On Assessing Trustworthy AI in Healthcare
Best Practice
Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in 112 Emergency Calls for the City of Copenhagen.

- Roberto V. Zicari 1, James Brusseau 21, Stig Nikolaj Blomberg 31, Helle Collatz Christensen 31, Megan Coffee 33, Marianna B. Ganapini 20, Sara Gerke 30, Thomas Krendl Gilbert 28, Eleanore Hickman 11, Elisabeth Hildt 25, Sune Holm 6, Ulrich Kühne 14, Vince Madai 8, Walter Osika 17, Andy Spezzatti 4, Eberhard Schnebel 1, Jesmin Jahan Tithi 32, Dennis Vetter 1, Magnus Westerlund 2, Renee Wurth 10, Julia Amann34, Vegard Antun 7, Valentina Beretta38, John Brodersen 29, Frédérick Bruneault 13, Erik Campano39 , Boris Düdder 3, Alessio Gallucci 9, Emmanuel Goffi 12 , Christoffer Bjerre Haase 16, Thilo Hagendorff 18, Georgios Kararigas 22,Pedro Kringen 1, Florian Möslein36 , Davi Ottenheimer 40, Matiss Ozols 5, Laura Palazzani 35, Martin Petrin 37, Karin Tafur19 , Jim Tørresen23, Holger Volland 24.

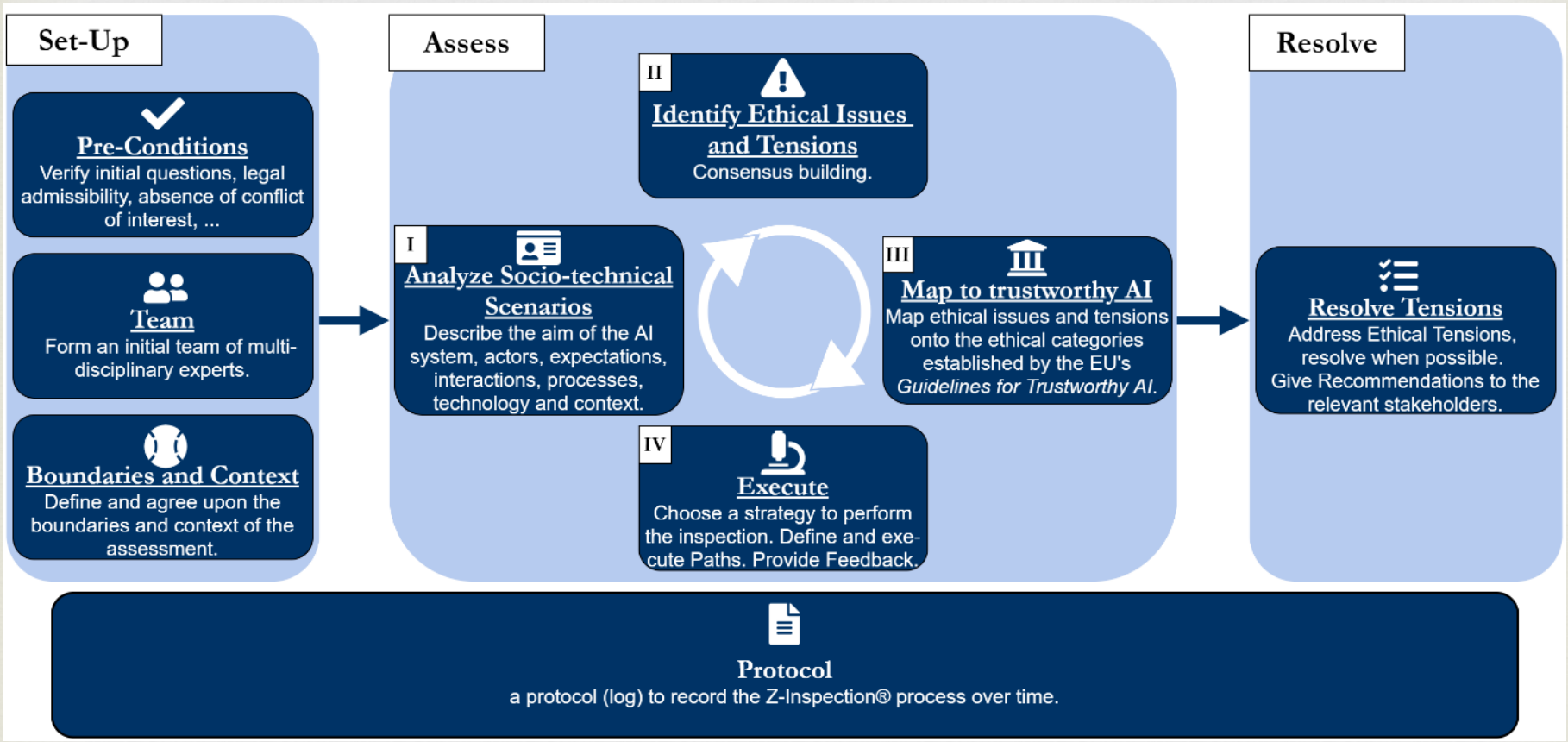
Assessing Trustworthy AI
Best Practice: Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls

- **Health-related emergency calls** (**112**) are part of the Emergency Medical Dispatch Center (EMS) of the **City of Copenhagen**, triaged by medical dispatchers (i.e., medically trained dispatchers who answer the call, e.g., nurses and paramedics) and medical control by a physician on-site (EMS).

Health-related emergency calls (112)

Image https://www.expatica.com/de/healthcare/healthcare-basics/emergency-numbers-in-germany-761525/

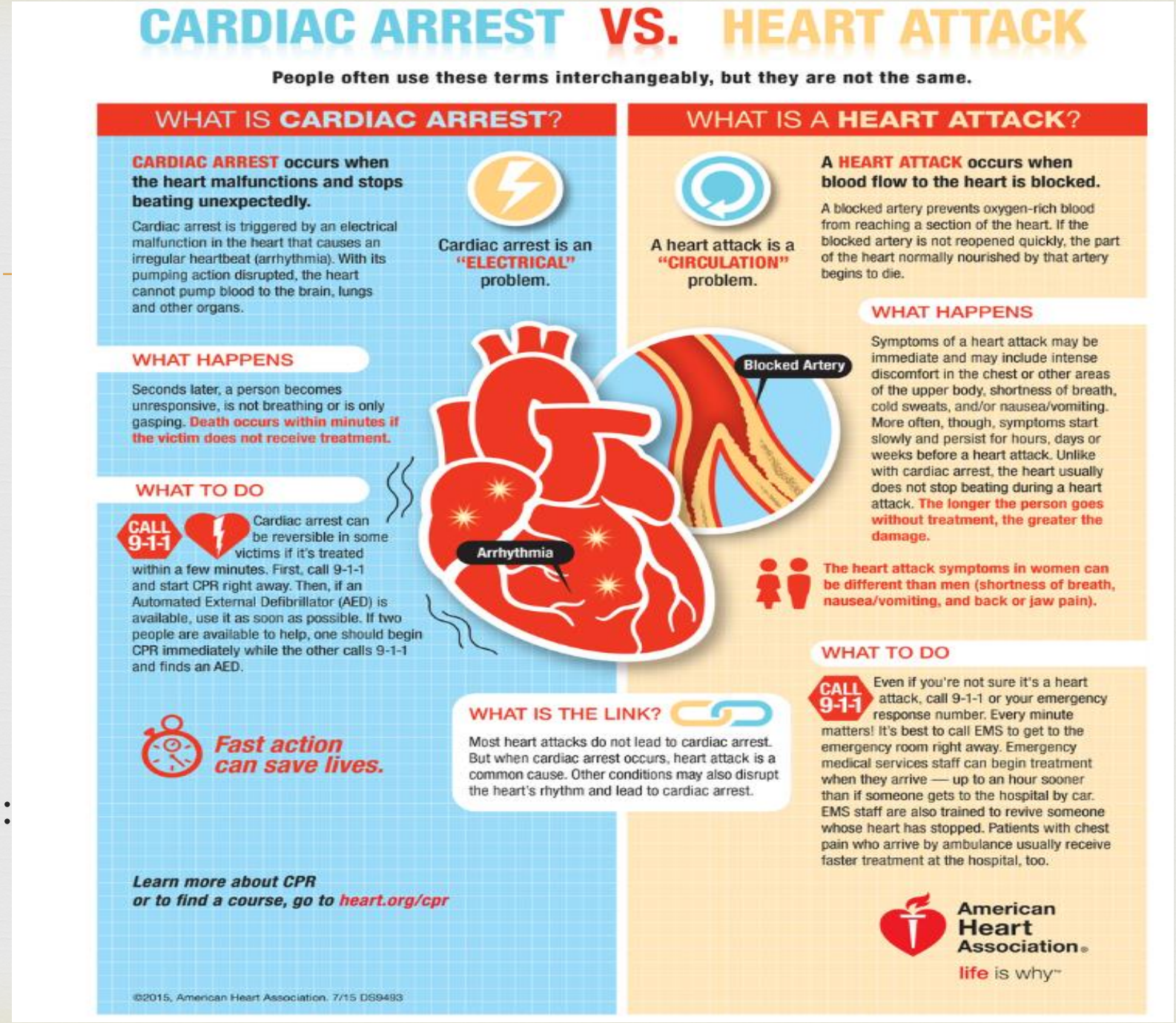
Z-inspection® Process in a Nutshell



The problem

- In the last years, the Emergency Medical Dispatch Center of the City of Copenhagen **has failed to identify approximately 25% of cases of out-of-hospital cardiac arrest (OHCA),** the last quarter has only been recognized once the paramedics/ambulance arrives at the scene .

Image: CPR


CARDIAC ARREST VS. HEART ATTACK
People often use these terms interchangeably, but they are not the same.
WHAT IS CARDIAC ARREST?
CARDIAC ARREST occurs when the heart malfunctions and stops beating unexpectedly.
Cardiac arrest is triggered by an electrical malfunction in the heart that causes an irregular heartbeat (arrhythmia). With its pumping action disrupted, the heart cannot pump blood to the brain, lungs and other organs.
Cardiac arrest is an "ELECTRICAL" problem.
WHAT HAPPENS
Seconds later, a person becomes unresponsive, is not breathing or is only gasping. Death occurs within minutes if the victim does not receive treatment.
WHAT TO DO
CALL 9-1-1
Cardiac arrest can be reversible in some victims if it's treated within a few minutes. First, call 9-1-1 and start CPR right away. Then, if an Automated External Defibrillator (AED) is available, use it as soon as possible. If two people are available to help, one should begin CPR immediately while the other calls 9-1-1 and finds an AED.
Fast action can save lives.
Learn more about CPR or to find a course, go to heart.org/cpr
©2015, American Heart Association. 7/15 DS9493
Arrhythmia
Blocked Artery
WHAT IS THE LINK?
Most heart attacks do not lead to cardiac arrest. But when cardiac arrest occurs, heart attack is a common cause. Other conditions may also disrupt the heart's rhythm and lead to cardiac arrest.
WHAT IS A HEART ATTACK?
A heart attack is a "CIRCULATION" problem.
A HEART ATTACK occurs when blood flow to the heart is blocked.
A blocked artery prevents oxygen-rich blood from reaching a section of the heart. If the blocked artery is not reopened quickly, the part of the heart normally nourished by that artery begins to die.
WHAT HAPPENS
Symptoms of a heart attack may be immediate and may include intense discomfort in the chest or other areas of the upper body, shortness of breath, cold sweats, and/or nausea/vomiting. More often, though, symptoms start slowly and persist for hours, days or weeks before a heart attack. Unlike with cardiac arrest, the heart usually does not stop beating during a heart attack. The longer the person goes without treatment, the greater the damage.
The heart attack symptoms in women can be different than men (shortness of breath, nausea/vomiting, and back or jaw pain).
WHAT TO DO
CALL 9-1-1
Even if you're not sure it's a heart attack, call 9-1-1 or your emergency response number. Every minute matters! It's best to call EMS to get to the emergency room right away. Emergency medical services staff can begin treatment when they arrive — up to an hour sooner than if someone gets to the hospital by car. EMS staff are also trained to revive someone whose heart has stopped. Patients with chest pain who arrive by ambulance usually receive faster treatment at the hospital, too.
American Heart Association.
life is why™

The Problem (cont.)

- Therefore, the Emergency Medical Dispatch Center of the City of Copenhagen **loses the opportunity to provide the caller instructions for cardiopulmonary resuscitation (CPR), and hence, impair survival rates.**
- OHCA is a life-threatening condition that needs to be recognized rapidly by dispatchers, and recognition of OHCA by either a bystander or a dispatcher in the emergency medical dispatch center is a prerequisite for initiation of cardiopulmonary resuscitation (CPR).

Cardiopulmonary resuscitation (CPR)

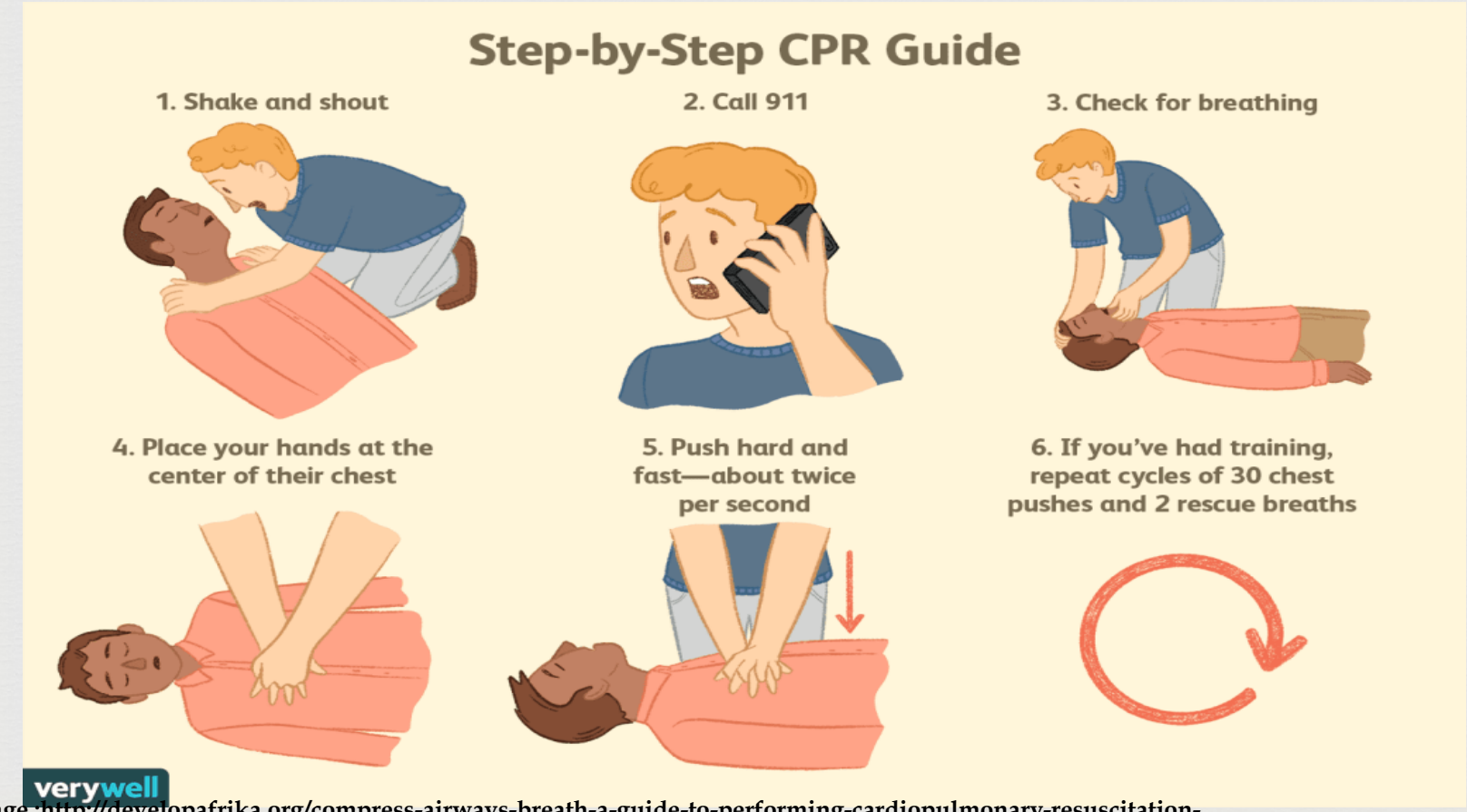


Image :http://developafrika.org/compress-airways-breath-a-guide-to-performing-cardiopulmonary-resuscitation-cpr/?utm_source=ReviveOldPost&utm_medium=social&utm_campaign=ReviveOldPost

Basic Terminology

- **Cardiac arrest** = heart stopped, the cessation of cardiac mechanical activity, as confirmed by the absence of signs of circulation (Perkins, Jacobs, et al., 2015).
- **Acute Coronary Syndromes** = "In those who have ACS, atheroma rupture is most commonly found 60% when compared to atheroma erosion (30%), thus causes the formation of thrombus which block the coronary arteries" (Banning et al., 2020, NICE).
- **Heart attack** = heart clogged (also known as myocardial infarctions) occur when a portion of the heart muscle does not receive adequate blood flow (Benjamin et al., 2017) (NICE).
- **Heart failure** = heart damaged; "is a clinical syndrome characterized by typical symptoms (e.g. breathlessness, ankle swelling and fatigue) (…) caused by a structural and/or functional cardiac abnormality, resulting in a reduced cardiac output" (Ponikowski et al., 2016).
- **Cardiopulmonary Resuscitation (CPR)** = compressions on the upper body to mechanically keep the blood flowing after the heart has stopped beating (Perkins, Handley, et al., 2015).

The Problem (cont.)

- Previous research has identified barriers to the recognition of OHCA (Møller et al., 2016; Sasson Comilla et al., 2010; Viereck et al., 2017).
- Improving early recognition is a goal for both the American Heart Association and the Global Resuscitation Alliance (Callaway et al., 2015; Eisenberg et al., 2018; Nadarajan et al., 2018).

Liability

- Who is responsible is something goes wrong?
- Medical Dispatchers are liable.

The AI solution

- A team lead by Stig Nikolaj Blomberg (Emergency Medical Services Copenhagen, and Department of Clinical Medicine, University of Copenhagen, Denmark) worked together **with a start-up** and examined whether a **machine learning (ML) framework could be used to recognize out-of-hospital cardiac arrest (OHCA) by listening to the calls** made to the Emergency Medical Dispatch Center of the City of Copenhagen.

The AI Solution

- The company designed and implemented the AI system and **trained and tested it by using the archive of audio files of emergency calls** provided by Emergency Medical Services Copenhagen in the year 2014.
- The prime aim of this AI system is **to assist medical dispatchers** when answering 112 emergency calls to help them to early detect OHCA during the calls, and therefore possibly saving lives.

Context and processes, where the AI system is used

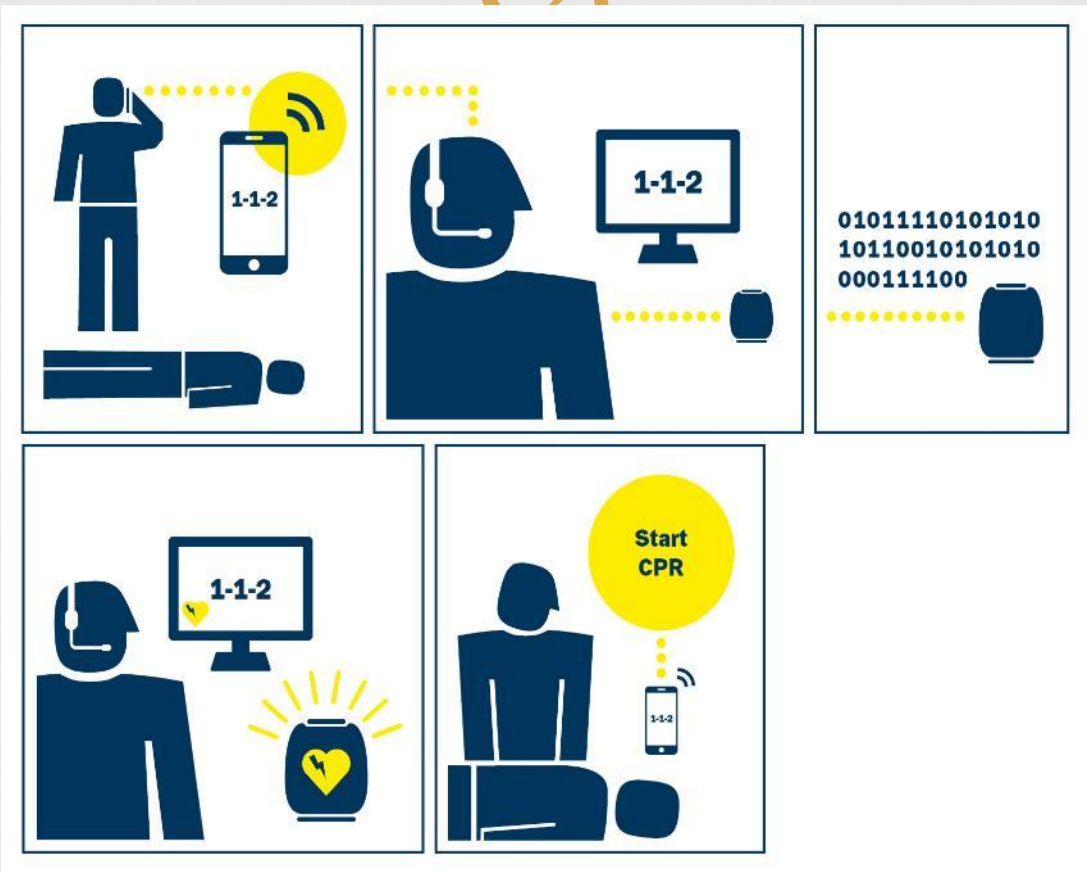


Figure . Ideal Case of Interaction between Bystander, Dispatcher, and the ML System. (with permission from Blomberg, S. N

Retrospective study

- The AI system **performed well in a retrospective study** (108,607 emergency calls audio files in 2014)

Sensitivity, Specificity

- **Sensitivity** (True Positive rate**) measures the proportion of positives that are correctly identified** (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition).

- **Specificity** (True Negative rate) **measures the proportion of negatives that are correctly identified** (i.e. the proportion of those who do not have the condition (unaffected) who are correctly identified as not having the condition).

Source:
https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Positive and Negative predictive values

- **The positive and negative predictive values** (**PPV** and **NPV** respectively) are the proportions of positive and negative results in statistics and diagnostic tests that are true positive and true negative results, respectively.

- PPV= Number of true positives/ Number of true positives + Number of false positives

- https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values

Retrospective study

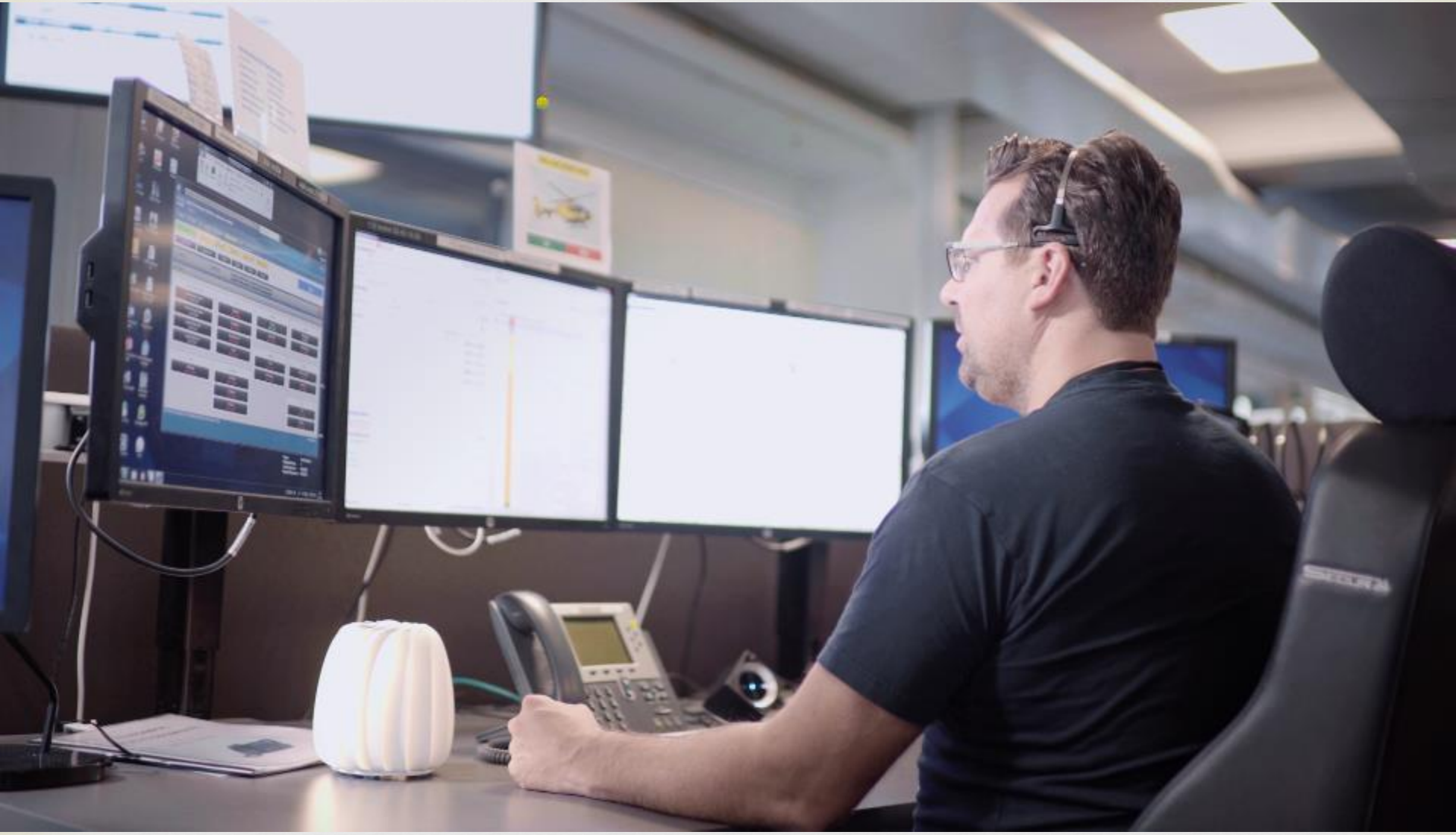
- The machine learning framework had a significantly **higher sensitivity** (72.5% vs. 84.1%, $p < 0.001$) with **lower specificity** (98.8% vs. 97.3%, $p < 0.001$).
- The machine learning framework had a **lower positive predictive** value than dispatchers (20.9% vs. 33.0%, $p < 0.001$). **Time-to- recognition** was significantly shorter for the machine learning framework compared to the dispatchers (median 44 seconds vs. 54 s, $p < 0.001$).

- Source RESUSCITATION 138(2019)322–329 Published 2019

AI System in Production

- **The AI system was put into production during Fall 2020.**

- Note: A responsible person at the Emergency Medical Dispatch Center authorized the use of the AI system.



- https://cordis.europa.eu/project/id/823383/reporting

Randomized clinical trial

- In a **randomized clinical** trial of 5242 emergency calls, a machine learning model listening to calls could alert the medical dispatchers in cases of suspected cardiac arrest.

Published January 2021, *JAMA Netw Open*. 2021;4(1):e2032320. doi:10.1001/jamanetworkopen.2020.32320

Randomized clinical trial (Cont.)

- **There was no significant improvement in recognition of out-of-hospital cardiac arrest during calls on which the model alerted dispatchers vs those on which it did not;** however, the machine learning model had higher sensitivity that dispatchers alone.

Key Questions

- Why dispatchers do not seem to *trust* the AI system?
- Is the AI system helping or harming people?

Motivation of our work.

- This is a self-assessment conducted jointly by a team of independent experts together with the prime stakeholder of this use case. The main motivation of this work is to verify if the rate of lives saved could be increased by using AI, and at the same time to identify possible risks and pitfalls of using the AI system assessed here, and to provide recommendations to key stakeholders.

Aim of the ML system

- We started by analyzing **the prime aim of this AI system**, namely **to assist medical dispatchers** (also referred to as call takers) when answering 112 emergency calls to help them to early detect OHCA during the calls, and increase the potential for saving lives.
- The system has been implemented because OHCA can be difficult for call takers to identify, possibly due to static, language barriers, unclear descriptions by callers, and misunderstandings, along with limited attention spans in calls.
- For OHCA, a specific problem (compared with other 112 calls) is that the caller is never the patient - as they are unresponsive at that time of the call (Safar, P. 1988)- but a bystander (i.e., spouse or passer-by).

Identification of actors

- For this use case, we identified three classes of actors: *primary*, *secondary*, and *tertiary*.
- We define *primary actors* as stakeholders in direct contact with the applied system.
- *Secondary actors* are stakeholders responsible for developing and implementing the system but not using it directly.
- *Tertiary actors* are part of the overall ecosystem where the AI system is used.

Actors

- *The primary actors are:* Stig Nikolaj and his team (who specified the requirements for the design of the AI system and supplied the training and test data) are the prime stakeholder of the use case; the patients; the patients' family members, the callers/bystanders; paramedics and the medically trained dispatchers who answer the call.

- *The secondary actors are*: the AI vendor, a start-up company, independent from the owner of the case who designed, implemented, and deployed the AI system. The CEO of the Emergency Medical Services who gave permission to put the system into deployment.

- *The tertiary actors are*: the Copenhagen Emergency Medical Services (EMS), which is an integrated part of the Health Care System for the Capital Region of Denmark, consisting of one hospital trust with six university hospitals in nine locations and one emergency medical service (Lippert 2018).

Actors Expectations and Motivations

The actors listed above share one common goal: saving the patient's life. Aside from this goal, the actors have some distinct expectations and motivations:

- *Caller/bystander:* receive easy to understand and follow instructions to help patient;
- *Dispatcher/call taker:* provide targeted support and instructions to caller based on correct information;
- *Paramedics:* receive correct information to be well prepared upon arrival to care for the patient;
- *Patients' family members:* know that everything was done to save the patient's life and that no error occurred in the process (human or machine); if the patient dies, they may look for someone to blame (the dispatcher/paramedic/AI system?);
- *AI vendor:* profit, reputation, satisfied clients, avoid malfunctioning of the system leading to poor performance (death of the patient);
- *Hospital system:* improve efficiency and efficacy (i.e., number of lives saved due to the system), reputational gains; and
- *Public Health System in Denmark:* improve efficiency and efficacy (i.e., number of lives saved due to the system).

AI pilot testing

- The system was introduced to the call takers by the primary investigator of research (i.e., the owner of the use case), who participated in four staff meetings, each of them consisting of an hour training session. During these sessions, the AI system was presented as well as the objectives of the research and the protocol the dispatchers should follow in case of an alert.
- There was a one-month pilot testing where none of the alerts were randomized. This was performed to allow most of the dispatchers to experience an alert prior to the randomization start. During this month, the primary investigator was present at the Emergency Medical Dispatch Center of the City of Copenhagen and available for dispatchers to question.

Develop of an evidence base

- It is important at this point to review and create an evidence base that we will use to verify/support any claims made by the producer of the AI system and other relevant stakeholders.

Develop of an evidence base

For this case, we summarize here the most relevant findings.

- **OHCA is a major health care and socioeconomic problem** with a total survival rate of generally below 10 percent (Berdowski et al., 2010; Gräsner et al., 2020; Virani et al., 2020). Time is of the essence when treating OHCA, with chances of survival decreasing rapidly in the first minutes after collapse.
- **Efficient emergency medical services should detect cardiac arrest within the first minute** (Perkins et al., 2015). Correct diagnosis and treatment are needed within minutes in order to increase the odds to attain a successful resuscitation**. Every minute without resuscitation decreases the probability of survival by ~10% and increases the risk of side-effects, such as brain damage** (Murphy et al., 1994).

Develop of an evidence base

- The chance of survival decreases rapidly after the onset of OHCA until the initiation of resuscitation efforts (CPR or defibrillation), with models illustrating a decrease of roughly 10% per minute, leaving close to zero percent chance of survival 15 minutes after collapse. Due to the loss of circulation following OHCA, imminent treatment is of the essence since the chance of survival rapidly decreases with increased time from collapse to treatment (Cummins et al., 1991; Hasselqvist-Ax et al., 2015; Larsen et al., 1993; Monsieurs et al., 2015).

Develop of an evidence base

- Survivors of OHCA may sustain brain injury due to inadequate cerebral perfusion during cardiac arrest. Anoxic brain damage after OHCA may result in a need for constant care or assistance with activities of daily living. Persons with anoxic brain damage may therefore require nursing home care after discharge (Middelkamp et al., 2007; Moulaert et al., 2009).

Context and processes, where the AI system is used

- We look at the context and process where the AI system is used, including **the interactions of actors with each other and with the ML.**

Context and processes, where the AI system is used

- The AI **system is listening** in to all calls made to the emergency medical services 112 emergency line. This includes calls for various other reasons (e.g., car accidents); **OHCA is only responsible for ~1% of all calls.**
- With ~65 dispatchers, every dispatcher only encounters 10-20 cases of cardiac arrest per year on average. It is reported that 1/2 of the human alerts were true cardiac arrests, ⅕ of the machine alerts were true cardiac arrests (Blomberg et al., 2021).

The technology used

- The prime stakeholder commissioned **an external start-up company to implement the AI system** because they discovered that off-the-shelf solutions did not work, due to poor sound quality of the calls, and abnormal vocals (e.g., emotionally distressed, shouting, etc.). Also, at that time (2018), no Danish language model was readily available.

The technology used

- For this use case, the ML system was designed and implemented with the **expectation to detect cardiac arrest in calls faster and more reliably than human operators.**
- An initial confirmation of this assumption was reported in a retrospective study conducted by the prime stakeholders (Blomberg et al., 2019).

The technology used

- **They used a language model for translating the audio to text based on a convolutional deep neural network** (LeCun et al 1989).
- The ML model was trained and tested on datasets of audio files of calls to the 112 emergency line made in 2014, provided by the prime stakeholder to the company.
- **Only the audio was used, so other personal data was explicitly not used.**

The technology used

- The text output of the language model was then fed **to a classifier that predicted whether a cardiac arrest was happening or not (Figure 3).**
- The AI system was applied directly on the audio stream where the only processing made was a short-term Fourier transformation (Havtorn et al., 2020), hence no explicit feature selection was made.

The technology used

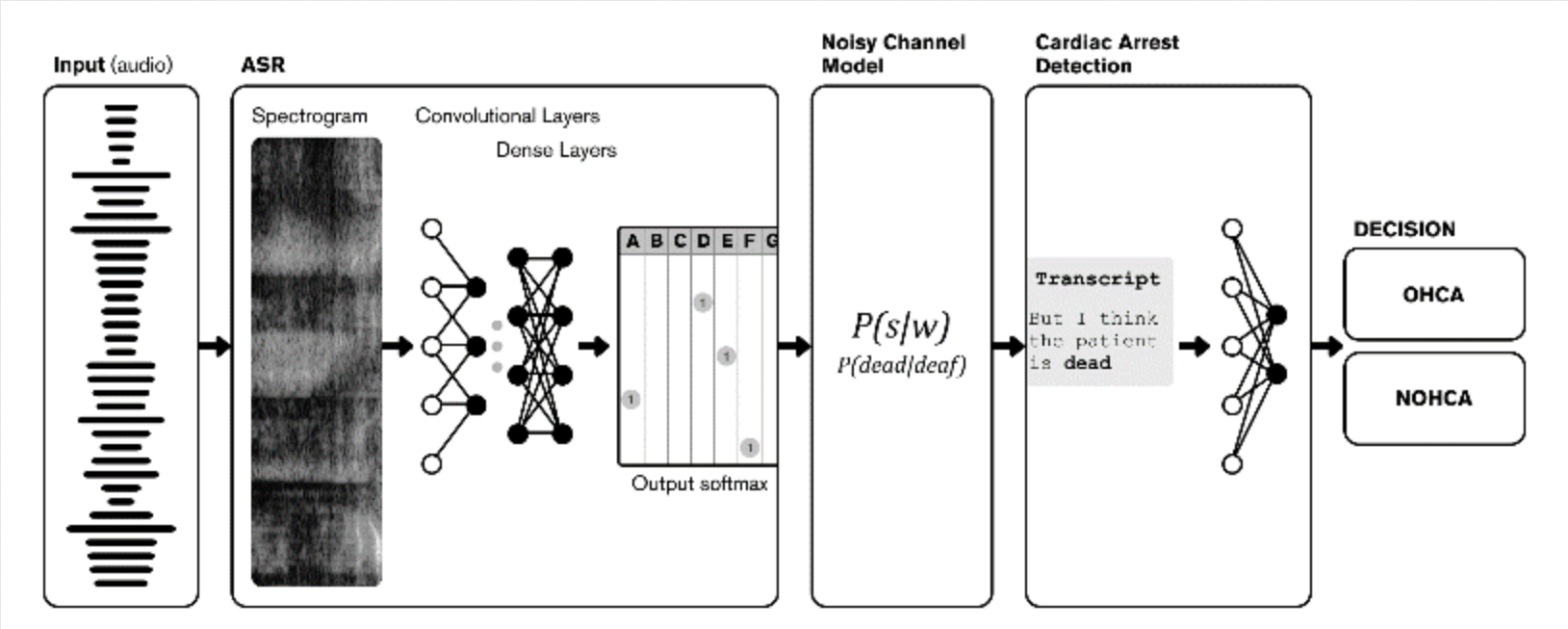


Figure 3 ML Data Flow (with permission of Stig Blomberg)

The technology used

- **The predictive model, working only on the text output of the automatic speech recognition model, was predicted based on the raw textual output**.
- When an emergency call was analyzed in real-time by the ML framework, the audio file was processed without any prior editing or transcription and **transformed to a textual representation of the call**, which was then analyzed and outputted as a prediction of cardiac arrest (Blomberg et al., 2021).

The technology used

- Using a **Danish language model** means that calls in other languages were interpreted in a way that the cardiac arrest model could not work with (i.e., trying to understand Danish words from English speech). In many cases, the model understood the calls anyways, **but in some cases not.**
- **So far, there is no explanation why some calls were seemingly not understood**.

The technology used

- **There is no explanation of how the ML makes its predictions.**
- The company that developed the AI system has some of their work in the open domain (Havtorn et al., 2020; Maaløe et al., 2019). However, the exact details on the ML system used for this use case are not publicly available.

The technology used

- The general principles used for this AI system are documented in the study by Borgholt L. 2020. The paper describes the AI model implemented for this use case.
- However, the paper presents the model trained using different data sets and therefore the results are not representative for this use case.
- **The details of the implementation of the AI system for this case are proprietary, and therefore not known to our team.**

Identification of possible ethical legal and technical issues

- In this step of the Assess Phase, we identified possible **ethical and technical and legal issues** for the use of the AI within the given boundaries and context (see list of *actors* above).

- For some **ethical issues, a tension** may occur.

Tensions

- We use the definition of *tension* from Whittlestone et al. (2019), which refers to different ways in which values can be in conflict
- – i.e., **tensions between the pursuit of different values in technological applications rather than an abstract tension between the values themselves.**

Tensions in the evidence base

- There is **a tension** between the conclusions from the **retrospective study** (Blomberg et al., 2019), indicating that **the ML framework performed better than emergency medical dispatchers** for identifying OHCA in emergency phone calls - and therefore with the expectation that the ML could play an important role as a decision support tool for emergency medical dispatchers- ,
- and the results of **a randomized control trial** performed later (September 2018 – January 2020) (Blomberg et al., 2021), **which did not show any benefits in using the AI system in practice.**

Possible lack of trust

- For our assessment, it is important to find out **whether and how the ML system influences the interaction between the human actors**,
- i.e., how it influences the conversation between the caller/bystander and the dispatcher, the duration of the call, and the outcome, and why during the clinical trial the use of the AI system did not translate into improved cardiac arrest recognition by dispatchers (Blomberg et al. 2021).

Lack of Trust?

- Some possible hypotheses that need to be verified:

- **The dispatcher possibly did not trust the cardiac arrest alert.** It might depend on how the system was introduced – how the well-known cognitive biases were presented/labeled – if the use of the system was labeled as a learning opportunity for the dispatcher, and not as a failure detection aid, that would disclose the incompetence of the dispatcher.

Lack of Trust?

- But it could be that **dispatchers did not sufficiently pay attention to the output of the machine**.
- It relates to the principle of ***human agency and oversight*** in trustworthy AI .
- Why exactly is this?

Lack of Trust?

- If one of the reasons why dispatchers are not following the system to the desired degree is that **they find the AI system to have too many false positives**, then this issue relates to the challenge of achieving a satisfactory interaction outcome between dispatchers and system.

Lack of Trust?

- Another tension concerns **whether dispatchers should be allowed to overrule a positive prediction made by the system and not just merely overrule a negative prediction by the system**.
- In particular, what exactly is the right interplay or form of interaction between system and human, given the goals of using the system and the documented performance of human and system?

Medical benefits – risks versus benefits

- *Possible risks and harm: false positives and false negatives*

- One of the biggest **risks** for this use case is **where a correct dispatcher would be overruled by an incorrect machine.**

Medical benefits – risks versus benefits

- We could not find a justification for choosing a certain **balance between sensitivity and specificity.**

- **If *specificity* is too low,** CPR is started on people who do not need it and administered CPR over a longer period of time can break the rib cage. However, it is unlikely that CPR would be performed on a conscious patient for a longer time, as the patient probably would fight back against it.

Medical benefits – risks versus benefits

- **If *sensitivity* is too low**, cardiac arrests may not be detected. This results in no CPR being administered and the patient remains dead.
- In this context "too low" is when the machine performs poorer than the dispatchers, hence will not be of any help.

Ethical tensions related to the design of the AI system

- ***Lack of explainability***

- The main issue here is that it is not apparent to the dispatchers how the system comes to its conclusions. **It is not transpa**rent to the dispatcher whether it is advisable to follow the system or not. Moreover, it is not transparent to the caller that an AI system is used in the process.

Diversity, non-discrimination, and fairness: possible bias, lack of fairness

- It was reported in one of the workshops that if the caller was not with the patient, such as in another room or in a car on their way to the patient, the AI system had more false negatives.
- **The same was found for people not speaking Danish or with a heavy dialect.**

Bias ,Fairness

- For this use case, concepts such as **“bias”** and **“fairness”** are domain-specific and should be considered at various levels of abstractions (e.g., from the viewpoint of the healthcare actors down to the level of the ML model).

Bias,Fairnes

- We look at possible bias in the use of the AI system. The AI system was only trained on Danish data, but the callers spoke more languages (i.e., English, German)**. Here, there is a risk of bias, as the system brings disadvantages for some groups, such as non-Danish speaking callers, callers speaking dialects, etc.**

Discrimination

- When we looked at the **data used to train the ML model,** we observed that the dataset used to train the ML system was created by collecting data from the Copenhagen Emergency Medical Services from 2014. The AI system was tested with data from calls between September 1, 2018, and December 31, 2019. **It appears to be biased toward older males, with no data on race and ethnicity.**

Liability

- For this use case, a problem is the **responsibility and liability of the dispatcher.**
- **What are the possible legal liability implications for ignoring an alert coming from a ML system?**

- The consequences of refuse or acceptance of an alert are central.

Risk of de-skilling

- There is a need of justification of choice: in this field, **the risk of de-skilling is possible (technological delegation also in order not to be considered reliable for ignoring/refusing it);** we also need to think about the cultural level of a dispatcher and the ethical awareness of the consequences of his/her choice:
- How could he/she decide against the machine? Sometimes it could be easier to accept than to ignore/refuse for many reasons.

Risk of alert fatigue

- In the randomized clinical trial it was reported that less than one in five alerts were true positives.

- **Such low sensitivity might lead to alert fatigue, and in turn, ignoring true alerts.**

- "The term **alert fatigue** describes how busy workers (in the case of health care, clinicians) become desensitized to safety alerts, and as a result ignore or fail to respond appropriately to such warnings"
- Source: https://psnet.ahrq.gov/primer/alert-fatigue

The legal framework

- The ML model in this use case is used by medical personnel to guide them in making an evaluation of the patient so that they can act accordingly.
- This process is not fully described by the Danish Health Act *Sundhedsloven* , but it is described to some extent by the Patient Danish Authority (STPS).

The legal framework

- As the use of AI in health services is fairly new, the Danish authorities have apparently (as mentioned to us by the prime stakeholder) not yet decided how this new technology be regulated.

- Our expert team was informed by the prime stakeholder that **the AI system does not have a CE-certification as a medical device.**

The legal framework

- Since the AI system processes personal data, the **General Data Protection Regulation** (GDPR) applies, and the prime stakeholder must comply with its requirements.

- From a data protection perspective, **the prime stakeholder** of the use case is in charge of fulfilling the legal requirements.

- From a risk-based perspective, it would be desirable if the **developers** of the system would also be responsible as they implemented the AI system. **But the responsibility of the vendors or developers of a system is not a requirement of the GDPR.**

Societal and environmental well-being

- We consider here broader implications, such as additional costs that could arise from an increase in false positives by the AI/ML system, resulting in unnecessary call taker assisted CPRs, and dispatching ambulances when they are not necessary, and trade-offs, by detracting resources from other areas.

Map Ethical issues and Flags to Trustworthy AI Areas of Investigation

- The basic idea of the Z-inspection® process in this step is to map the above list of "issues" described with an *open vocabulary*, to some or all of the seven requirements for trustworthy AI.
- We guide the discussion to reach a consensus by using a *closed vocabulary*, i.e., using the four ethical principles and the seven requirements for trustworthy AI.

Four pillars of the AI HLEG trustworthy AI guidelines

- Respect for Human Autonomy,
- Prevention of Harm,
- Fairness,
- Explicability

7 Requirements

- REQUIREMENT #1 **Human Agency and Oversight**

Sub-requirements:

  - *Human Agency and Autonomy*
  - *Human Oversight*

- REQUIREMENT #2 **Technical Robustness and Safety**

Sub-requirements:

  - *Resilience to Attack and Security General Safety*
  - *Accuracy*
  - *Reliability,*
  - *Fall-back plans and Reproducibility*

- REQUIREMENT #3 **Privacy and Data Governa**nce

Sub-requirements:

  - *Privacy*
  - *Data Governance*

7 Requirements (cont.)

- REQUIREMENT #4 **Transparency**

Sub-requirements:

  - *Traceability*
  - *Explainability*
  - *Communication*

- REQUIREMENT #5 **Diversity, Non-Discrimination and Fairness**

Sub-requirements:

  - *Avoidance of Unfair Bias*
  - *Accessibility and Universal Design*
  - *Stakeholder Participation*

7 Requirements (cont.)

- REQUIREMENT #6 **Societal and Environmental Well-Being**

Sub-requirements:

  - *Environmental Well-Being*
  - *Impact on Work and Skills*
  - *Impact on Society at Large or Democracy*

- REQUIREMENT #7 **Accountability**

Sub-requirements:

  - *Auditability*
  - *Risk Management*

Catalog of predefined ethical tensions

- To help the process, especially as a help to experts who might have not sufficient knowledge in ethics, we used a sample of catalog of predefined ethical tensions.
- We have chosen the catalog defined by the Nuffield Foundations (Whittlestone et al., 2019)

Catalog of predefined ethical tensions

- Quality of services *versus* privacy;
- Personalisation *versus* solidarity;
- Convenience *versus* dignity;
- Privacy *versus* transparency;
- Accuracy *versus* explainability;
- Accuracy *versus* fairness;
- Satisfaction of preferences *versus* equality;
- Efficiency *versus* safety and sustainability.

Source: *Sample Catalog of Ethical Tensions* (Whittlestone et al., 2019)

Ethical tensions

- When a specific “issue” did not correspond to one or more of the predefined ethical tensions, experts described them with their own words.

Classification of ethical tensions

From [1]:

- **true dilemma**, i.e. "a conflict between two or more duties, obligations, or values, both of which an agent would ordinarily have reason to pursue but cannot";
- **dilemma in practice**, i.e. "the tension exists not inherently, but due to current technological capabilities and constraints, including the time and resources available for finding a solution";
- **false dilemmas**, i.e. "situations where there exists a third set of options beyond having to choose between two important values".

ID Ethical Issue: E4, Fairness in the Training Data.

- **Description:**

The training data is likely not sufficient to account for relevant differences in languages, accents, and voice patterns, potentially generating unfair outcomes.

MAP TO ETHICAL Pillars/Requirements/Sub-requirements (closed vocabulary):

- **Fairness** > *Diversity, Non-Discrimination*

and

- **Fairness** > *Avoidance of Unfair Bias*

NARRATIVE RESPONSE

- There is likely empirical bias since the tool was developed in a predominantly white Danish patient group. It is unclear how the tool would perform in patients with accents, different ages, gender, and other specific subgroups.
- There is also a concern that this tool is not evaluated for fairness with respect to outcomes in a variety of populations. Given the reliance on transcripts, non-native speakers of Danish may not have the same outcome. It was reported that Swedish and English speakers were well represented but would need to ensure a broad training set. It would also be important to see if analyses show any bias in results regarding age, gender, race, nationality, and other sub-groups. The concern is that the training data may not have a diverse enough representation.

ID Ethical Issue: E5, Potential harm resulting from tool performance.

- **Description:**

The tool's characteristic performance, such as a higher rate of false positives compared to human dispatchers, could adversely affect health outcomes for patients.

MAP TO ETHICAL Pillars/Requirements/Sub-requirements (closed vocabulary):

- **Prevention of Harm**

> *Technical Robustness and Safety* > *Accuracy*

NARRATIVE RESPONSE

- The algorithm did not appear to reduce the effectiveness of emergency dispatchers but also did not significantly improve it. The algorithm, in general, has a higher sensitivity but also leads to more false positives. There should be a firm decision on thresholds for false positive versus false negatives. The risk of not doing CPR if someone needs CPR exceeds the risk of doing CPR if not needed. On the other hand, excessive false positives put a strain on healthcare resources by sending out ambulances and staff to false alarms. This potentially harms other patients in need of this resource. The gold standard to assess whether the tool is helpful for the given use case is to analyze its impact on outcome. Given, however, the low likelihood of survival from out of hospital cardiac arrest, there wasn't an analysis attempting to assess the impact on survival, as it would take years in a unicentric study.

ID Ethical Issue: E6, The AI tool is not interpretable.

- **Description:** The system outputs cannot be interpreted, leading to challenges when dispatcher and tool are in disagreement.

- **Explicability** > *Transparency* > *Explainability*

NARRATIVE RESPONSE

- The tool lacks explainability, which might lead to several challenges. First, outcomes are based on a transcription of the conversation between dispatcher and caller. It is not clear what is used from these transcripts to trigger an alert. This lack of transparency may have contributed to the noted lack of trust among the dispatchers, as well as the limited training of the users. Second, there is a lack of transparency regarding whether and which value judgments went into the design of the model. Such value judgments are important because explaining the output is partly a matter of accounting for the design decisions that humans have made.

The Resolve Phase Verification of Requirements

- Start from the list of consolidated ethical and technical and legal issues, **priorize them by urgency.**
- Verify claims, using a mixed approach, consisting in adapting concepts from the **Claims, Arguments, Evidence (CAE) framework** and using the **ALTAI web tool.**
- As result (revise) the “issues” and give recommendations to relevant stakeholders.

Recommendations to the key stakeholders

- The output of the assessment will be a report containing recommendations to the key stakeholders. Such recommendations should be considered as a source of qualified information that help decision makers make good decisions, and that help the decision-making process for defining appropriate trade-offs. They would also help continue the discussion by engaging additional stakeholders in the decision- process.

EU guidelines for trustworthy AI

- In order to bring some clarity and define a general framework for the use of AI Systems, the High-Level Expert Group on AI (AI HLEG) set up by the European Commission published ethics guidelines for trustworthy AI in April 2019 (AI HLEG, 2019).
- These guidelines are aimed at a variety of stakeholders, especially guiding practitioners towards more ethical and more robust applications of AI.

Trustworthy artificial intelligence

EU High-Level Expert Group on AI presented their ethics guidelines for *trustworthy* artificial intelligence:

- (1) **lawful** - respecting all applicable laws and regulations
- (2) **ethical** - respecting ethical principles and values
- (3) **robust** - both from a technical perspective while taking into account its social environment

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

European Commission. Independent High-Level Experts Group on AI.

Four ethical principles, rooted in fundamental rights

(i) Respect for human autonomy

(ii) Prevention of harm

(iii) Fairness

(iv) Explicability

- There may be **Tensions** between the principles

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Resources

Z-Inspection®: A Process to Assess Trustworthy AI.

Roberto V. Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas , Pedro Kringen, Melissa McCullough, Florian Möslein, Karsten Tolle, Jesmin Jahan Tithi, Naveed Mushtaq, Gemma Roig , Norman Stürtz, Irmhild van Halem, Magnus Westerlund.

IEEE Transactions on Technology and Society, (Early Access)

Print ISSN: 2637-6415

Online ISSN: 2637-6415

Digital Object Identifier: 10.1109/TTS.2021.3066209

Date of Publication: 17 March 2021

Link to Download the paper:

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9380498

http://z-inspection.org