How to Assess Trustworthy AI *with* Z-inspection®

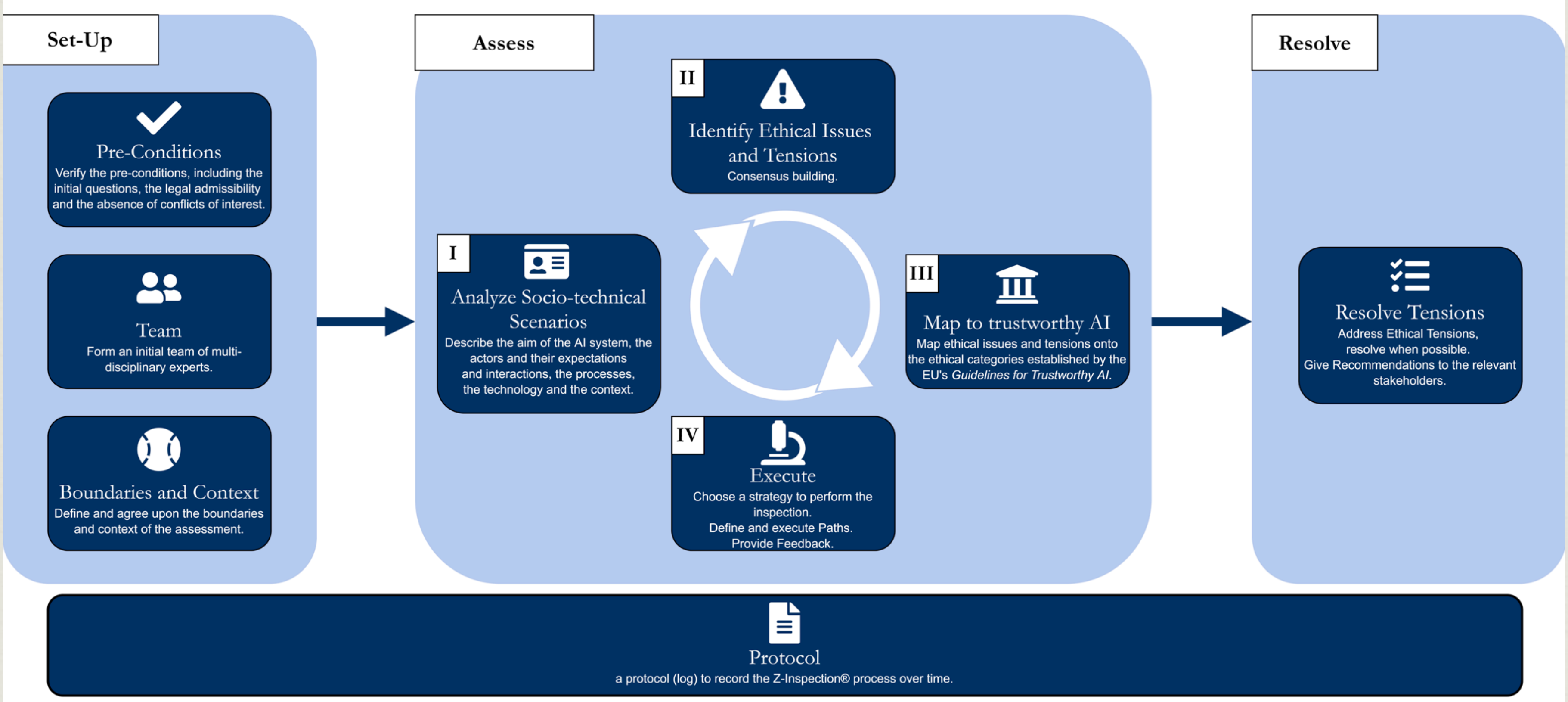
Roberto V. Zicari

Z-Inspection® Initiative

http://z-inspection.org

GSDS SNU, Seoul, November 3, 2022

Z-inspection® Process in a Nutshell



Set-Up

Pre-Conditions
Verify the pre-conditions, including the initial questions, the legal admissibility and the absence of conflicts of interest.

Team
Form an initial team of multi-disciplinary experts.

Boundaries and Context
Define and agree upon the boundaries and context of the assessment.

Assess

I
Analyze Socio-technical Scenarios
Describe the aim of the AI system, the actors and their expectations and interactions, the processes, the technology and the context.

II
Identify Ethical Issues and Tensions
Consensus building.

III
Map to trustworthy AI
Map ethical issues and tensions onto the ethical categories established by the EU's *Guidelines for Trustworthy AI.*

IV
Execute
Choose a strategy to perform the inspection.
Define and execute Paths.
Provide Feedback.

Resolve

Resolve Tensions
Address Ethical Tensions, resolve when possible.
Give Recommendations to the relevant stakeholders.

Protocol
a protocol (log) to record the Z-Inspection® process over time.

Lessons Learned

There may be tensions in building a stronger evidence base on the current uses and impacts (*domain specific*)

- **Different View Points among Domain Experts**
- **Who is "qualified" to give a strong evidence?**

Identifying further tensions

„Thinking about tensions could also be enhanced by systematically considering different *ways* that tensions are likely to arise.

We outline some conceptual lenses that serve this purpose“

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Identifying further tensions

- Whittlestone et al. : points to three axis for such considerations: **power, time and locus.**

Identifying further tensions *cont.*

- ***Winners versus losers***. *Tensions sometimes arise because the costs and benefits of ADA-based technologies are unequally distributed across different groups and communities.*
- ***Short term versus long term***. *Tensions can arise because values or opportunities that can be enhanced by ADA-based technologies in the short term may compromise other values in the long term.*
- ***Local versus global***. *Tensions may arise when applications that are defensible from a narrow or individualistic view produce negative externalities, exacerbating existing collective action problems or creating new ones.*

Identifying further tensions *cont.*

This helps to flag the broader issues linked to

Winners versus *Losers* i.e. **power,**
Short/Long term i.e. **time**
and
Local versus Global - i.e. **the locus** or **scope**

- or the "good for whom?" question.

Mappings to the framework of trustworthy AI

- The "*issues*" described in free text should be then mapped using templates (called rubrics) to some of the four ethical principles and the seven requirements (sub-requirements) defined in the EU framework for trustworthy AI.

Mappings to the framework of trustworthy AI

- With this mapping the reports developed **from an open vocabulary to a closed vocabulary** (i.e. the templates). We call these *mappings*.

- There are different strategies to perform such *mappings*.

Seven Requirements (and sub-requirements) for Trustworthy AI

1 Human agency and oversight

Including fundamental rights, human agency and human oversight

2 Technical robustness and safety

Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility

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

Seven Requirements (and sub-requirements) for Trustworthy AI

3 Privacy and data governance

Including respect for privacy, quality and integrity of data, and access to data

4 Transparency

Including traceability, explainability and communication

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

Seven Requirements (and sub-requirements) for Trustworthy AI

5 Diversity, non-discrimination and fairness

Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

6 Societal and environmental wellbeing

Including sustainability and environmental friendliness, social impact, society and democracy

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

Seven Requirements (and sub-requirements) for Trustworthy AI

7 Accountability

Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

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

Mapping: Use of Rubrics

- **ID Ethical Issue:**
- **Short Description:**
- **Narrative Response:**
- **MAP TO Ethical Pillars / Requirements/ Sub-requirements (closed vocabulary):**

Example

- *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.

Example *cont.*

- *Narrative Response* (*Open Vocabulary*)

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, sex, 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.

Example *cont.*

- *Map to Ethical Pillars/Requirements/Sub-Requirements (Closed Vocabulary)*

Fairness

> *Diversity, Non-Discrimination and Fairness*

> *Avoidance of Unfair Bias*

Examples of a Mapping Strategy

In one working group (**WG Ethics and Healthcare**), the mapping of issues identified in the WG report was organized using the following process:

Examples of a Mapping Strategy

- At the **initial meeting**, they made a list of the key issues that they found to be present in the WG report. The list merely stated key words, no description of the issues. They then divided the issues between them and each member of the group made a description of her selection of the issues. The descriptions formed the basis of another meeting at which they initiated the mapping of issues to ethical pillars, requirements and sub requirements.

Example of a Mapping Strategy, *cont.*

At the **second meeting** they discussed the mapping of a couple of the issues identified. This involved quite a bit of clarification and discussion of their understanding of the pillars and requirements.

Moreover, the discussion of what was covered by the pillars and requirements shaped and structured the way they understood the issues. At the meeting they did not get around to mapping all the issues to the pillars, requirements, and sub-requirements.

Example of a Mapping Strategy, *cont*

Instead they decided that they would each map the issues they had described and then meet and discuss these suggested mappings.

They did this at the **3rd meeting**. At this point they seemed to have reached a common understanding of the pillars and requirements as well as of the issues described.

Example of consolidation process of mappings *cont*.

- The consolidated lists of WG issues for each of the seven requirements were reviewed so commonalities and differences could be identified and discussed before final consolidation.
- The method highlighted how different perspectives could lead to similar issues being mapped to different requirements.

Challenges of Mapping

- The difficulty is that it is often not obvious which of the pillars or requirements applies, in many cases multiple pillars or requirements can apply or a decision is made which one is the most applicable.
- The working group team found that the mapping of an issue is often debatable and strongly depends on the background of the person performing the mapping. Disagreements regarding the mappings within the groups were resolved by group consensus.
- In one use case, across the different working groups, the whole team identified a large number of issues (over 50) which need further consolidation.

Offer Recommendations

- The resolve phase completes the process by addressing ethical tensions and by giving recommendations to the key stakeholders.

Recommendations to the key stakeholders

- The output of the assessment is 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.

Monitor over time

- It is crucial to monitor that the AI system that fulfilled the trustworthy AI requirement at launch continues to do so over time.
- Therefore, when required, the resolve phase includes conducting a trustworthy monitoring over time of the AI system (we call it “ethical maintenance”).

Shortcomings

We recognize that our self assessment for this use case has some limitations, namely:

- Both the mappings and the consolidation of the mappings involve subjective decision-making components;

Shortcomings

- The mapping process relies on European guidelines. On the one hand, this clearly shapes the process, as it provides a framework for assessment.
- On the other hand, as the guidelines stress certain ethical concepts and principles in their pillars and requirements, the whole assessment tends to stress those ethical concepts and principles. This may bear the risk of disregarding other ethical concepts and principles relevant in the context of a use case;

Shortcomings

- Overall, the assessment is shaped and also limited by the team members’ focus of work and expertise. While a very large interdisciplinary team may work on a use case, it is not possible to ensure that every perspective is covered or is equally covered.

Resources

How to Assess Trustworthy AI in Practice.

Roberto V. Zicari, Julia Amann, Frédérick Bruneault, Megan Coffee, Boris Düdder, Eleanore Hickman, Alessio Gallucci, Thomas Krendl Gilbert, Thilo Hagendorff, Irmhild van Halem, Elisabeth Hildt, Georgios Kararigas, Pedro Kringen, Vince I. Madai, Emilie Wiinblad Mathez, Jesmin Jahan Tithi, Dennis Vetter, Magnus Westerlund, Renee Wurth
On behalf of the Z-Inspection® initiative (2022)

Cite as: arXiv:2206.09887 [cs.CY]
The full report is available on *arXiv* .

Link: https://arxiv.org/abs/2206.09887

Use Case

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

together with the Emergency Medical Dispatch Center (EMS) of the City of Copenhagen.

Health-related emergency calls (112)

- **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).

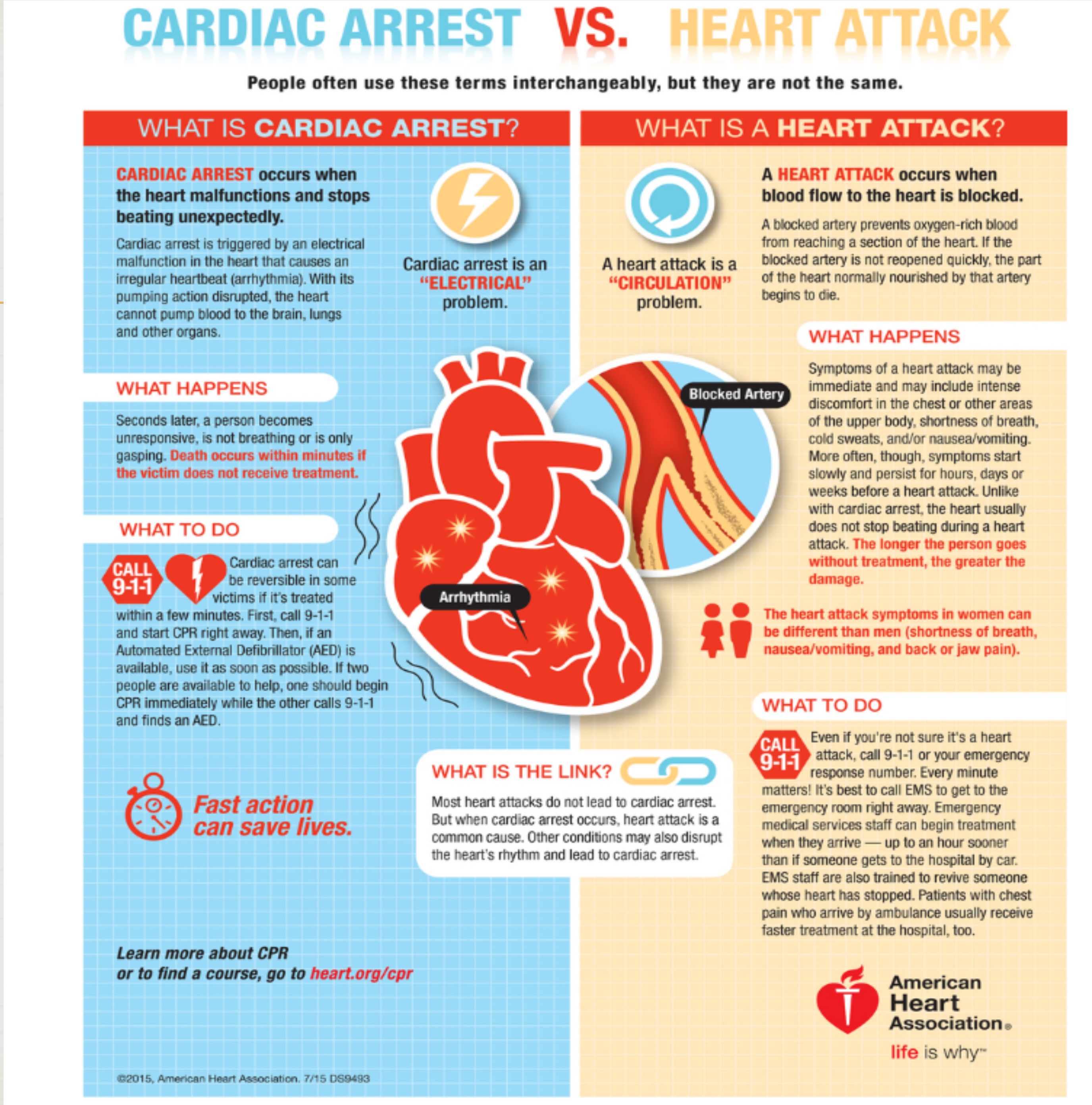
Health-related emergency calls (112)

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

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

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.

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.

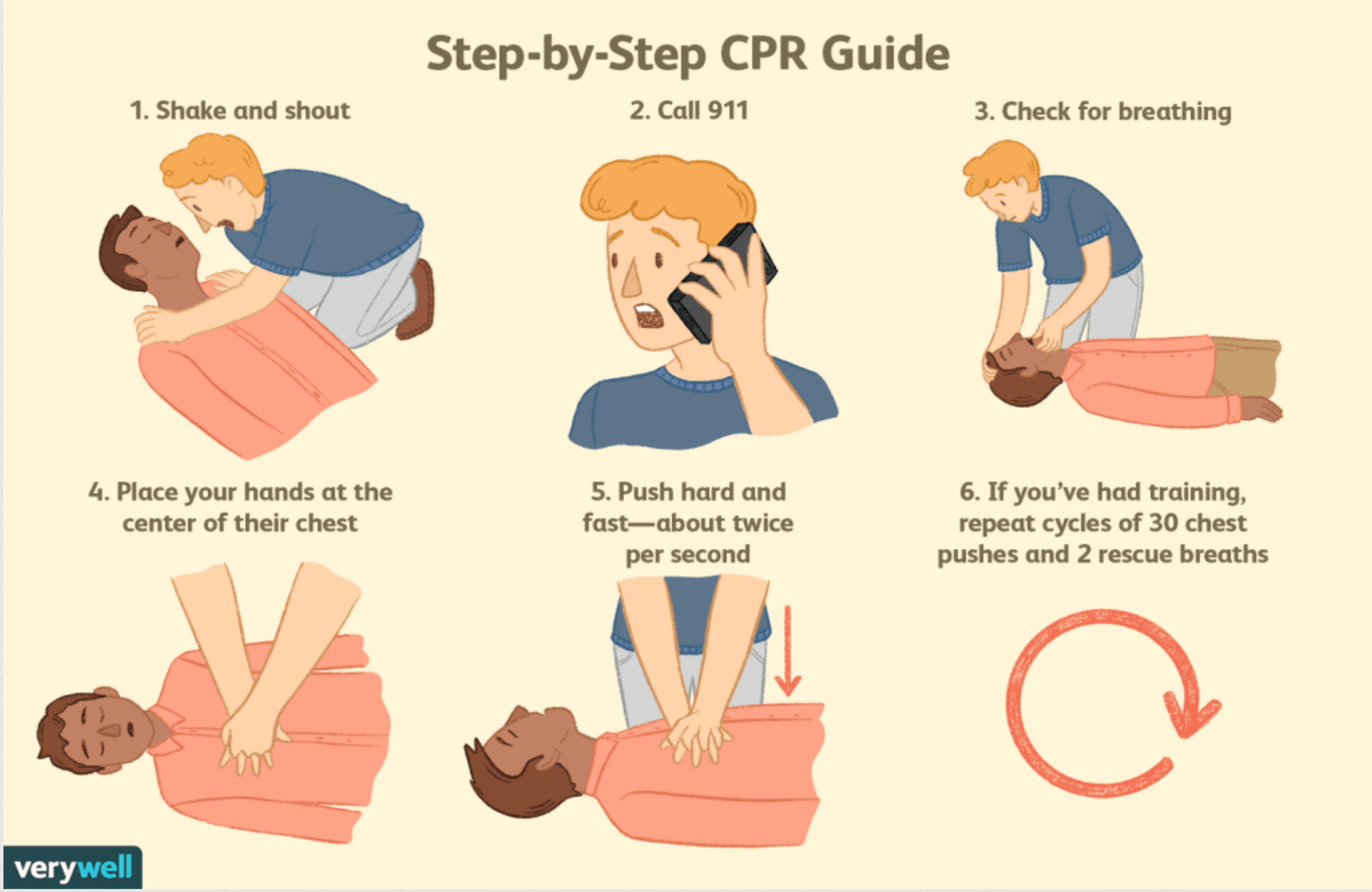
American Heart Association®

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)



Step-by-Step CPR Guide

1. Shake and shout

2. Call 911

3. Check for breathing

4. Place your hands at the center of their chest

5. Push hard and fast—about twice per second

6. If you've had training, repeat cycles of 30 chest pushes and 2 rescue breaths

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

Liability

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

The AI “solution”

- A team of medical doctors of the Emergency Medical Services Copenhagen, and the 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.

Context and processes, where the AI system is used

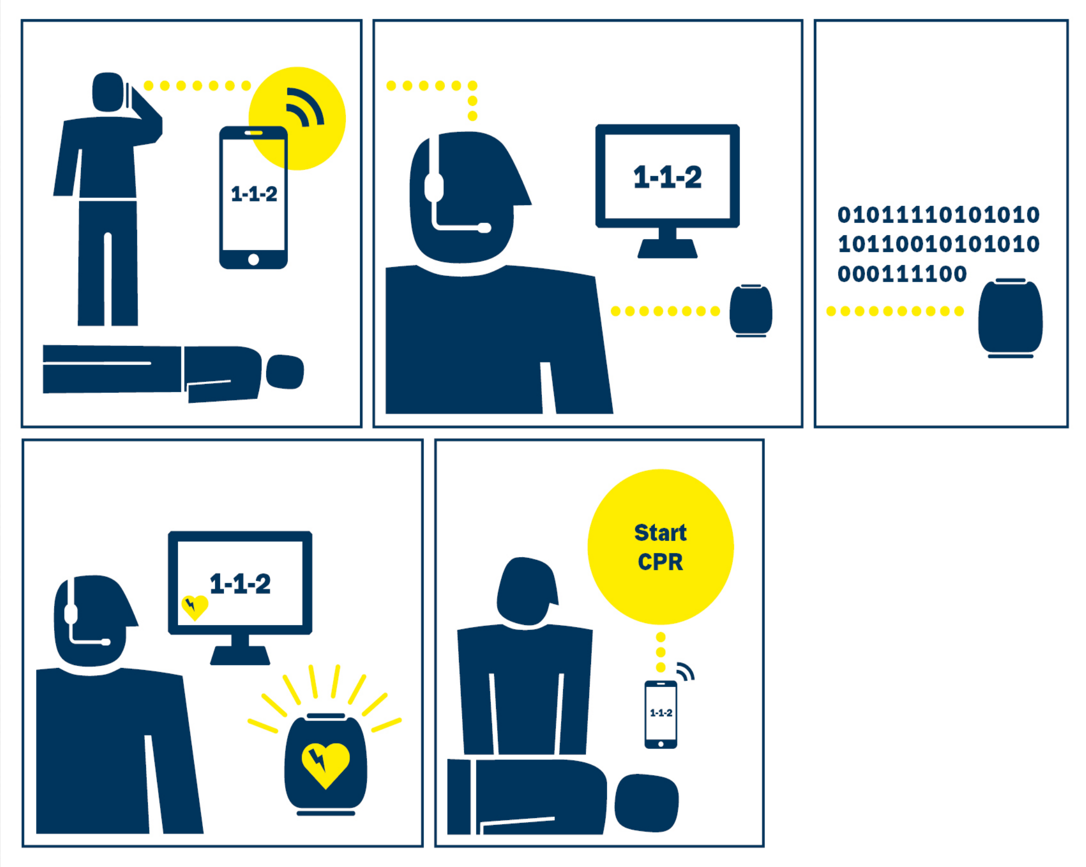


Figure . Ideal Case of Interaction between Bystander, Dispatcher, and the ML System. (with permission from Blomberg, S. N 2019b)

Retrospective study

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

Retrospective study

- The machine learning framework had a significantly **higher sensitivi**ty (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

Randomized clinical trial

- In 2020 it was conducted 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

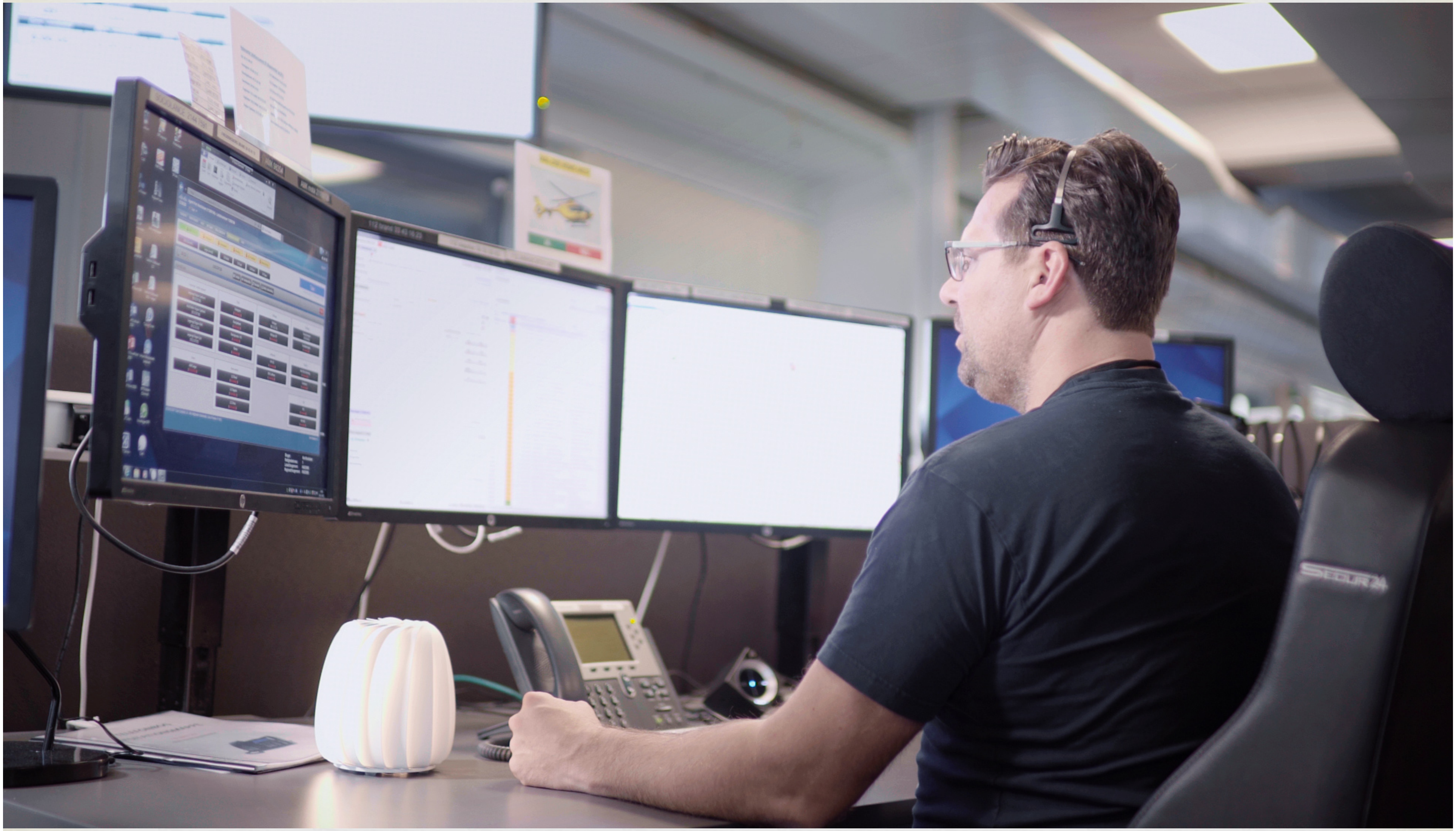
Randomized clinical trial (Cont.)

- **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.

The AI system was put 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

Motivation

- We agreed to conduct a *self-assessment* jointly by our team of independent experts together with the prime stakeholder of this use case.
- The main motivation of this work is to study if the rate of lives saved could be increased by using AI, and at the same time to identify ***how trustworthy is the use of the AI system*** assessed here, and to provide recommendations to key stakeholders.

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- ,

Tensions in the evidence base

- 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 was 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 needed 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 bene*fits*

- 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.

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, Fairness

- 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

- 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.

Example of mapping

- **Dispatcher Accept/Reject Prompt**

ID Ethical Issue: E1, Dispatcher Accept/ Reject Prompt

- **Description:**

It is unclear whether the dispatcher should be advised or controlled by the AI, and it is unclear how the ultimate decision is made.

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

- Respect for Human Autonomy > Human Agency and Oversight > Human Agency and Autonomy.

NARRATIVE RESPONSE

- Importantly, any use of an AI system in the healthcare system needs to be accompanied by a clear definition of its use. In the current setting, it is unclear how the decision support tool, should be used by the dispatchers. Should they defer to the tool's decision (especially since the performance seems to surpass human capabilities)?
- And if they do not defer to the tool, do they need to justify the decision? We also need to take into account that the dispatchers in Denmark are highly trained professionals that will not easily defer to an automated tool without a certain level of clinical validation and trust in the system. Despite the fact that the dispatchers are the primary users, they were not involved in the system design.

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 vs. 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 Tension (Open Vocabulary): ET4

- Kind of tension: True dilemma.
- Trade-off: *Fairness vs. Accuracy*.
- Description: The algorithm is accurate on average but may systematically discriminate against specific minorities of callers and/or dispatchers due to ethnic and gender bias in the training data.

Example of Recommendations

- *Recommendation 1*: **It is important to ensure that dispatchers understand the model predictions so that they can identify errors and detect biases that could discriminate against certain populations.**