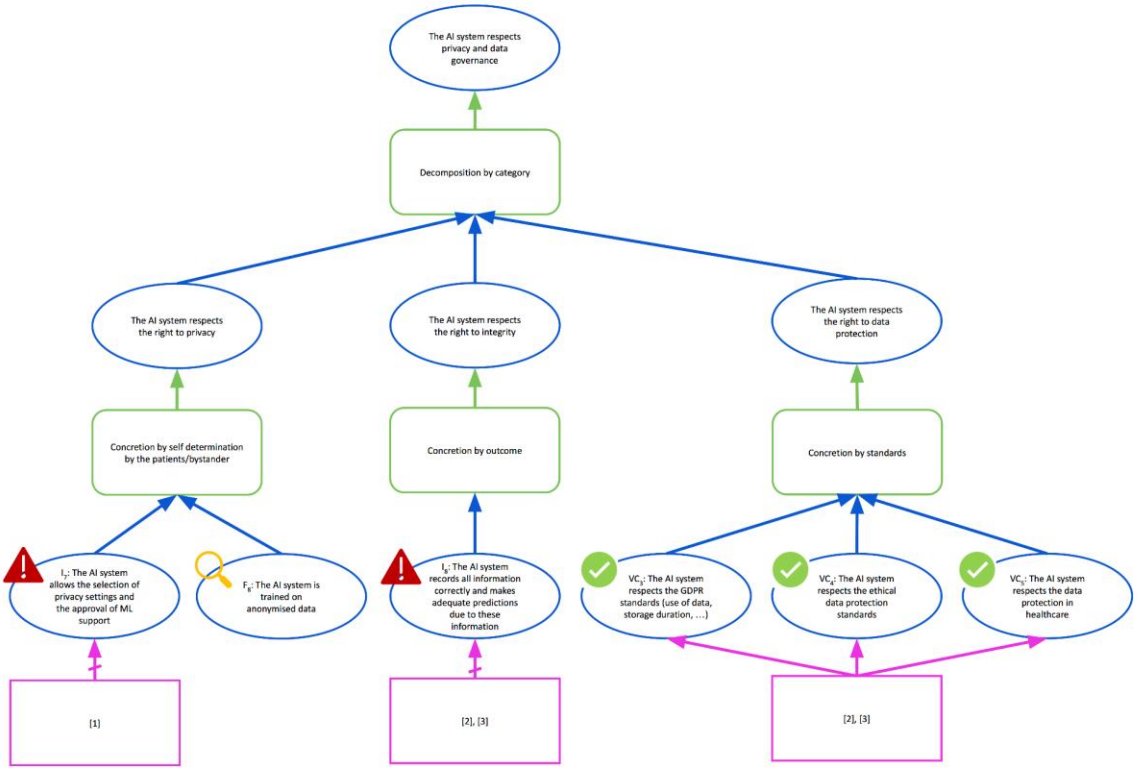
Appendix A: CAE - The AI system respects human agency and oversight



Appendix B: CAE - The AI system is technical robust and safe

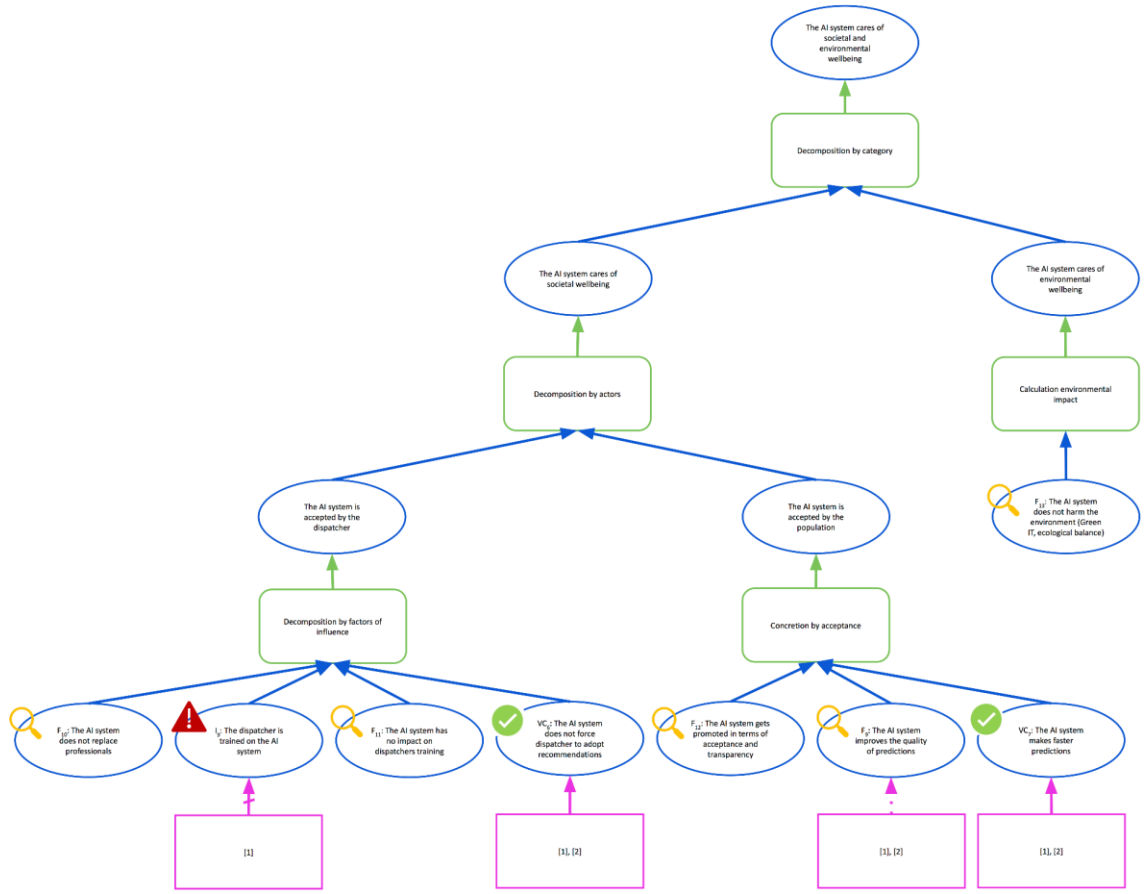


Appendix C: CAE - The AI system respects privacy and data governance

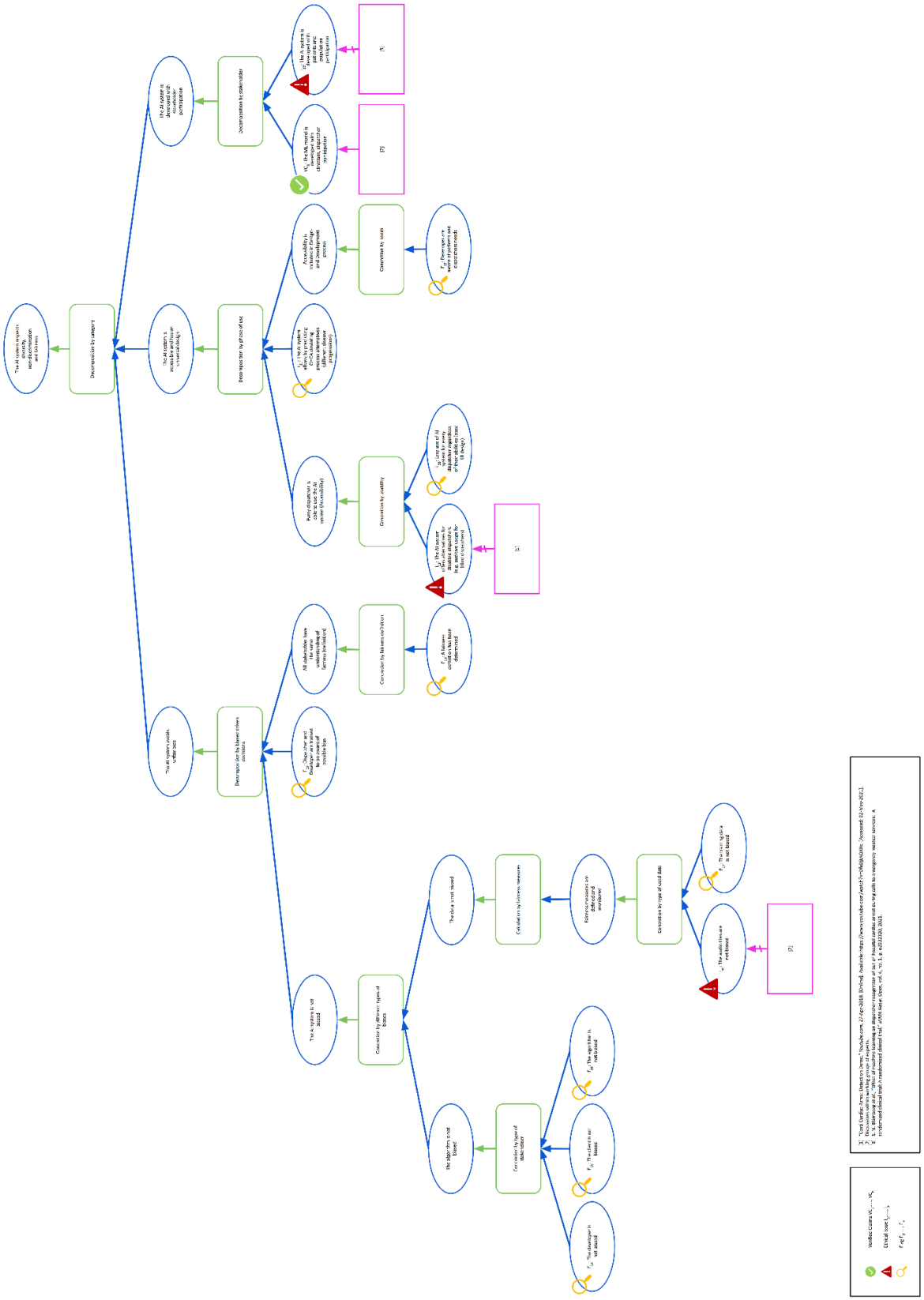


[1] "Corti Cardiac Arrest Detection Demo," *Youtube.com*, 27-Apr-2018. [Online]. Available: https://www.youtube.com/watch?v=DfeDjtAQKRc. [Accessed: 02-May-2021].
[2] S. N. Blomberg *et al.*, "Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: A randomized clinical trial: A randomized clinical trial," *JAMA Netw. Open*, vol. 4, no. 1, p. e2032320, 2021.
[3] S. N. Blomberg *et al.*, "Machine learning as a supportive tool to recognize cardiac arrest in emergency calls," *Resuscitation*, vol. 138, pp. 322–329, 2019.

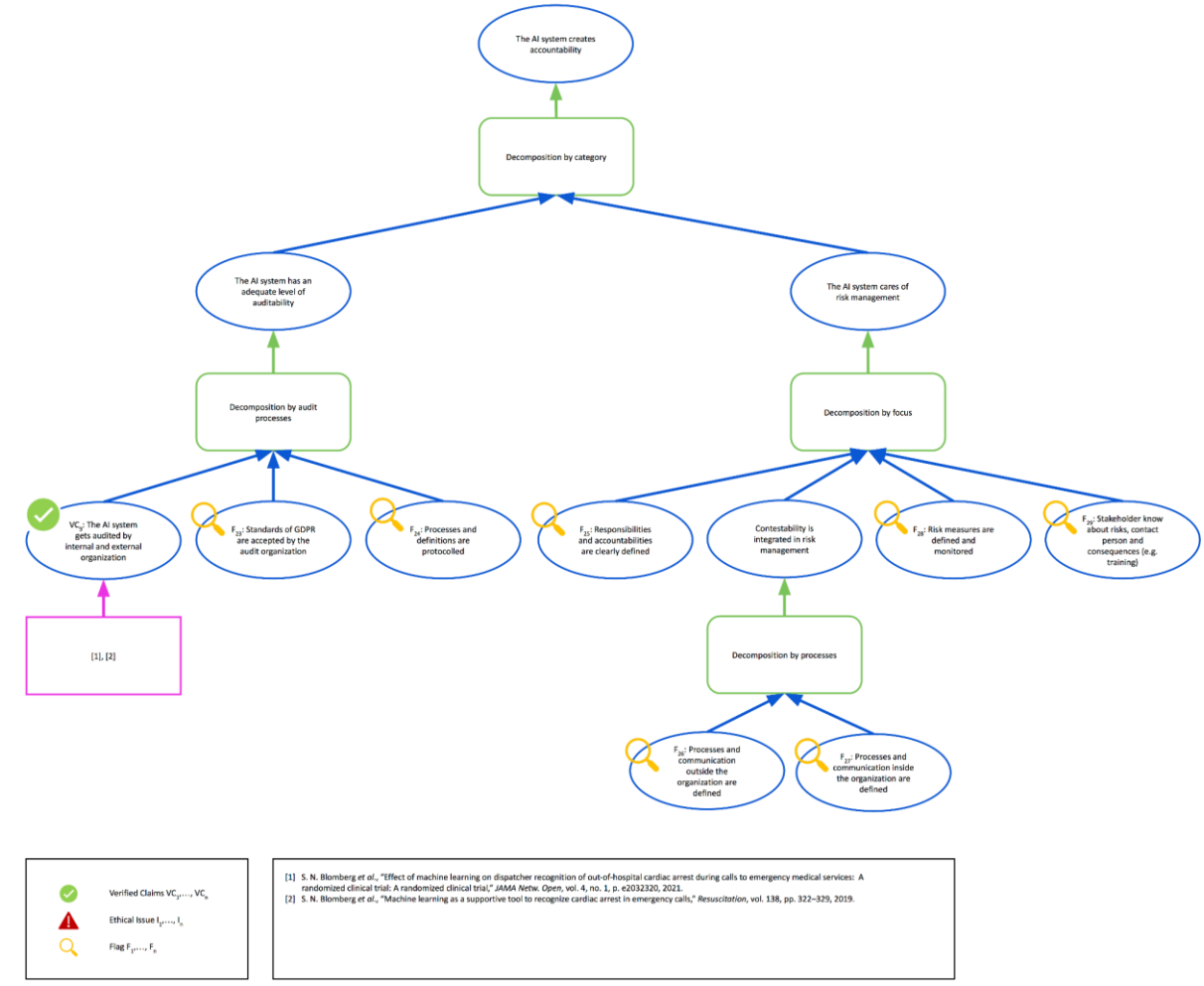
Appendix D: CAE - The AI system cares of societal and environmental wellbeing



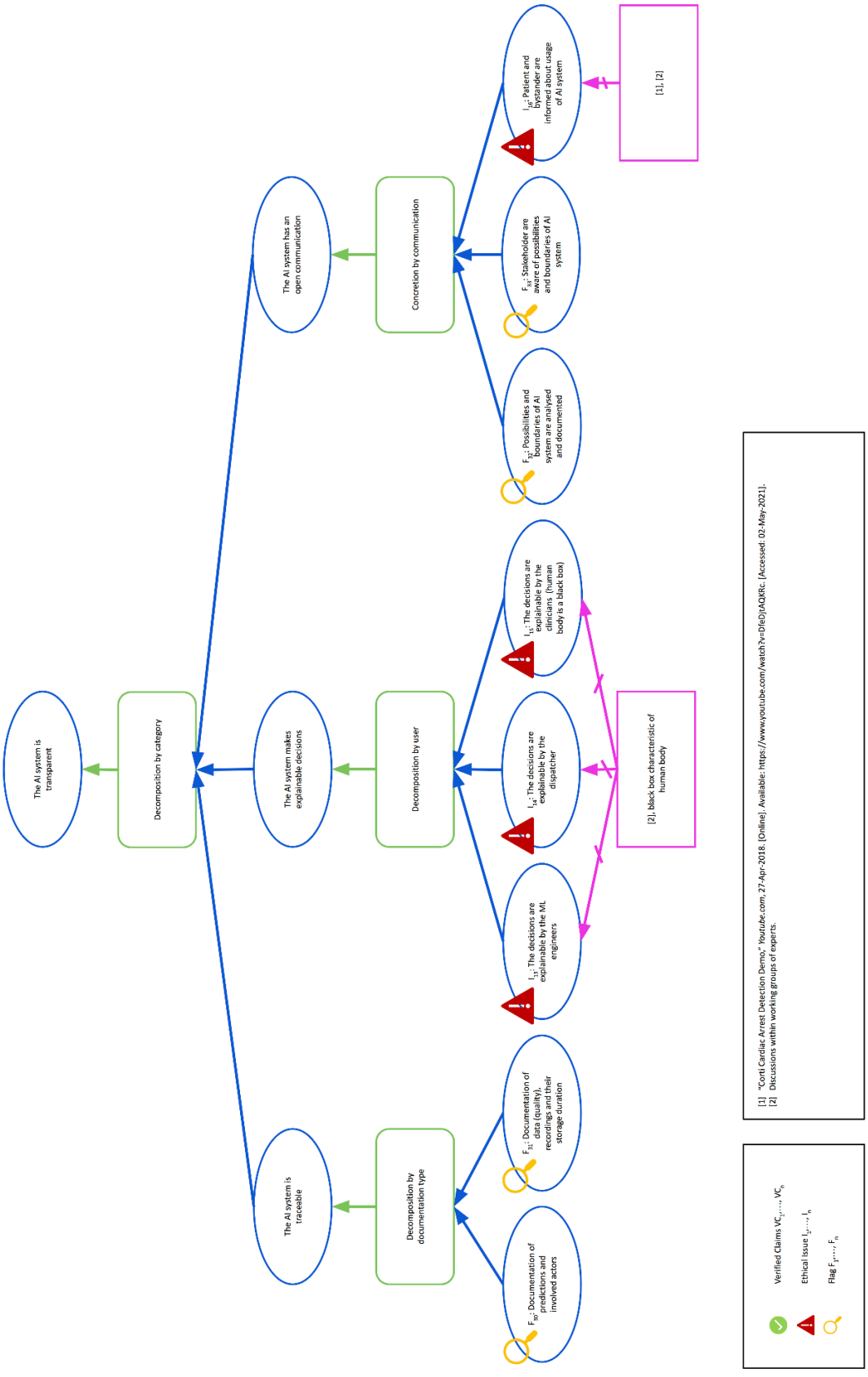
Appendix E: CAE - The AI system respects diversity, non-discrimination and fairness



Appendix F: CAE - The AI System creates accountability



Appendix G: CAE - The AI system is transparent



Appendix H: Data evaluation - Jupyter Notebook

Masterthesis_Analysis_Laufenberg_Werthn

Appendix I: Data evaluation - Comparison of control group, intervention group and total

	Control group (N=314)	**Intervention group (N=306)**	**Total (N=620)**
Patient age (mean, sd)	69.14, 16.27	71.36, 15.62	70.23, 15.98
Patient gender (total, %) Male Female	 214, 68.15% 100, 31.85%	 202, 66.01% 104, 33.99%	 416, 67.10% 204, 32.90%
Caller gender (total, %) Male Female	 101, 32.17% 213, 67.83%	 92, 30.07% 214, 69.93%	 193, 31.13% 427, 68.87%
Caller relation (total, %) Relative Healthcare professional Other	 180, 57.32% 73, 23.25% 61, 19.43%	 160, 52.29% 89, 29.08% 57, 18.63%	 340, 54.84% 162, 26.13% 118, 19.03%
Access of caller to patient (total, %) Patients side Access No Access	 287, 91.40% 21, 6.69% 6, 1.91%	 274, 89.54% 17, 5.56% 15, 4.90%	 561, 90.48% 38, 6.13% 21, 3.39%
Caller alone (total, %)	142, 45.22%	153, 50.00%	295, 47.58%
Area of call (total, %) Private Public (Traffic and Natural area) Other	 257, 81.85% 27, 8.60% 30, 9.55%	 267, 87.25% 23, 7.52% 16, 5.23%	 524, 84.52% 50, 8.06% 46, 7.42%
OHCA recognition by dispatcher (total, %)	282, 89.81%	281, 91.83%	563, 90.81%
DA-CPR instructions started (total, %)	197, 62.74%	198, 64.71%	395, 63.71%
Length of call, min (mean, sd)	6.72, 3.41	6.99, 3.36	6.85, 3.38
Time to recognition of ML model, min (mean, sd)	1.30, 1.46	1.37, 1.32	1.34, 1.39
Time to recognition of dispatcher, min (mean, sd)	1.70, 1.59	1.72, 1.65	1.71, 1.62
Uninstructed time of dispatcher, min (mean, sd)	2.45, 1.92	2.58, 1.83	2.52, 1.87

Appendix J: Description of patients and call(er) characteristics selected for the analysis [146]

Feature	Characteristics	Description
patient_gender	female	The patient is female.
	male	The patient is male.
age_group	<=35	The patient is younger than 36.
	36-64	The patient is between 36 and 64 years old.
	65-79	The patient is between 65 and 79 years old.
	>=80	The patient is older than 79.
caller_relation	Healthcare professional	Caller with medical education.
	Relative	The caller is relative to patient (e.g. spouse or family).
	Other	The caller is neither healthcare professional nor relative to patient.
area_call	Private area	The call is made in the private environment of the patient.
	Natural area	The call is made in a natural environment.
	Traffic area	The call is made in a traffic environment.
	Other area	The call is made neither in a private nor in a natural or in a traffic area.
patient_breathing_normally	No	The caller states that patient is not breathing normally (no breathing or abnormal breathing).
	Yes	The caller states that patient is breathing normally.
patient_conscious	No	The caller states that the patient is not conscious.
	Yes	The caller states that the patient is conscious.
caller_emotional_distressed	No	The caller is emotionally stable.
	Yes	The caller is emotional distressed (e.g. he/she is afraid or crying).
dispatcher_assertive_passive	Assertive	The dispatcher gives instructions instead of questions.
	N/A	No information available (not stated).
	Passive	The dispatcher asks more questions and has a passive role.
caller_gender	female	The caller is female.
	male	The caller is male.

caller_alone	No	The caller is not alone when calling.
	Yes	The caller is alone when calling.
caller_access_patient	Patients side	The caller has visual and physical contact to the patient.
	Access	The caller has visual but no directly physical contact (he/she has to leave the phone to touch the patient).
	No Access	The caller has no access to patient (e.g. he/she is on the way to the patient).
symptom	Asystoli	Our assumption: Stop of heart interval interaction.
	PEA	Our assumption: Pulseless Electrical Activity.
	VF	Our assumption: Ventricular Fibrillation.
	Other	Our assumption: Other Symptoms.

Bitte dieses Formular zusammen mit der Abschlussarbeit abgeben!

Erklärung zur Abschlussarbeit

gemäß § 23, Abs. 12 der Ordnung für den Masterstudiengang Wirtschaftsinformatik vom 14. Februar 2011

Hiermit erkläre ich ~~Herr~~ / Frau

Frederike Laufenberg

Meinen entsprechend gekennzeichneten Anteil der Arbeit habe ich ich selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst.

Frankfurt am Main, den 20.05.2021

F. L.

Unterschrift der Studentin / des Studenten

Bitte dieses Formular zusammen mit der Abschlussarbeit abgeben!

Erklärung zur Abschlussarbeit

gemäß § 23, Abs. 12 der Ordnung für den Masterstudiengang Wirtschaftsinformatik vom 14. Februar 2011

Hiermit erkläre ich ~~Herr~~/ Frau

Teresa Sophia Gabriele Werthmann

Meinen entsprechend gekennzeichneten Anteil der Arbeit habe ich ich selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst.

Frankfurt am Main, den 20.05.2021 Werthmann

Unterschrift der Studentin / des Studenten