

How to Assess Trustworthy AI in Practice



Roberto V. Zicari

**Graduate School of Data Science, SNU
October 12, 2022**

Z-inspection® is a registered trademark.

The content of this work is open access distributed under the
terms and conditions of the Creative Commons

(Attribution-NonCommercial-ShareAlike
CC BY-NC-SA) license

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Structure of the Talk



Part I: Quick intro to the EU Framework for Trustworthy Artificial Intelligence

Part II: Quick intro to the Z-Inspection®: A Process to Assess Trustworthy AI

Part III: Illustration of a Use Case

*Part I: Quick intro to the EU Framework for
Trustworthy Artificial Intelligence*



The EU considers the view of contemporary Western European democracy

☞ "The essence of a modern democracy is based on respect for others, expressed through support for **fundamental human rights.** "

-- **Christopher Hodges**, *Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford*

We use the EU Framework for Trustworthy Artificial Intelligence



The EU High-Level Expert Group on AI defined 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.

Four Ethical Principles



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 these principles.

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

Seven 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 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 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 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.

AI life cycle: Evaluation



Challenges and Limitations of the EU Framework for Trustworthy AI



They offer a static checklist and web tool (ALTAI) for self-assessment, but *do not validate claims, nor take into account changes of AI over time.*

The AI HLEG trustworthy AI *guidelines are not a law and are not contextualized by the domain they are involved in.* The meaning of some of the seven requirements is not anchored to the context (e.g., fairness, wellbeing, etc.).

How to assess Trustworthy AI in *practice*?



photo RVZ

Part II



Z-Inspection[®]: A process to assess
trustworthy AI



Z-inspection®



I lead a team of international experts who created an *orchestration (participatory) process* to help stakeholders to assess the *ethical, technical, domain specific and legal* implications of the use of an AI-product/services within given *contexts*.

Z-inspection® is a registered trademark.

This work is distributed under the terms and conditions of the Creative Commons (Attribution-NonCommercial-ShareAlike CC BY-NC-SA) license.

Based on our research work

On Assessing Trustworthy AI in Practice

☞ **Assessing Trustworthy AI. Best Practice:**

AI for Predicting Cardiovascular Risks (completed. Jan. 2019-August 2020)

☞ **Assessing Trustworthy AI. Best Practice:**

Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. (1st phase completed. September 2020-March 2021)

☞ **Co-design of Trustworthy AI. Best Practice:**

Deep Learning based Skin Lesion Classifiers. (1st phase completed. November 2020-March 2021)

☞ **Assessing Trustworthy AI in times of COVID-19.**

Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients.(completed April- Dec. 2021)

We use a holistic approach



☞ We use a *holistic* approach, rather than monolithic and static checklists.

We include a broader set of stakeholders



- ∞ To perform an assessment, we create a *interdisciplinary* team of experts.
- ∞ At all stages of the AI life cycle, it is important to bring together a broader set of stakeholders.

Z-inspection® can be applied to the Entire AI Life Cycle



- ❧ **Design** (*together with AI designers and Domain Experts*)
- ❧ **Development** (*together with AI designers and Domain Experts*)
- ❧ **Deployment** (*self-assessment with stakeholders owning the case or independent audit*)
- ❧ **Monitoring** (*with stakeholders owning the case or independent audit*)

We use Socio-technical Scenarios



By collecting relevant resources, the team of interdisciplinary experts create socio-technical scenarios and analyze them to describe:

**the aim of the AI systems,
the actors and their expectations and interactions,
the process where the AI systems are used,
the technology and the context (*ecosystem*).**

Resulting in a number of *issues* to be assessed.

Depending on the use case we perform a Fundamental Rights Assessment



- ❧ **The Fundamental Rights and Algorithm Impact Assessment (FRAIA)** helps to map the risks to human rights in the use of algorithms and to take measures to address this. In all stages, respect for fundamental rights must be ensured.
- ❧ The FRAIA includes a special sub-section that pays attention to **identifying risks of infringing fundamental rights and to the need to provide a justification for doing so.**
- ❧ <https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms>

We use Clusters of Fundamental rights



❧ Fundamental rights relating to **the person**

❧ **Freedom-related** fundamental rights

❧ **Equality** rights

❧ **Procedural** fundamental rights

❧ <https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms>



We develop an evidence base



This is an iterative process among experts with different skills and background with goal to:

- ❧ Understand technological capabilities and limitations
- ❧ **Build a stronger evidence base to support claims and identify tensions** (*domain specific*)
- ❧ Understand the perspective of different members of society

Claims, Arguments and Evidence (CAE)



Claims – “assertions put forward for general acceptance. They are typically statements about a property of the system or some subsystem.

Claims that are asserted as true without justification become **assumptions** and claims supporting an argument are called subclaims. “

Claims, Arguments and Evidence (CAE)



Evidence “that is used as the basis of the justification of the claim.

Sources of evidence may include the design, the development process, prior field experience, testing, source code analysis or formal analysis”, peer-reviewed journals articles, peer-reviewed clinical trials, etc.

Claims, Arguments and Evidence (CAE)



Arguments link the evidence to the claim.

They are “statements indicating the general ways of arguing being applied in a particular case and implicitly relied on and whose trustworthiness is well established”, together with the validation for the scientific and engineering laws used.

Identify Tensions and Trade-offs



- ✧ **Tensions may arise between ethical principles, for which there is no fixed solution.**
- ✧ “In line with the EU fundamental commitment to democratic engagement, due process and open political participation, *methods of accountable deliberation to deal with such tensions should be established.*”

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

We use a 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)
- ❧ Source: *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.

Ethical Tensions



- ☞ “We use the umbrella term ‘tension’ to refer to different ways in which values can be in conflict, some more fundamentally than others.”
- ☞ “When we talk about tensions between values, we mean tensions between the pursuit of **different values in technological applications** rather than an abstract tension between the values themselves.”

Source:[1] *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.

Catalogue of Examples of Tensions



From [1]:

- ❧ *Accuracy vs. Fairness*
- ❧ *Accuracy vs. Explainability*
- ❧ *Privacy vs. Transparency*
- ❧ *Quality of services vs. Privacy*
- ❧ *Personalisation vs. Solidarity*
- ❧ *Convenience vs. Dignity*
- ❧ *Efficiency vs. Safety and Sustainability*
- ❧ *Satisfaction of Preferences vs. Equality*

[1] 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.

We use a consensus process based on *mappings*. *Open to close vocabulary*



We map *issues* freely described (*open vocabulary*) by the interdisciplinary team of experts) to some of the 4 ethical principles and 7 requirements for Trustworthy AI (*closed vocabulary*)

We rank *mapped issues* by relevance depending on the context. (e.g. Transparency, Fairness, Accountability)



We give Recommendations



☞ **Appropriate use;**

☞ **Remedies:** If risks are identified, we recommend ways to mitigate them (when possible);

☞ **Ability to redress.**

Part III



Use Case



Self Assessment

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

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

The Team



✧ [Roberto V. Zicari](#)^{1,2*}, James Brusseau³, Stig Nikolaj Blomberg⁴, Helle Collatz Christensen⁴, Megan Coffee⁵, Marianna B. Ganapini⁶, Sara Gerke⁷, Thomas Krendl Gilbert⁸, Eleanore Hickman⁹, Elisabeth Hildt¹⁰, Sune Holm¹¹, Ulrich Kühne¹², Vince I. Madai^{13,14,15}, Walter Osika¹⁶, Andy Spezzatti¹⁷, Eberhard Schnebel¹⁸, Jesmin Jahan Tithi¹⁹, Dennis Vetter¹⁸, Magnus Westerlund¹, Renee Wurth²⁰, Julia Amann²¹, Vegard Antun²², Valentina Beretta²³, Frédéric Bruneault²⁴, Erik Campano²⁵, Boris Düdler²⁶, Alessio Gallucci²⁷, Emmanuel Goffi²⁸, Christoffer Bjerre Haase²⁹, Thilo Hagendorff³⁰, Pedro Kringen¹⁸, Florian Möslein³¹, Davi Ottenheimer³², Matiss Ozols³³, Laura Palazzani³⁴, Martin Petrin^{35,36}, Karin Tafur³⁷, Jim Tørresen³⁸, Holger Volland³⁹ and Georgios Kararigas⁴⁰

Socio-technical Scenarios: The Context



☞ **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/>

Socio-technical Scenarios: *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 .

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

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™

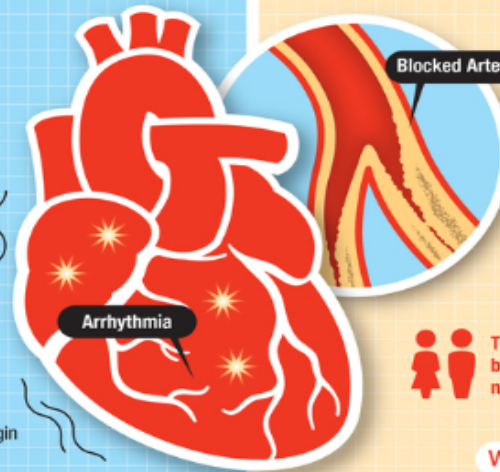


Image:
CPR

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.

The Problem (cont.)



- ∞ 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



verywell

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 if 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.

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 was to identify *how trustworthy was the use of the AI system* assessed, and to provide recommendations to key stakeholders.

Context and processes, where the AI system is used

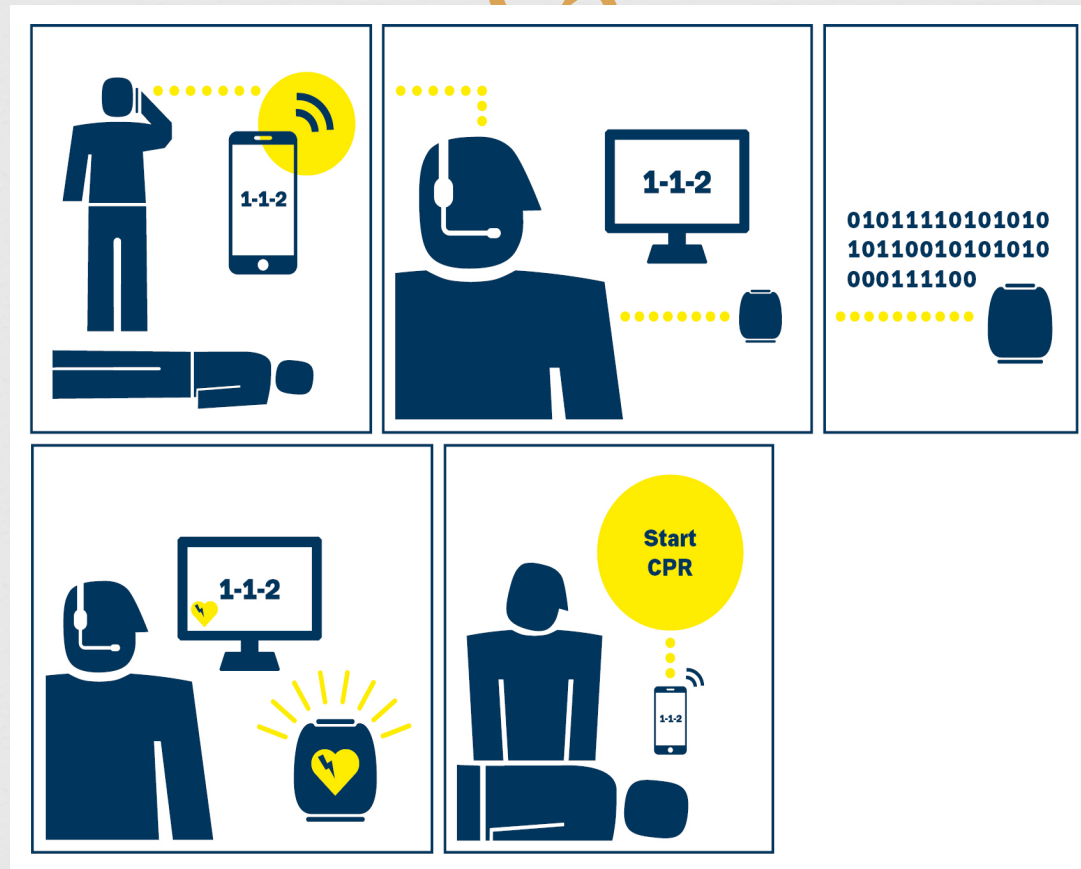


Figure. Ideal Case of Interaction between Bystander, Dispatcher, and the ML System. (with permission from Blomberg, S. N 2019b)

Develop an Evidence: 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 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, $p < 0.001$).

*Develop an Evidence:
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.)



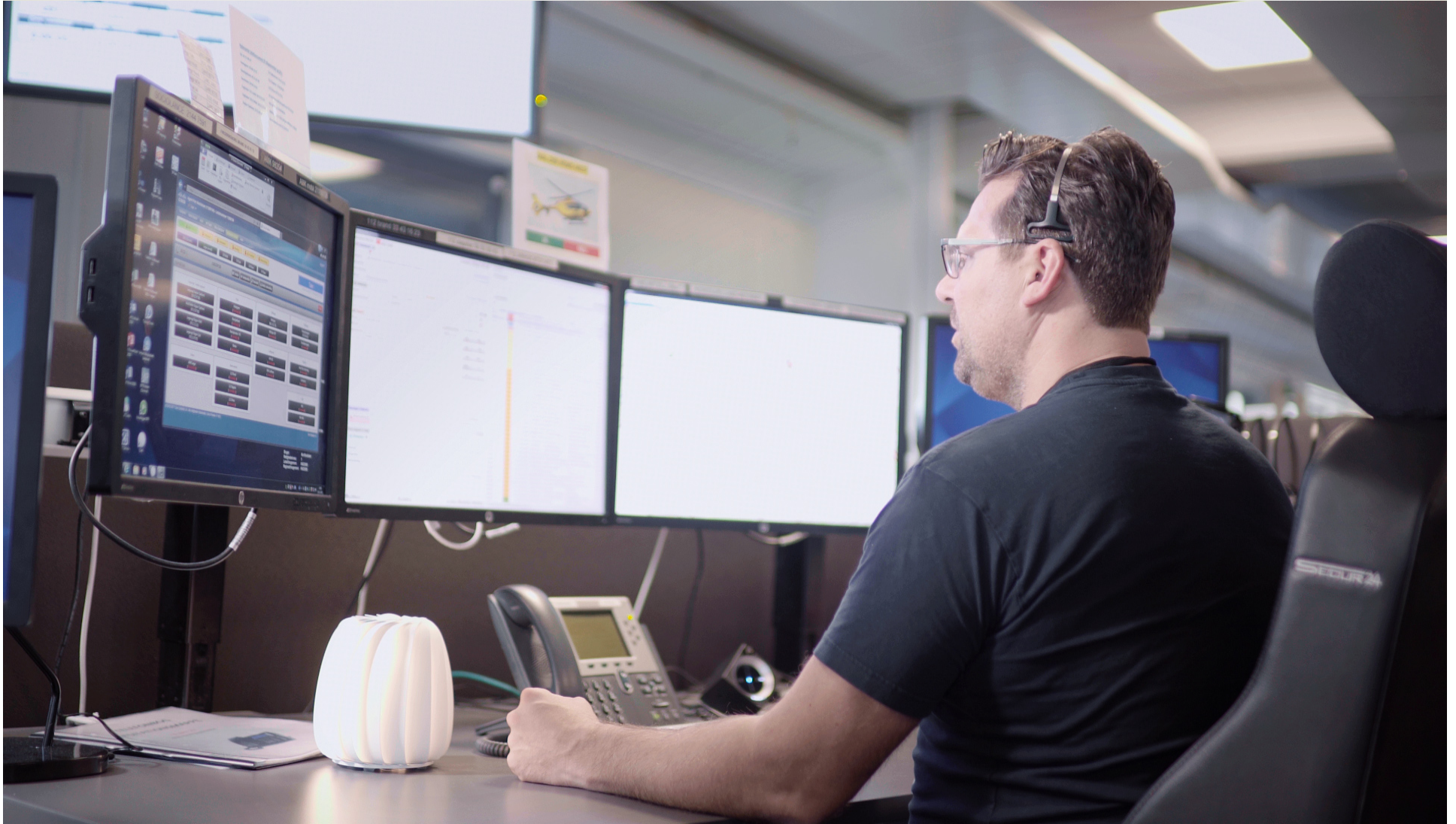
✧ **There was no significant improvement in recognition of out-of-hospital cardiac arrest during calls on which the model alerted dispatchers vs those on which it did not; however, the machine learning model had higher sensitivity than dispatchers alone.**


The AI system was put in production



☞ The AI system was put into production during Fall 2020.

☞ Note: The responsible person at the Emergency Medical Dispatch Center **authorized the use** of the AI system.



 <https://cordis.europa.eu/project/id/823383/reporting>

Tensions in the evidence base



- ✧ We discovered 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).

Possible 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.

Possible 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?

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.**

AI Design decision



- ❧ 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.
- ❧ (*) **this was a design decision** made by the medical doctors and implemented by the AI engineers.

Consider the Design Decision as a claim



- ❧ 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.
- ❧ The design argument was that 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 transparent** 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



- ❧ We looked 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.

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.

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.

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.**
- ❧ Here, the model is a statistical black-box, and the clinical trial conducted with the model showed an important lack of trust that had an impact on the outcome of the trial. An improvement to the model would include interpretable local approximations [such as SHAP ([Lundberg and Lee, 2017](#))], which are easy for stakeholders to understand and provide different levels of interpretation for judging the relevance of an individual prediction. In our example, explanation may involve words that were more predictive, tone of voice, or breath sounds.



Recommendations



- ❧ *Recommendation 2:* We believe that the team should either intentionally sample the entire training set in order to prevent discrimination, or define a heuristic that could inform dispatchers when to use and when not to use the model. ..
- ❧ *Recommendation 3:* Involve stakeholders. The group of (potential future) patients and (potential future) callers could be interested in how the system functions and is developed. User involvement/stakeholder involvement could be very helpful in the process of re-designing the AI system. ..
- ❧ *Recommendation 4:* It is important to learn how the protocol (what questions, how many, etc.) does or does not influence the accuracy of the ML output. Further research work should be performed to answer this question....
- ❧ *Recommendation 5:* Although we did not assess the legal aspects of the AI system, we suggest to the prime stakeholder to verify with legal local competent authorities if the AI system needed a CE-certification as a medical device...

Reference



On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls

Roberto V. Zicari • James Brusseau • Stig Nikolaj Blomberg • Helle Collatz Christensen • Megan Coffee • Marianna B. Ganapini • Sara Gerke • Thomas Krendl Gilbert • Eleanore Hickman • Elisabeth Hildt • Sune Holm • Ulrich Kühne • Vince I. Madai • Walter Osika • Andy Spezzatti • Eberhard Schnebel • Jesmin Jahan Tithi • Dennis Vetter • Magnus Westerlund • Renee Wurth • Julia Amann • Vegard Antun • Valentina Beretta • Frédérick Bruneault • Erik Campano • Boris Düdder • Alessio Gallucci • Emmanuel Goffi • Christoffer Bjerre Haase • Thilo Hagendorff • Pedro Kringen • Florian Möslin • Davi Ottenheimer • Matiss Ozols • Laura Palazzani • Martin Petrin • Karin Tafur • Jim Tørresen • Holger Volland • Georgios Kararigas

Front. Hum. Dyn., 08 July 2021 |

Current Pilot Project



❧ Pilot Project: Assessment for Responsible Artificial Intelligence together with Rijks ICT Gilde -Ministry of the Interior and Kingdom Relations (BZK)- and the province of Fryslân (The Netherlands)

Background for the pilot

❧ The practical application of a deep learning algorithm from the province of Fryslân will be investigated and assessed. The algorithm maps heathland grassland by means of satellite images for monitoring nature reserves. The testing of this algorithm is done in collaboration with an international interdisciplinary team, based on the 'Z-Inspection® method' - a process to test AI for reliability.