What did we learn in assessing Trustworthy AI in practice?

Roberto V. Zicari
Z-Inspection® Initiative
http://z-inspection.org

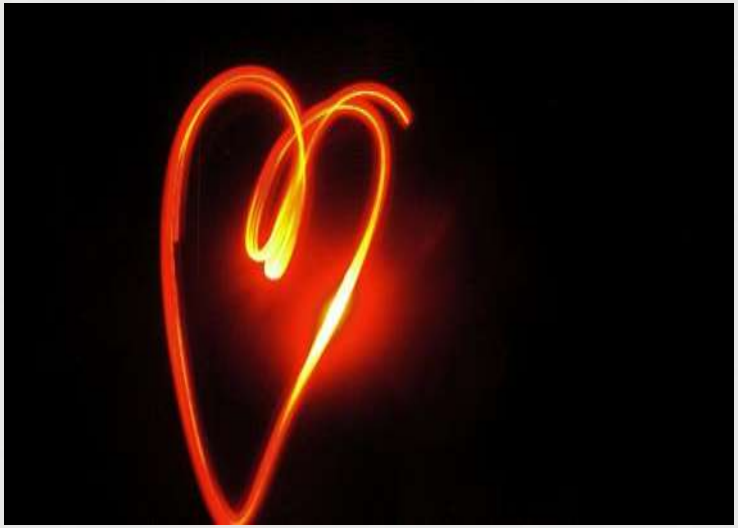
AI Ethics online--Chalmers
April 20, 2021

Artificial Intelligence (AI)

"Everything we love about civilization is a product of intelligence, so amplifying our human intelligence with artificial intelligence has the potential of helping civilization flourish like never before – as long as we manage to keep the technology ***beneficial.***"

Max Tegmark, President of the Future of Life Institute

How do we "know" when an AI system is "***beneficial" or not?***



- Our approach is inspired by both theory and practices (" learning by doing").

photo CZ

Best Practices http://z-inspection.org/best-practices/

- Assessing Trustworthy AI. Best Practice: **AI for Predicting Cardiovascular Risk**s (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. Best Practice: **Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients**.(Start April15, 2021)

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 112 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/

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 .

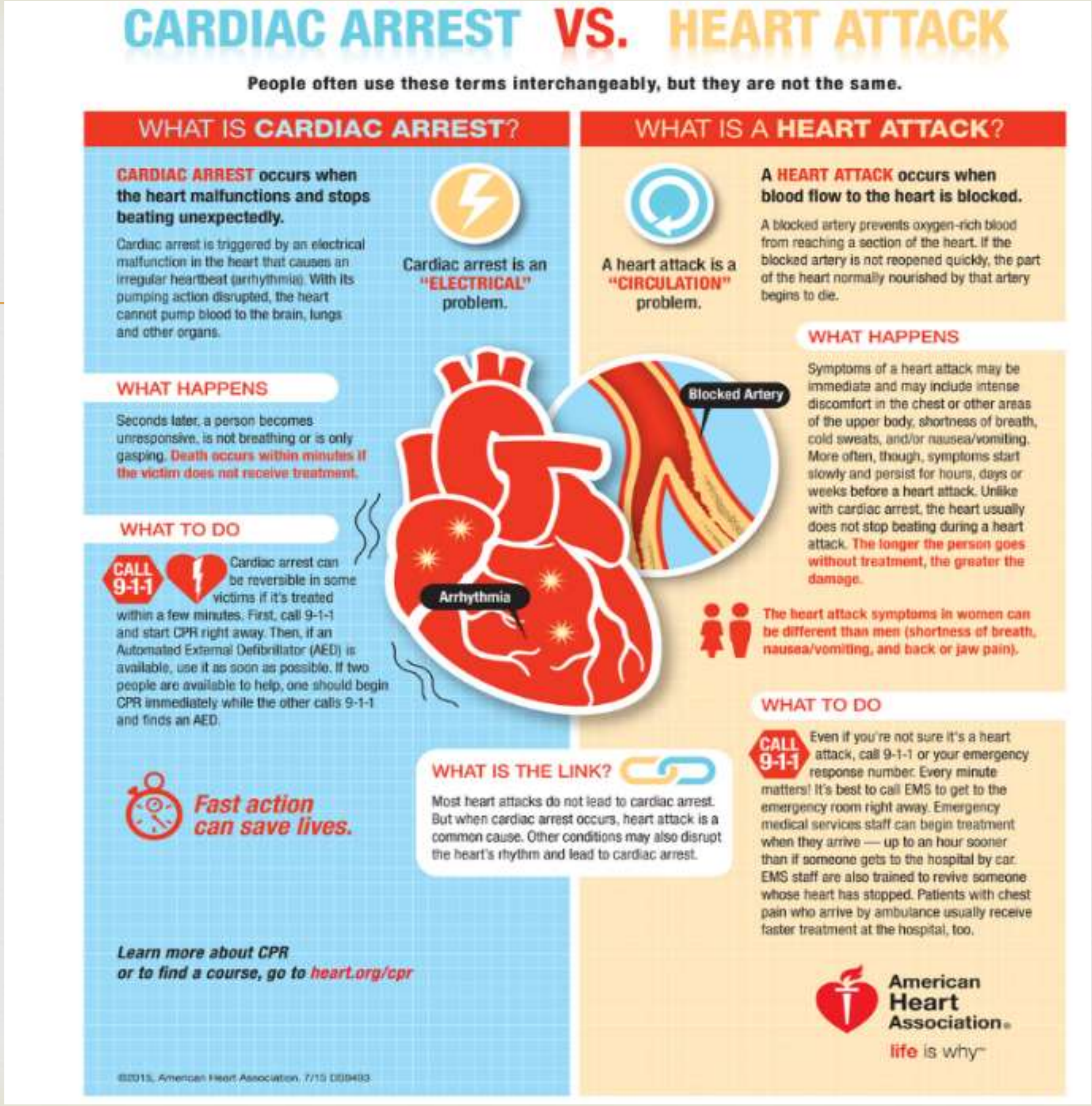

CARDIAC ARREST VS. HEART ATTACK
People often use these terms interchangeably, but they are not the same.
WHAT IS CARDIAC ARREST?
CARDIAC ARREST occurs when the heart malfunctions and stops beating unexpectedly.
Cardiac arrest is triggered by an electrical malfunction in the heart that causes an irregular 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
Arrhythmia
Blocked Artery
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

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.**
- 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

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

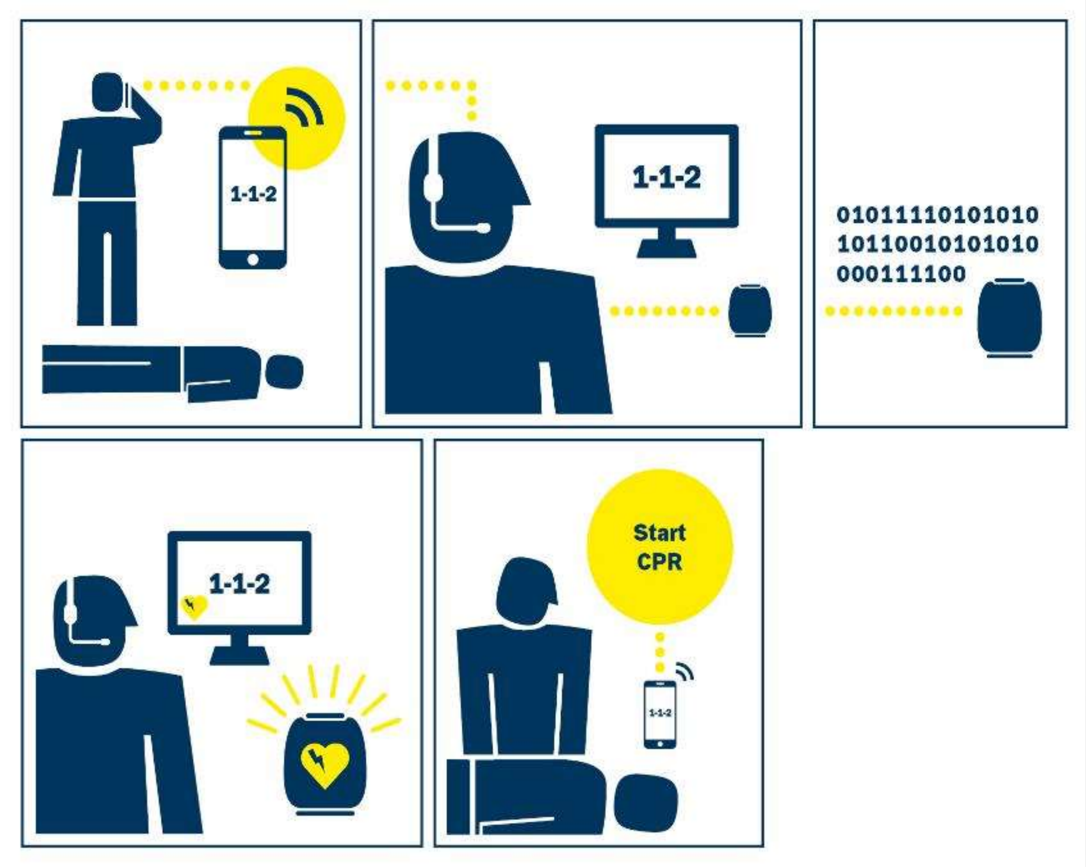


Figure . Ideal Case of Interaction between Bystander, Dispatcher, and the ML System. (with permission from Blomberg, S. N

Retrospective study

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

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 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.

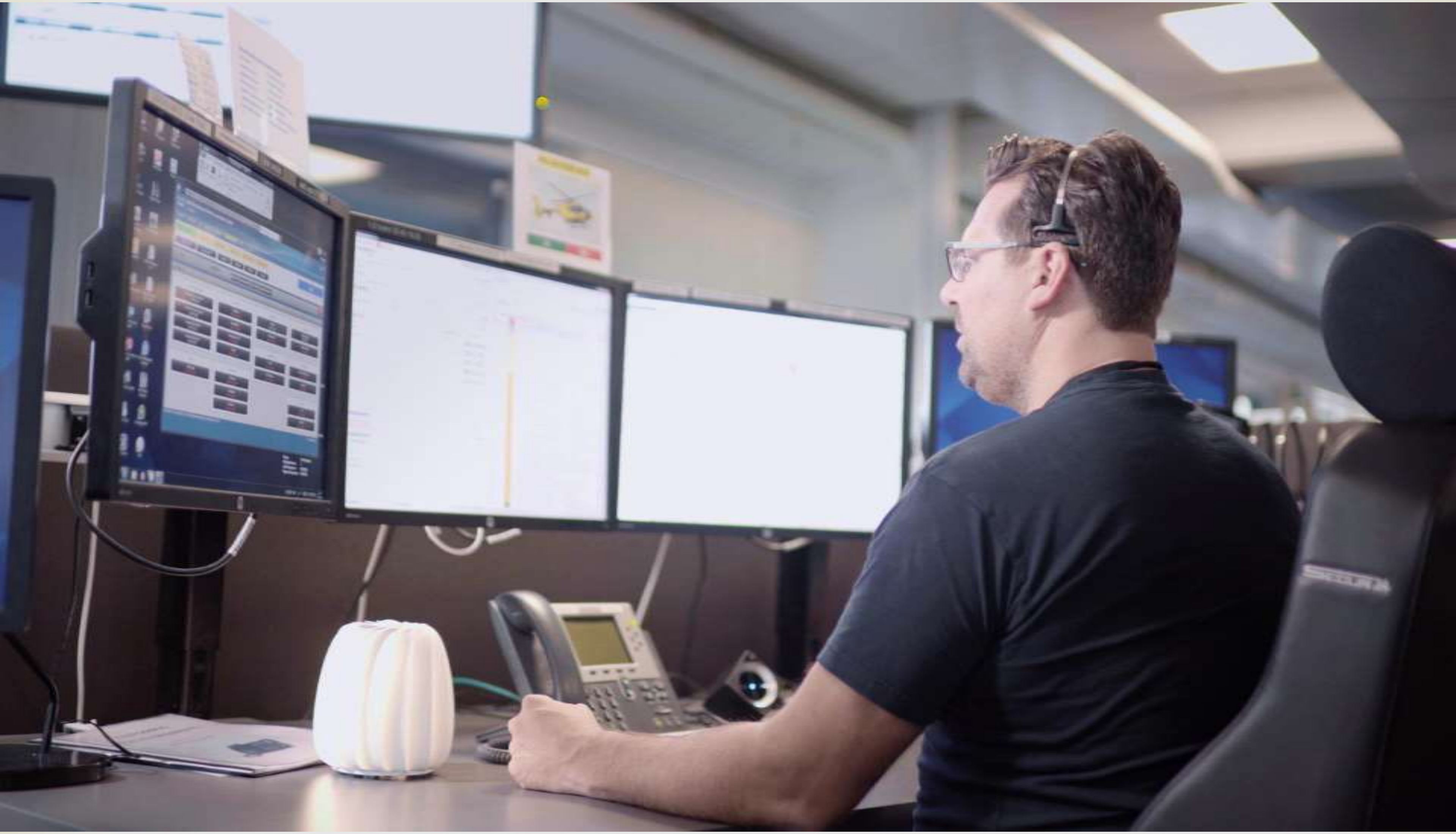
Status

- Although **the AI system performed well in a retrospective study**,
- **In randomized clinical trial it did not contribute to an improvement for detection of OHCA**

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

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.**

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.

Identification of possible ethical, legal and technical “issues”

- Using the Z-inspection® process we identified possible **ethical and technical and legal issues** for the use of the AI within the given boundaries and context.

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

Trustworthy AI Framework

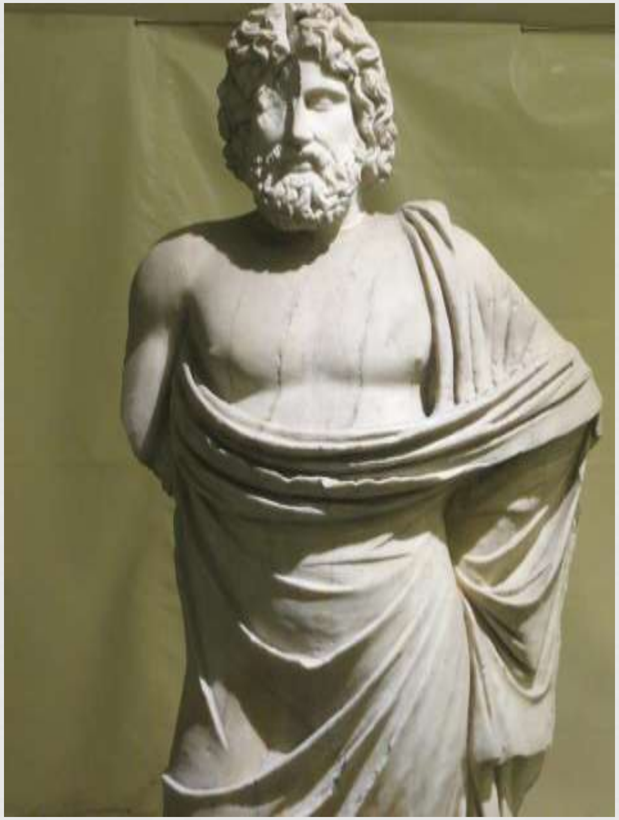


Photo RVZ

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.

The principle of explicability (cont.)

- An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible.
- These cases are referred to as **'black box' algorithms** and **require special attention**.
- In those circumstances, **other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental right**s.
- The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.

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

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

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.

Requirements of 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

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.

Limitations

Although these requirements are a welcome first step towards enabling an assessment of the societal implication of the use of AI systems, there are some challenges in the practical application of requirements, namely:

- **The AI guidelines are not domain specific** and the meaning of some of the seven requirements is not anchored to the context (e.g. fairness, wellbeing etc.)
- **They mainly offer a static checklist** and do not distinguish different applicability of the AI guidelines (e.g. during design vs. after production) as well as different stages of algorithmic development, starting from business and use-case development, design phase, training data procurement, building, testing, deployment, and monitoring.
- **There are not available best practices** to show how to implement such requirements an be applied in practice.

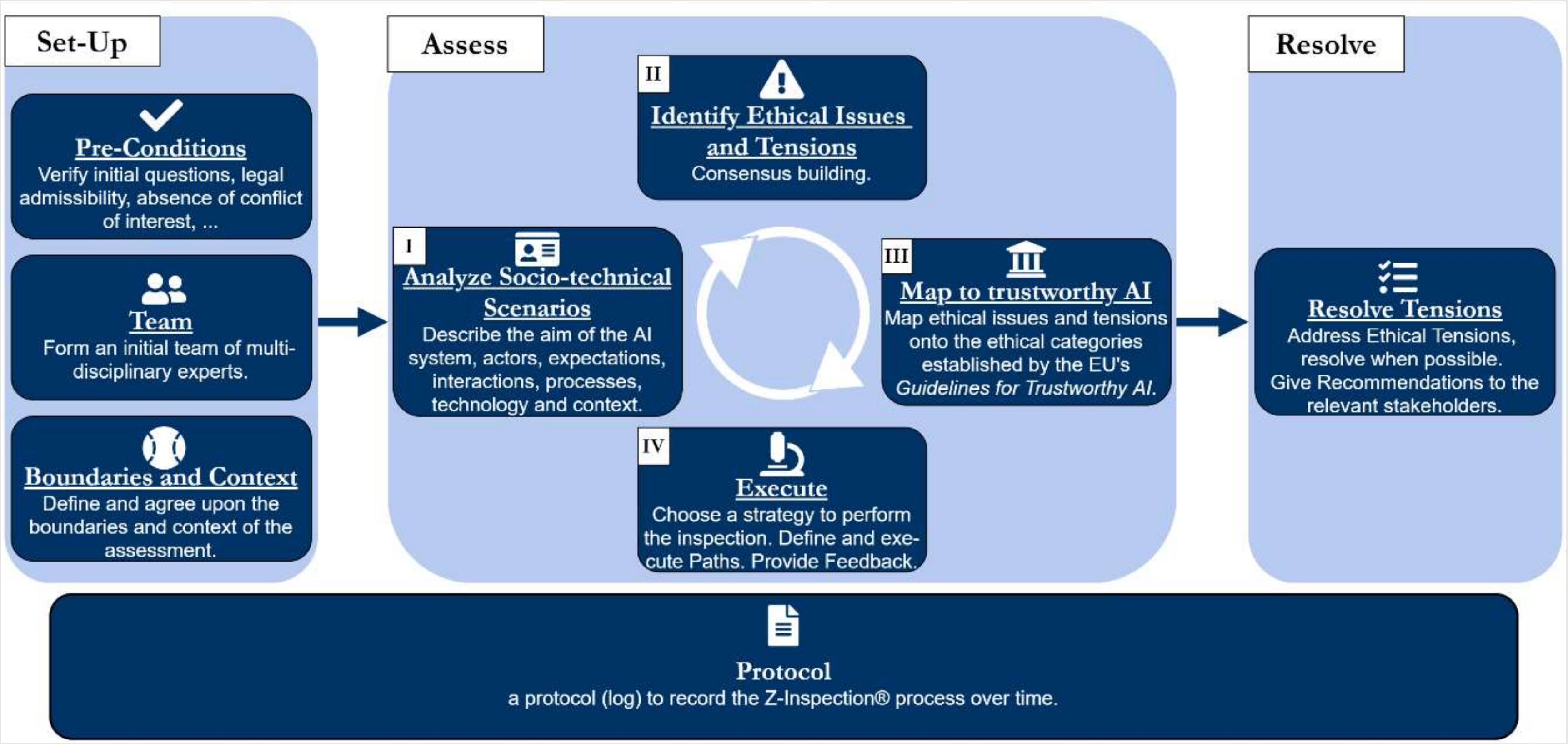
Z-inspection® : A Process to Assess Trustworthy AI.

Roberto V. Zicari *(1)*, John Brodersen *(4)(9)*, James Brusseau *(8)*, Boris Düdder (*6*), Timo Eichhorn *(1)*, Todor Ivanov *(1)*, Georgios Kararigas *(3)*, Pedro Kringen *(1)*, Melissa McCullough *(1)*, Florian Möslein *(7)*, Naveed Mushtaq *(1)*, Gemma Roig *(1)*, Norman Stürtz *(1)*, Karsten Tolle *(1)*, Jesmin Jahan Tithi (*2*), Irmhild van Halem *(1)*, Magnus Westerlund (*5*).

(1) Frankfurt Big Data Lab, Goethe University Frankfurt, Germany; (2) Intel Labs, Santa Clara, CA, USA; (3) German Centre for Cardiovascular Research, Charité University Hospital, Berlin, Germany; (4) Section of General Practice and Research Unit for General Practice, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Danemark; (5)Arcada University of Applied Sciences, Helsinki, Finland; (6) Department of Computer Science (DIKU), University of Copenhagen (UCPH), Denmark; (7) Institute of the Law and Regulation of Digitalization, Philipps-University Marburg, Germany; (8) Philosophy Department, Pace University, New York, USA; (9) Primary Health Care Research Unit, Region Zealand, Denmark

IEEE Transactions on Technology and Society. *2021*

Z-inspection® Process in a Nutshell



Focus of Z-inspection®

Z-inspection® covers the following:

- Ethical and Societal implications;
- Technical robustness;
- Legal/Contractual implications.

Note1: *Illegal and unethical are not the same thing*.
Note2: *Legal and Ethics depend on the context*
Note 3*:* Relevant/accepted for the ecosystem(s) of the AI use case.

“Embedded” Ethics into AI.

- When designing, training and testing an AI-system (e.g. Machine-Learning algorithm) we do “embed” into the system notions such as “good”, “bad”, “healthy”, “disease”, etc. mostly not in an explicit way.

“Embedded” Ethics into AI: Medical Diagnosis

"In case medical diagnosis or treatment recommendations are being deferred to machine learning algorithms, it is the algorithm who sets the bar about how a disease is being defined.”

-- Thomas Grote , Philipp Berens

Source: Grote T, Berens P.
J Med Ethics Epub ahead of print: [please include Day Month Year]. doi:10.1136/ medethics-2019-105586

A holistic approach

- We use a holistic approach, rather than monolithic and static ethical checklists.

Orchestration Process

- The core idea of our assessment is to create an *orchestration process* to help teams of skilled experts to assess the *ethical, technical* and *legal* implications of the use of an AI-product/services within given *contexts*.

- Wherever possible Z-inspection® allows us to use existing frameworks, check lists, "plug in" existing tools to perform specific parts of the verification. The goal is to customize the assessment process for AIs deployed in different domains and in different contexts.

Set Up

Who? Why? For Whom?

We defined a catalogue of questions to help clarify the expectation between stakeholders, before the Z-Inspection assessment process starts:

- *Who* requested the inspection?
- *Why carry* out an inspection?
- For *whom* is the inspection relevant?
- Is it *recommended or required* (mandatory inspection)?
- What are the *sufficient vs. necessary* conditions that need to be analysed?
- How to *use the results* of the Inspection? There are different, possible uses of the results of the inspection: e.g. verification, certification, and sanctions (if illegal).

What to do with the assessment?

Lesson Learned

- A further important issue to clarify upfront is if the results will be shared (public), or kept private.

- In the latter case, the key question is: why keeping it private? This issue is also related to the definition of IP as it will be discussed later.

No conflict of interests: Go, NoGo

1. Ensure *no conflict of interests* exist between the inspectors and the entity/organization to be examined
2. Ensure *no conflict of interests* exist between the inspectors and vendors of tools and/toolkits/frameworks/platforms to be used in the inspection.
3. Assess *potential bias* of the team of inspectors.

- → GO if all three above are satisfied
- → Still GO with restricted use of specific tools, if 2 is not satisfied.
- → NoGO if 1 or 3 are not satisfied

Responsible use of AI

- The responsible use of AI (processes and procedures, protocols and mechanisms and institutions to achieve it) **inherit properties from the wider political and institutional contexts**.

AI, Context, Trust, Ethics, Democracy

- From a Western perspective, the terms context, trust and ethics are closely related to our concept of democracy.

There is a "*Need of examination of the extent to which the function of the system can affect the function of democracy, fundamental rights, secondary law or the basic rules of the rule of law*".

-- German Data Ethics Commission (DEK)

What if the Ecosystems are not Democratic?

If we assume that the definition of the boundaries of ecosystems is part of our inspection process, then a key question that needs to be answered before starting any assessment is the following:

What if the Ecosystems are not Democratic?

Political and institutional contexts

- We recommend that the decision-making process as to whether and where AI-based products/ services should be used must include, as an integral part, the political assessment of the "democracy" of the ecosystems that define the context.

We understand that this could be a debatable point.

What if the AI consolidates the concentration of power?

"The development of the data economy is accompanied by economic concentration tendencies that allow the emergence of new power imbalances to be observed.

Efforts to secure digital sovereignty in the long term are therefore not only a requirement of political foresight, but also an expression of ethical responsibility."

-- German Data Ethics Commission (DEK)

Should this be part of the assessment?

We think the answer is yes.

How to handle IP

- Clarify *what is* and *how to handle* the *IP* of the AI and of the part of the entity/company to be examined.

- Identify possible restrictions to the Inspection process, in this case assess the consequences (if any)

- Define if and when *Code Reviews* is needed/possible.
  For example, check the following preconditions (*):
  - There are no risks to the security of the system
  - Privacy of underlying data is ensured
  - No undermining of intellectual property

  Define the implications if any of the above conditions are not satisfied.

(*) Source: *"Engaging Policy Shareholders on issue in AI governance" (Google)*

Implication of IP on the Investigation

Lesson Learned

- There is an inevitable trade off to be made between disclosing all activities of the inspection vs. delaying them to a later stage or not disclosing them at all.

Build a Team

A team of multi-disciplinary experts is formed. The composition of the team is a dynamic process. Experts with different skills and background can be added at any time of the process.

Lesson Learned

The choice of experts have an ethical implication!

Create a Log

- A protocol (log) of the process is created that contains over time several information, e.g. information on the teams of experts, the actions performed as part of each investigation, the steps done in data preparation and analyses and the steps to perform use case evaluation with tools.
- *The protocol can be shared to relevant stakeholders at any time to ensure transparency of the process and the possibility to re-do actions;*

Define the Boundaries and Context of the inspection

- In our assessment the concept of ***ecosystems*** plays an important role, they define the boundaries of the assessment.

- Our definition of ecosystem generalizes the notion of "*sectors and parts of society, level of social organization, and publics*" defined in [1], by adding the political and economic dimensions.

[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.

AI and the Context

It is important to clarify what we wish to investigate. The following aspects need to be taken into consideration:

- AI is not a single element;
- AI is not in isolation;
- AI is dependent on the domain where it is deployed;
- AI is part of one or more (digital) ecosystems;
- AI is part of Processes, Products, Services, etc.;
- AI is related to People, Data.

Assess

Socio-technical Scenarios

- Socio-technical scenarios are created (or given to) by the team of experts to represent possible scenarios of use of the AI. This is a process per se, that involves several iterations among the experts, including using *Concept Building*.

Socio-technical Scenarios

By collecting relevant resources, socio-technical scenarios are created and analyzed by the team of experts:

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**.

Lessons Learned

- **Use an Open Vocabulary** to analyze socio-technical scenarios by the team of experts
- A **Consolidation process must be in place**

Develop an evidence base

This is an iterative process among experts with different skills and background.

- Understand technological capabilities and limitations
- Build a stronger evidence base on the current uses and impacts (*domain specific*)
- Understand the perspective of different members of society

Source: Whittlestone, J et al (2019)

Lessons Learned

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

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

On Developing an evidence base

Our experience in practice (e.g. domain healthcare/cardiology) suggests that this is a non obvious process.

Identify Tensions: For the same domain, there may be different point of views among "experts" of what is evidence; different view points of what constitutes a "neutral" and "not biased" evidence; and "who" is qualified to produce such evidence without being personally "biased".

Identification of Ethical issues and tensions.

- An appropriate *consolidated building* process is chosen that involves several iterations among the experts of different disciplines and backgrounds and result in identifying ethical issues and ethical tensions.

Tensions and Trade-offs

- 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.**

Identify Ethical Issues and Tensions, and Flags

- As a result of the analysis of the scenarios, **Ethical issues** and **Flags** are identified .

- An Ethical issue or tension refers to different ways in which values can be in conflict.

- A **Flag** is an issue that needs to be assessed further.
  (it could be a technical, legal, ethical issue)

Describe Ethical issues and Tensions

- *Confirm*, *describe* and *classify* if such Ethical Issues represent ethical tensions and if yes, describe them.
- This is done by a selected number of members of the inspection team, who are experts on ethics and/or the specific domain.
- Goal is to reach a "consensus" among the experts (when possible) and agree on a common definition of Ethical tensions to be further investigated in the Z-Inspection process.

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)

EXAMPLES OF TENSIONS BETWEEN VALUES

- **Quality of services** ***versus*** **privacy**: using personal data may improve public services by tailoring them based on personal characteristics or demographics, but compromise personal privacy because of high data demands.
- **Personalisation** ***versus*** **solidarity**: increasing personalisation of services and information may bring economic and individual benefits, but risks creating or furthering divisions and undermining community solidarity.
- Source: *Sample Catalog of Ethical Tensions* (Whittlestone et al., 2019)

EXAMPLES OF TENSIONS BETWEEN VALUES

- **Convenience *versus* dignity**: increasing automation and quantification could make lives more convenient, but risks undermining those unquantifiable values and skills that constitute human dignity and individuality.
- **Privacy *versus* transparency**: the need to respect privacy or intellectual property may make it difficult to provide fully satisfying information about an algorithm or the data on which it was trained.

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

EXAMPLES OF TENSIONS BETWEEN VALUES

- **Accuracy *versus* explainability**: the most accurate algorithms may be based on complex methods (such as deep learning), the internal logic of which its developers or users do not fully understand.
- **Accuracy *versus* fairn**ess: an algorithm which is most accurate on average may systematically discriminate against a specific minority.

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

EXAMPLES OF TENSIONS BETWEEN VALUES

- **Satisfaction of preferences *versus* equality:** automation and AI could invigorate industries and spearhead new technologies, but also exacerbate exclusion and poverty.
- **Efficiency *versus* safety and sustainability**: pursuing technological progress as quickly as possible may not leave enough time to ensure that developments are safe, robust and reliable.

- 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".

Lessons Learned

- **Use a Close Vocabulary to arrive to a consolidated list of “issues”.**

Mapping to Trustworthy AI.

- Once the ethical issues and tensions have been agreed upon among the experts, the consensus building process among experts continue by asking them to map ethical issues and tensions onto
- **the four ethical categories**, and
- **the seven requirements established by the EU High Level Experts Guidelines 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

Do a Pre-Check

- At this point in some cases, it is already possible to come up with an initial ethical pre-assessment that considers the level of abstraction of the domain, with no need to go deeper into technical levels (i.e. considering the AI as a black box).
- This is a kind of pre-check, and depends on the domain.

Back to Our Use Case

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

- 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*

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.

*Example: **Verify "fairness"***

Step 1. **Clarifying what kind of algorithmic "fairness" is most important for the domain** *(*)*

Step 2. **Identify Gaps/Mapping conceptual concepts between**:

a. *Context-relevant Ethical values,*

b. Domain-specific metrics,

c. Machine Learning fairness metrics.

(*) Source: Whittlestone, J et al (2019) *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* London: Nuffield Foundation.

Choosing Fairness criteria
(domain specific)

For *healthcare,* one possible approach is to use ***Distributive justice*** (from philosophy and social sciences) options for machine learning (*)

Define *Fairness* criteria, e.g.

Equal Outcomes
Equal Performance
Equal Allocation

(*) Source. Alvin Rajkomar et al. Ensuring, Fairness in Machine Learning to Advance Health, Equity, Annals of Internal Medicine (2018). DOI: 10.7326/M18-1990
Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594166/

Fairness criteria and Machine Learning

- ***Equal patient outcomes*** refers to the assurance that protected groups have equal benefit in terms of patient outcomes from the deployment of machine-learning models
- ***Equal performance*** refers to the assurance that a model is equally accurate for patients in the protected and non protected groups.
- ***Equal allocation*** (also known as demographic parity), ensures that the resources are proportionately allocated to patients in the protected group.

To verify these *Fairness* criteria we need to have access to the Machine Learning Model.

From Domain Specific to ML metrics

Several Approaches in Machine Learning:

Individual fairness , Group fairness, Calibration, Multiple sensitive attributes, Casuality.

In Models : Adversarial training, constrained optimization. regularization techniques,….

(*) Source *Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements*
Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, Ed H. Chi (Submitted on 14 Jan 2019)

Mapping Domain specific "Fairness" to Machine Learning metrics

- **Resulting Metrics** **Formal "non-discrimination" criteria**
  - Statistical parity Independence
  - Demographic parity (DemParity) Independence
  (average prediction for each group should be equal)
  - Equal coverage Separation
  - No loss benefits
  - Accurate coverage
  - No worse off
  - Equal of opportunity (EqOpt) Separation
  (comparing the false positive rate from each group)
  - Equality of odds Separation
  (comparing the false negative rate from each group)
  - Minimum accuracy
  - Conditional equality, Sufficiency
  - Maximum utility (MaxUtil)

(*) Source *Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements*
Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, Ed H. Chi (Submitted on 14 Jan 2019)

Trust in Machine Learning "Fairness" metrics

Some of the ML metrics depend on the training labels (*):

- When is the *training data trusted*?
- When do we have *negative legacy*?
- When *labels are unbiased*? (Human raters)

Predictions in conjunction with other "signals"

These questions are highly related to *the context* (e.g. ecosystems) in which the AI is designed/ deployed.
They cannot always be answered technically…

→ *Trust in the ecosystem*

(*) Source *Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements*
Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, Ed H. Chi
(Submitted on 14 Jan 2019)

Incompatible types of fairness

Known Trade Offs (**Incompatible types of fairness**):
- Equal positive and negative predictive value vs. equalized odds
- Equalized odds vs. equal allocation
- Equal allocation vs. equal positive and negative prediction value

Which type of fairness is appropriate for the given application and what level of it is satisfactory?

It requires not only Machine Learning specialists, but also clinical and ethical reasoning.

Source. Alvin Rajkomar et al. Ensuring, Fairness in Machine Learning to Advance Health, Equity, Annals of Internal Medicine (2018). DOI: 10.7326/M18-1990
Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594166/

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.

Re-asses Ethical Issues and Flags

- Execution of Paths may imply that Ethical issues and Flags are re-assessed and revised;
- The process reiterates from until a *stop* is reached.

Resolve

Next Steps

- (Optional) Scores/Labels are defined;
- Address, Resolve Tensions;
- Recommendations are given;
- (Optional) Trade off decisions are made;
- (Optional) Ethical maintenance starts.

Decide on Trade offs

- **Appropriate use:** Assess if the data and algorithm are appropriate to use for the purpose anticipated and perception of use.
  - Suppose we assess that the AI is technically *unbiased* and *fair* –this does not imply that it is acceptable to deploy it.
- **Remedies**: If risks are identified, define ways to mitigate risks (when possible)
- **Ability to redress**

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.

Possible (un)-wanted side-effects

- Assessing the ethics of an AI, may end up resulting in an ethical inspection of the entire *context* in which AI is designed/deployed…

- Could raise issues and resistance..

Co-design of a Trustworthy AI System in Healthcare. Best Practice: Deep Learning based Skin Lesion Classifier.

- Roberto V. Zicari (1), Sheraz Ahmed (39), Julia Amann (48), Stephan Alexander Braun (23)(58), John Brodersen (4)(9), Frédérick Bruneault (36), James Brusseau (8), Erik Campano (55) ,Megan Coffee (18), Andreas Dengel (39), Boris Düdder (6), Alessio Gallucci (28), Thomas Krendl Gilbert (15), Philippe Gottfrois (33), Emmanuel Goffi (16), Christoffer Bjerre Haase (34), Thilo Hagendorff (29), Eleanore Hickman (45), Elisabeth Hildt (17), Sune Holm (25), Pedro Kringen (1), Ulrich Kühne (32), Adriano Lucieri (39), Vince I. Madai (27)(56)(57), Pedro A. Moreno-Sánchez(53) , Oriana Medlicott(54), Matiss Ozols (14)(59), Eberhard Schnebel (1), Andy Spezzati (11), Jesmin Jahan Tithi (2), Steven Umbrello (52), Dennis Vetter (1), Holger Volland (40), Magnus Westerlund (5), Renee Wurth.(42)

Co-design of trustworthy AI

- Co-design of trustworthy AI in healthcare using a holistic approach, rather than monolithic ethical checklists.

The Initial Aim of the AI prototype

- A team led by Prof. Andreas Dengel at the German Research Center for Artificial Intelligence (DFKI) developed a framework for the domain-specific explanation of arbitrary Neural Network (NN)-based classifiers.
- Dermatology has been chosen as a first use case for the system.

The Initial Aim of the AI prototype

Status: AI System in early design phase.

The Research Questions

- How do we help engineers to design and implement a trustworthy AI system for this use case?
- What are the potential pitfalls of the AI system and how might they be mitigated at the development stage?

Lessons Learned

- **Co-Design: Think Holistically**
- **Re-evaluate and understand what is the "aim" of the system.**
- **Consider Different Viewpoints**

Lessons Learned

- **Measure the risk of harming**

- **Look for Similarities**

- **Consider the *aim* of the future AI system as a *claim***

Lessons Learned

- **Is Bias justifiable?**
- **Verify if Transparency is a prerequisite for Explanation**
- **Involve Patients**
- **Consider the Legal, Technical and Ethical Perspectives**

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