Trustworthy AI

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

Zurich Insurance Group
October 25, 2021

Session Content

1. Why Trustworthy AI?

2. The EU "Framework for Trustworthy AI" defined by the independent High-Level Expert Group on Artificial Intelligence set by the European Commission: Basic concepts.

3. European Commission: **Proposed Legal Framework on AI**.
Risks Classification. Plans for a **Conformity Assessment for AI.**

4. Introduction to **Trustworthy AI self-assessment and the ALTAI web tool.**

50 minutes session+ Interactive Q&A (40 minutes)

.

1. Why Trustworthy 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

The *Promises* of AI for the insurance industry

- The insurance industry includes numerous manual tasks that can be automated with AI and machine learning.

- With the advances in AI, insurance companies can provide faster services, ensuring customer satisfaction.

- As a result, the interest in AI insurance has tripled since 2012, according to Google Trends.

Source: https://research.aimultiple.com/ai-insurance/

Potential benefits of AI for the insurance industry

Potential benefits of implementing AI into insurance processes are:

- Time and cost-saving
- Improved customer experience
- Increasing profitability due to more accurate customer pricing and reduced fraudulent claims

Source: https://research.aimultiple.com/ai-insurance/

Possible use cases of AI in insurance

- Claims processing
- Appeals processing
- Insurance pricing
- Document creation
- Responding to customer queries
- Claim fraud detection
- Personalized services

Source: https://research.aimultiple.com/ai-insurance/

AI technologies that can be used for the insurance industry

- **Deep Learning**

With the increasing amount of customer data, insurance companies can build machine learning models to evaluate customer risk profiles more accurately and provide optimal insurance prices.

- **Document Processing**
- **Chatbots**
- **Affective Computing**

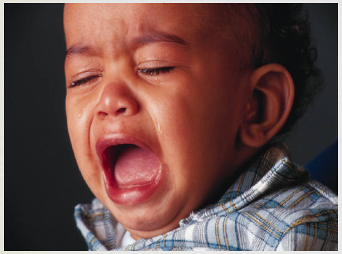
Source: https://research.aimultiple.com/ai-insurance/

The impact of AI on the future of insurance

- *Insurance 2030 – The impact of AI on the future of insurance*

  McKinsey] Company March 12, 2021 | Article

https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance



Do no harm
Can we explain decisions?

What if the decision made using AI-driven algorithm harmed somebody, and you cannot explain how the decision was made?

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

Based on our research work
Assessing Trustworthy AI: Best Practices

- 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. Best Practice: **Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients**.(On going 2021)

Consider the Entire AI Life Cycle

- **Design**
- **Development**
- **Deployment**
- **Monitoring**

Include the Full Community of Stakeholders

- At all stages of the AI life cycle, it is important to bring together a broader set of stakeholders.

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**

Assessing Trustworthy AI Best Practice: Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in 112 Emergency Calls

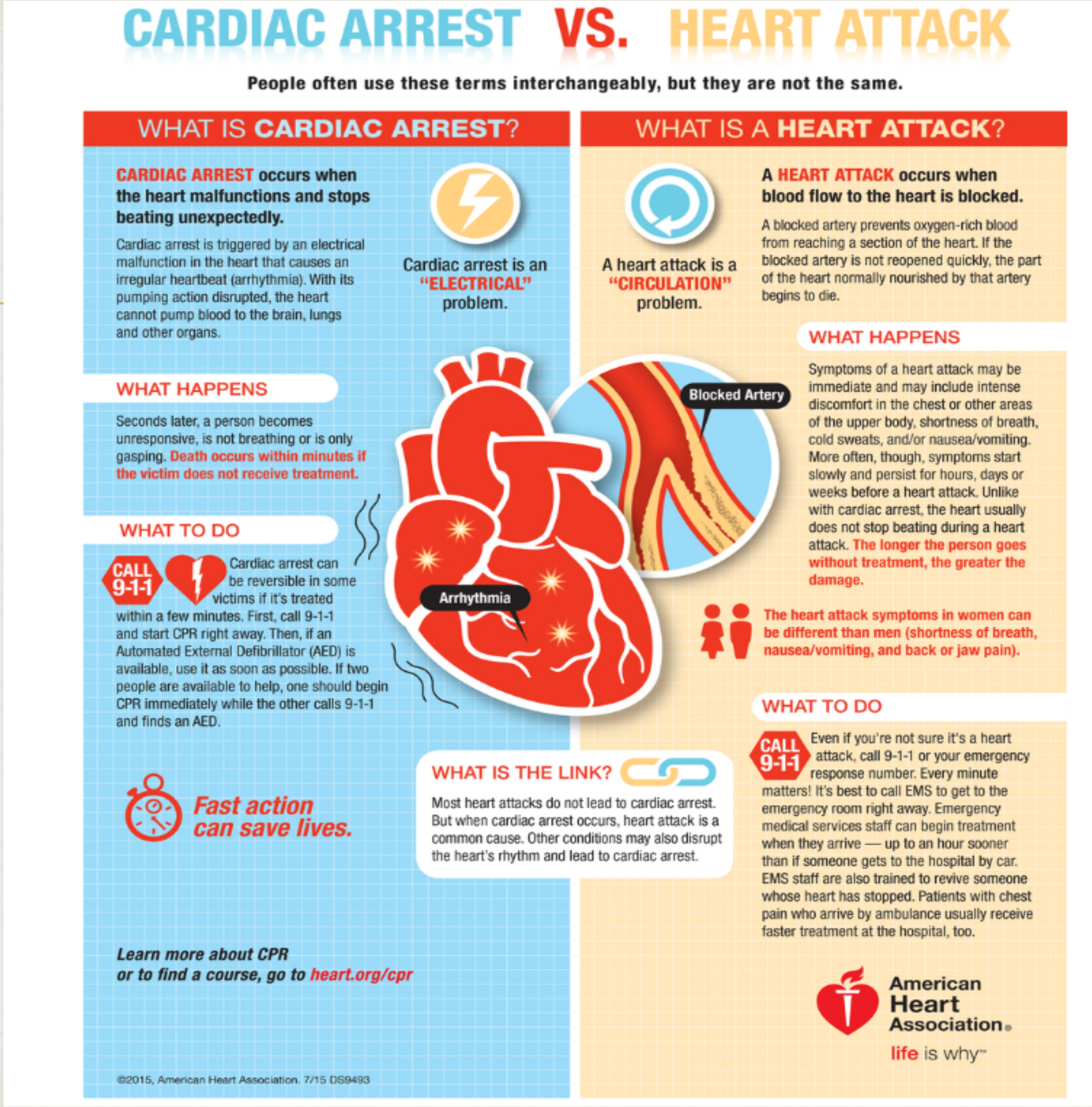
- **Health-related emergency calls** (**112**) are part of the Emergency Medical Dispatch Center (EMS) of the **City of Copenhagen**, triaged by medical dispatchers (i.e., medically trained dispatchers who answer the call, e.g., nurses and paramedics) and medical control by a physician on-site (EMS).

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

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.

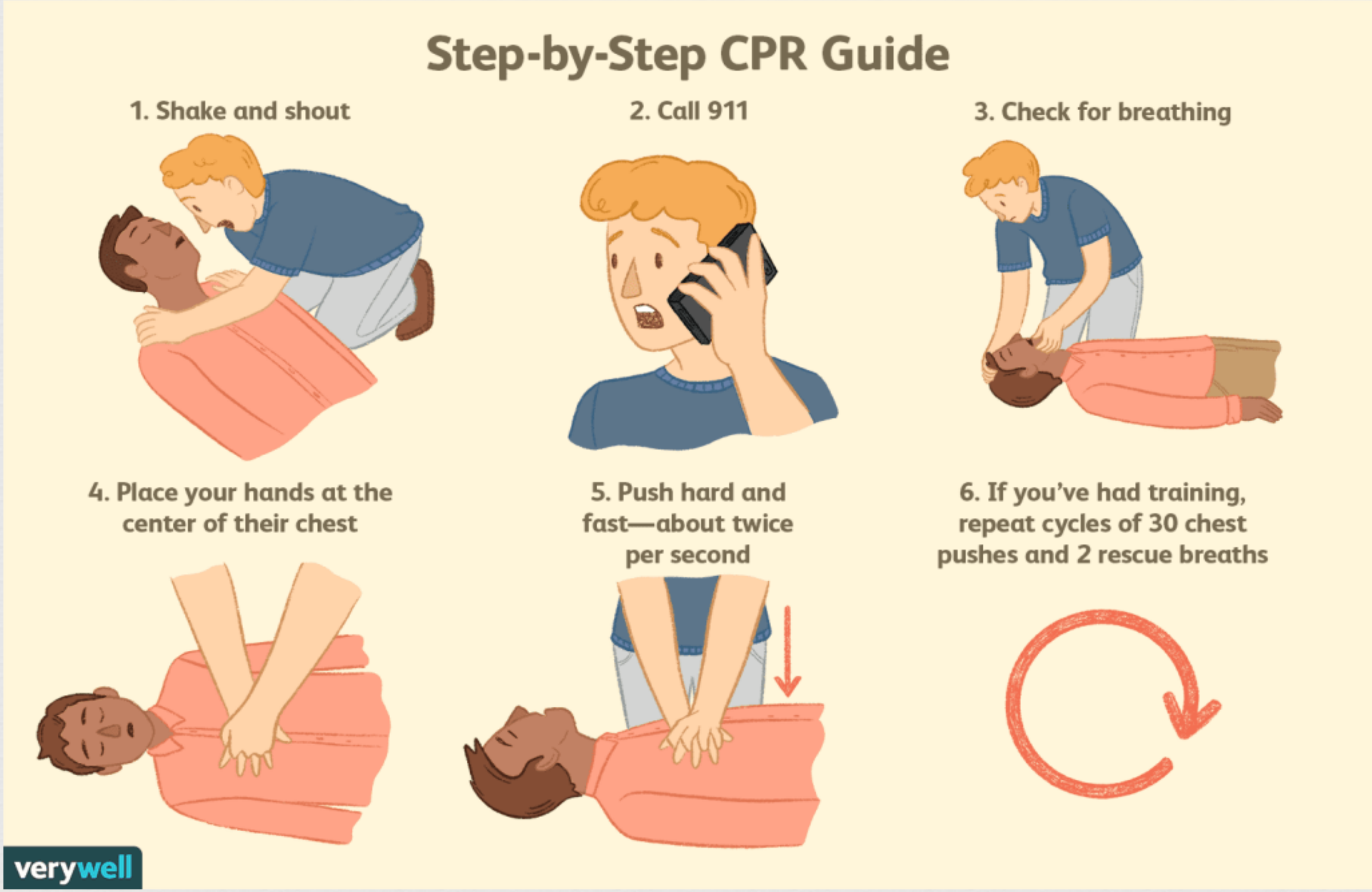
American Heart Association®
life is why™

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)



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

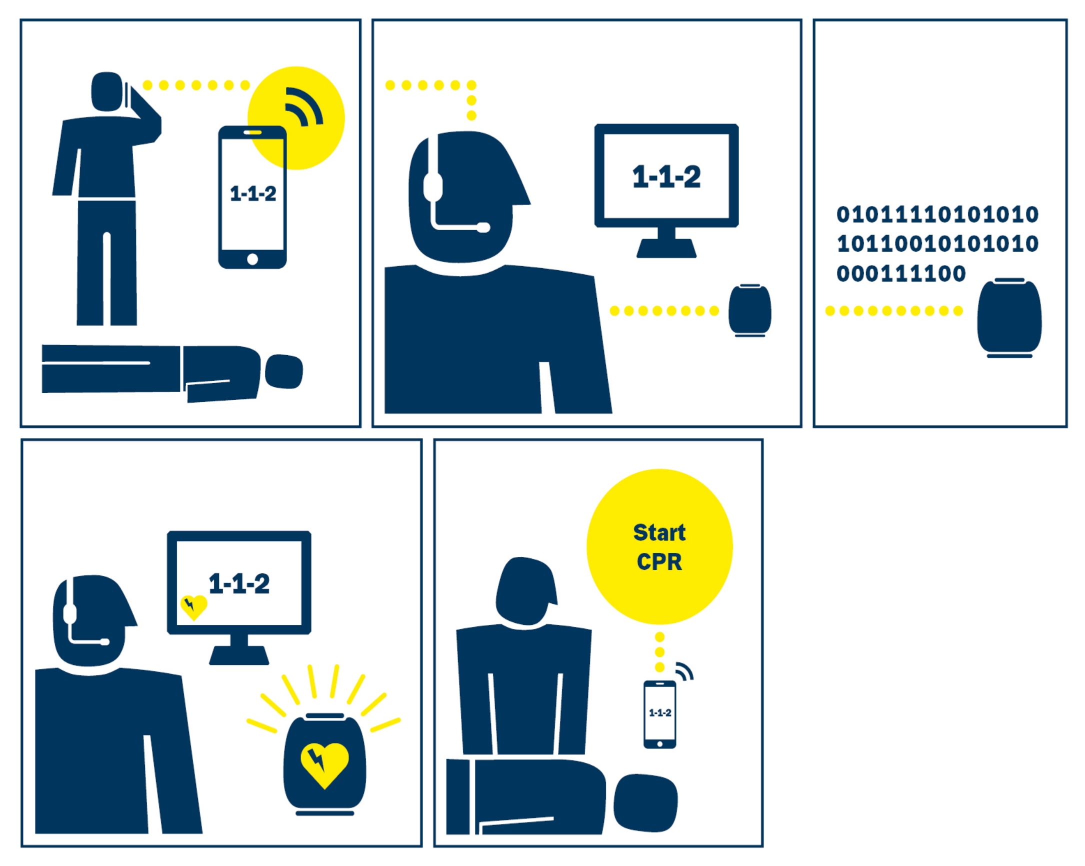


Figure . Ideal Case of Interaction between Bystander, Dispatcher, and the ML System. (with permission from Blomberg, S. N 2019b)

Retrospective study

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

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.

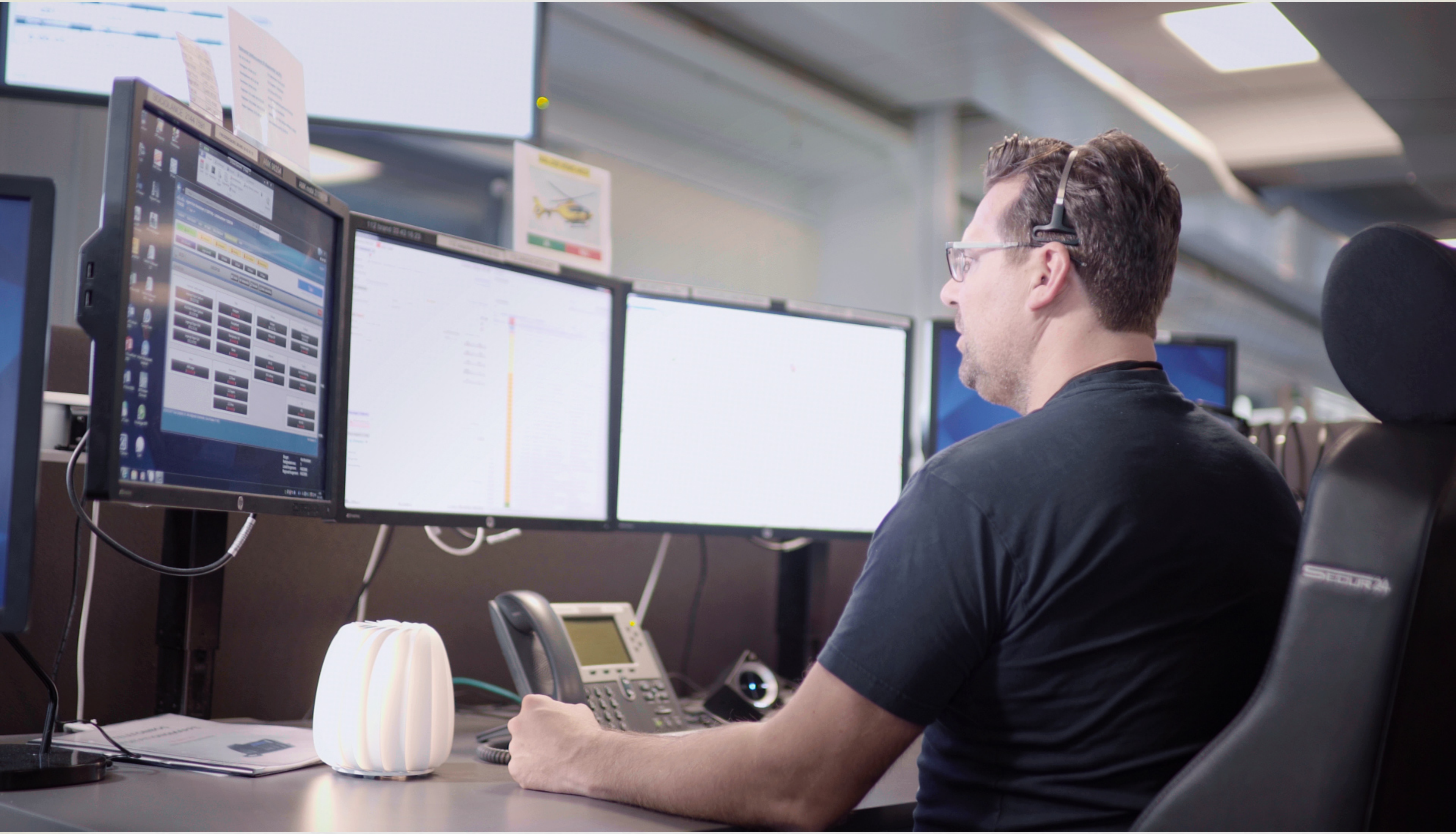
Randomized clinical trial (Cont.)

- **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.

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

Key Questions

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

Motivation

- We conducted a **self-assessment** conducted jointly by our team of independent experts together with the prime stakeholder of this use case.
- The main motivation of this work was to study 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.

Assess

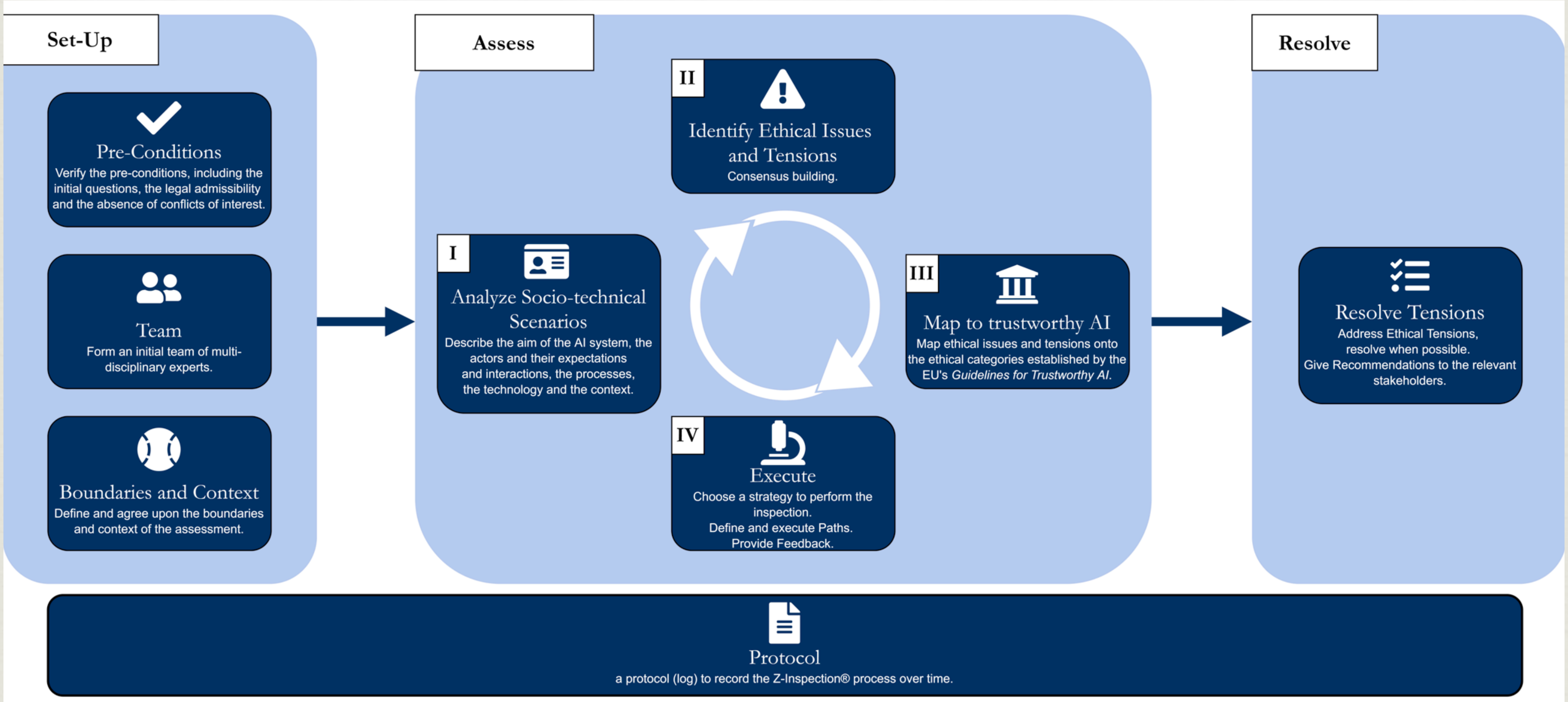
A holistic approach

- 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*.

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.

Identification of possible "issues"

Using the Z-inspection® process we identified

- **Ethical, Technical, Legal and Domain specific *issues.***

Ethics

- A branch of philosophy
  - ➨An appraisal of what is right (good) and wrong (bad)
  - ➨An assessment of human actions

Source: **The Ethics of Artificial Intelligence (AI) (Dr. Emmanuel Goffi)**
http://www.bigdata.uni-frankfurt.de/wp-content/uploads/2020/01/Goffi-2020-The-Ethics-of-Artificial-Intelligence-AI_Ethical-Implications-of-AI-SS2020.pdf

2. EU Framework for Trustworthy Artificial Intelligence

Fundamental values

- "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*

Ethical Business Regulation: Understanding the Evidence

- **Christopher Hodges** wrote a report (February 2016) for the UK Department for Business Innovation & Skills addressing the following question:

- *How public regulators in a contemporary* ***Western European democracy*** *should seek to affect the market behaviour of traders ?*

5 Key Recommendations

- 1.

Compliant behaviour cannot be guaranteed by regulation alone.

Ethical culture in business is an essential component that should be promoted and not undermined.

Key Recommendations

- 2.

Regulatory and other systems need to be designed **to provide evidence of business commitment to ethical behaviour**, on which **trust** can be based.

Key Recommendations

- 3.

Systemic learning has to be based on capture of information, and that **maximising the reporting of problems requires a no blame culture.**

Key Recommendations

- 4.

Regulation will be most effective where it is based on **the collaborative involvement of all parties.**

Key Recommendations

- 5.

Society needs to be protected from those who seek to break laws, and that people expect that wrongdoing deserves proportionate sanctions.

How can businesses behave ethically?

- **“The requirement is for a business to adopt ethical business practices in everything that is done throughout the organisation**.
- Codes on individual aspects, such as production, waste, marketing or social responsibility, are not enough: **the approach has to be holistic**.
- It has to be led from the top, but to exist at every level of the social groups within an organisation. “

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Who establishes Core Values?

- “Studies on the causes of sustained long term business success have concluded that it is critical **to establish clear *core values***, which are shared by ***all*** members of the workforce, form an ideology that is enduring and able to be applied consistently in different trading and geographical circumstances, whilst operational goals are constantly examined and developed. “

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Evidence of trust

- “It is essential to **provide *evidence of trust*** that an organisation operates **with *ethical values*,** to support independent judgment on whether an expectation of ethical behaviour is warranted.

- **Mere claims by a company that it can be trusted will clearly not suffice**. *Mechanisms should be designed to produce reliable evidence of trust*. “

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Response to adverse events

- "In most cases, where people are acting *in good faith*, the appropriate response to adverse events is to support them to analyse and learn rather than to blame. Failures should be noted and acknowledged, rather than ignored, and an appropriate response made. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Response to adverse events

- "Where actions are *immoral, or accountability has not been observed*, a proportionate response should be made. Enforcement policies should generally avoid the concept of deterrence, since it has limited effect on behaviour, conflicts with a learning-based performance culture, and is undemocratic. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges

Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Response to adverse events

- “Where sanctions are imposed, the totality of the sanctions should be proportionate to the degree of moral culpability involved. That requires an equalisation as between all of the various factors: social response, reputational and public response, employment discipline response, civil redress response, and regulatory or criminal response. “

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

“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 in AI for healthcare: *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.”**

“The deployment of machine learning in medicine might resurge the debate between *naturalists* and *normativists*.

-- 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

Implication of IP on Trustworthy AI

- There is an inevitable trade off to be made between disclosing all activities of a trustworthy AI assessment vs. delaying them to a later stage.

Benjamin Haibe-Kains, et al. The importance of transparency and reproducibility in artificial intelligence research. (Submitted on 28 Feb 2020 (v1), last revised 7 Mar 2020 (this version, v2))
https://arxiv.org/pdf/2003.00898.pdf

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)

2. EU Framework for Trustworthy Artificial Intelligence

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.

EU Framework for Trustworthy Artificial Intelligence

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.

The principle of respect for human autonomy

“The fundamental rights upon which the EU is founded are directed towards ensuring **respect for the freedom and autonomy of human beings.**

- Humans interacting with AI systems must be able to keep full and effective self- determination over themselves, and be able to partake in **the democratic process**. “

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

The principle of respect for human autonomy (cont.)

- " AI systems **should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans.**

- Instead, **they should be designed to augment, complement and empower human cognitive, social and cultural skills.**

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

Human oversight

“The allocation of functions between humans and AI systems should follow human-centric design principles and **leave meaningful opportunity for human choice**.

This means **securing human oversight** over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work. “

The principle of prevention of harm

8 April, 2019.

- "AI systems **should neither cause nor exacerbate harm or otherwise adversely affect human beings.**
- This entails the **protection of human dignity as well as mental and physical integrity**.
- AI systems and the environments in which they operate must **be safe and secur**e. They must be **technically robust** and it should be ensured that they are not open to malicious use."

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

The principle of prevention of harm (cont.)

- “ **Vulnerable persons** should receive greater attention and be included in the development, deployment and use of AI systems.
- Particular attention must also be paid to situations where AI systems can **cause or exacerbate adverse impacts due to asymmetries of power or information,** such as between employers and employees, businesses and consumers or governments and citizens.
- Preventing harm also entails consideration of the **natural environment and all living beings**."

The principle of fairness

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

- **“The development, deployment and use of AI systems must be fair.**

There are many different interpretations of fairness; fairness has both a (i) substantive and a (ii) procedural dimension.

(i) The substantive dimension implies a commitment to: **ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.”**

The principle of fairness (cont.)

- **If unfair biases can be avoided, AI systems could even increase societal fairness**.

Equal opportunity in terms of access to education, goods, services and technology should also be fostered.

- **The use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice.**
- Additionally, fairness implies that AI practitioners should respect the **principle of proportionality** between means and ends, and consider carefully how to balance competing interests and objectives**. "**

The principle of fairness (cont.)

- (ii) The procedural dimension **of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them**.

In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

Fairness is closely linked to the rights to

Non-discrimination, Solidarity and Justice

The principle of explicability

- “Explicability is crucial for building and maintaining users’ **trust in AI systems.**

- This means that **processes need to be transparent**, the capabilities and purpose of AI systems openly communicated, **and decisions – to the extent possible – explainable to those directly and indirectly affected.** Without such information, a decision cannot be duly contested. “

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

AI as a ‘black box’

- “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 rights.
- 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. “

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.

Tensions and Trade-offs

- „**Tension**s may arise between the above 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

Example of Tensions and Trade-offs

- In various application domains, *the principle of prevention of harm* and *the principle of human autonomy* may be in conflict.

- Consider as an example the use of AI systems for 'predictive policing', which may help to reduce crime, but in ways that entail surveillance activities that impinge on individual liberty and privacy. "

Tensions and Trade-offs (cont.)

- "AI practitioners can hence not be expected to find the right solution based on the ethical principles, yet they should approach ethical dilemmas and trade-offs via reasoned, evidence-based reflection rather than intuition or random discretion.

- There may be situations, however, where no ethically acceptable trade-offs can be identified."

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

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

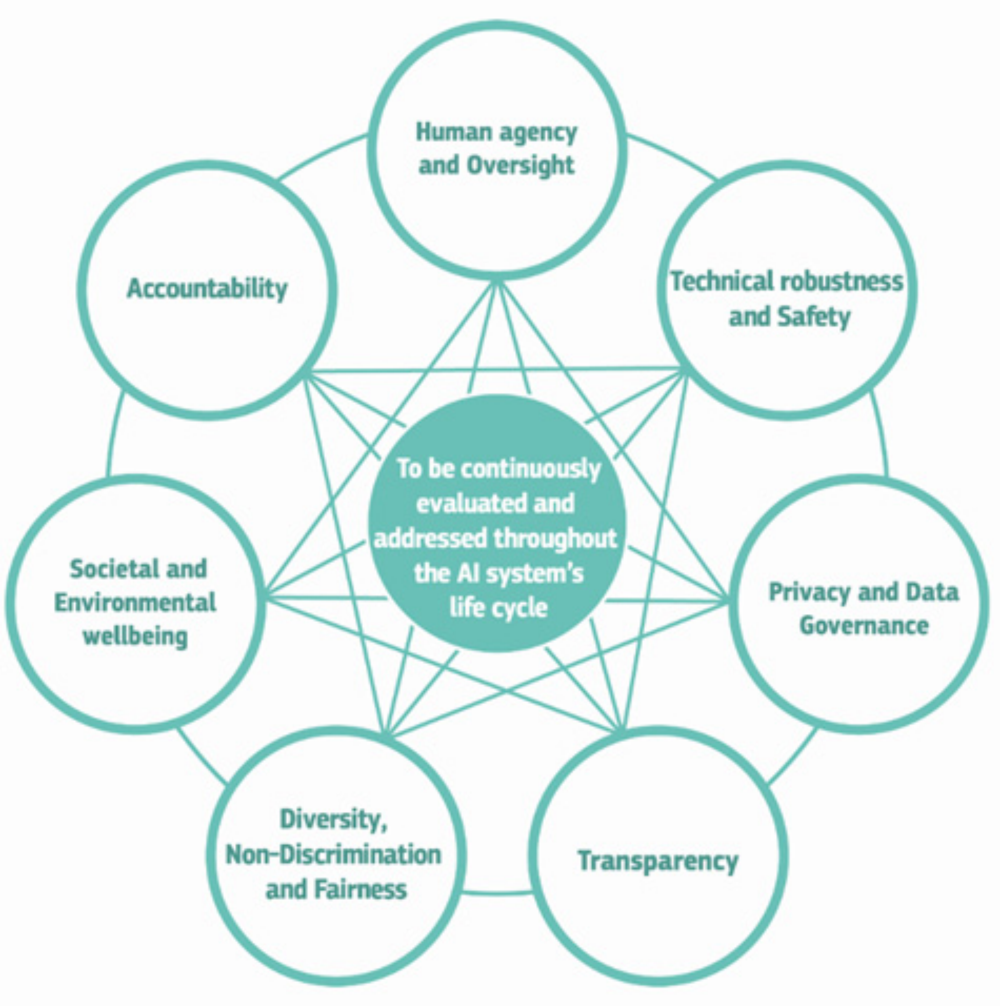
Including sustainability and environmental friendliness, social impact, society and democracy

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.

Requirements of Trustworthy AI


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

Human agency and oversight

- All potential impacts that AI systems may have on fundamental rights should be accounted for and that the human role in the decision- making process is protected.

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.

Technical robustness and safety

- AI systems should be secure and resilient in their operation in a way that minimizes potential harm, optimizes accuracy, and fosters confidence in their reliability;

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.

Privacy and data governance

- Given the vast quantities of data processed by AI systems, this principle impresses the importance of protecting the privacy, integrity, and quality of the data and protects human rights of access to it;

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.

Transparency

- AI systems need to be understandable at a human level so that decisions made through AI can be traced back to their underlying data. If a decision cannot be explained it cannot easily be justified;

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.

Diversity, non-discrimination, and fairness

- AI systems need to be inclusive and non- biased in their application. This is challenging when the data is not reflective of all the potential stakeholders of an AI system;

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications

Societal and environmental wellbeing

- In acknowledging the potential power of AI systems, this principle emphasizes the need for wider social concerns, including the environment, democracy, and individuals to be taken into account;

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications

Accountability

- This principle, rooted in fairness, seeks to ensure clear lines of responsibility and accountability for the outcomes of AI systems, mechanisms for addressing trade-offs, and an environment in which concerns can be raised. However, the interpretation, relevance, and implementation of trustworthy AI depends on the domain and the context where the AI system is used.

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications

Challenges and Limitations

The AI HLEG trustworthy AI guidelines 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.).

They mainly offer a static checklist (AI HLEG 2020) and do not take into account changes of the AI over time.

Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021

Challenges and Limitations

They do not distinguish different applicability of the AI HLEG trustworthy 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, monitoring.

There are not available best practices to show how to implement such requirements and apply them in practice.

The guidelines do not explicitly address the lawful part of the assessment.

Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021.

When Assessing Trustworthy AI Consider AI in the Context

- 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.

Use Socio-technical Scenarios

By collecting relevant resources, socio-technical scenarios are created and analyzed by a team of *interdisciplinary* 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.

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

Use a Catalogue 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.

Catalogue of Tensions

- **Accuracy *versus* fairness**:

an algorithm which is most accurate on average may systematically discriminate against a specific minority.

Catalogue of Tensions

- **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.

Catalogue of Tensions

- **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.

Catalogue of Examples of Tensions

- **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.

Catalogue of Examples of Tensions

- **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.

Catalogue of Examples of Tensions

- **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.

Catalog of Examples of Tensions

- **Satisfaction of preferences** *versus* **equality**:

automation and AI could invigorate industries and spearhead new technologies, but also exacerbate exclusion and poverty.

Catalog of Examples of Tensions

- **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.

e.g. AI deployed during COVID pandemic.

Use a 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".

Back to Our Use Case

Risks versus Benefits

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

Lack of Trust?

- 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 .

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.

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.**

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.

3. EU Proposed Legal Framework on AI

- “ The Commission proposes new rules and actions aiming to turn Europe into the global hub for trustworthy Artificial Intelligence (AI).
- The combination of the first-ever **legal framework on AI** and a new Coordinated Plan with Member States will guarantee the safety and fundamental rights of people and businesses, while strengthening AI uptake, investment and innovation across the EU.
- New rules on Machinery will complement this approach by adapting safety rules to increase users' trust in the new, versatile generation of products.”

Source EU press release
https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682

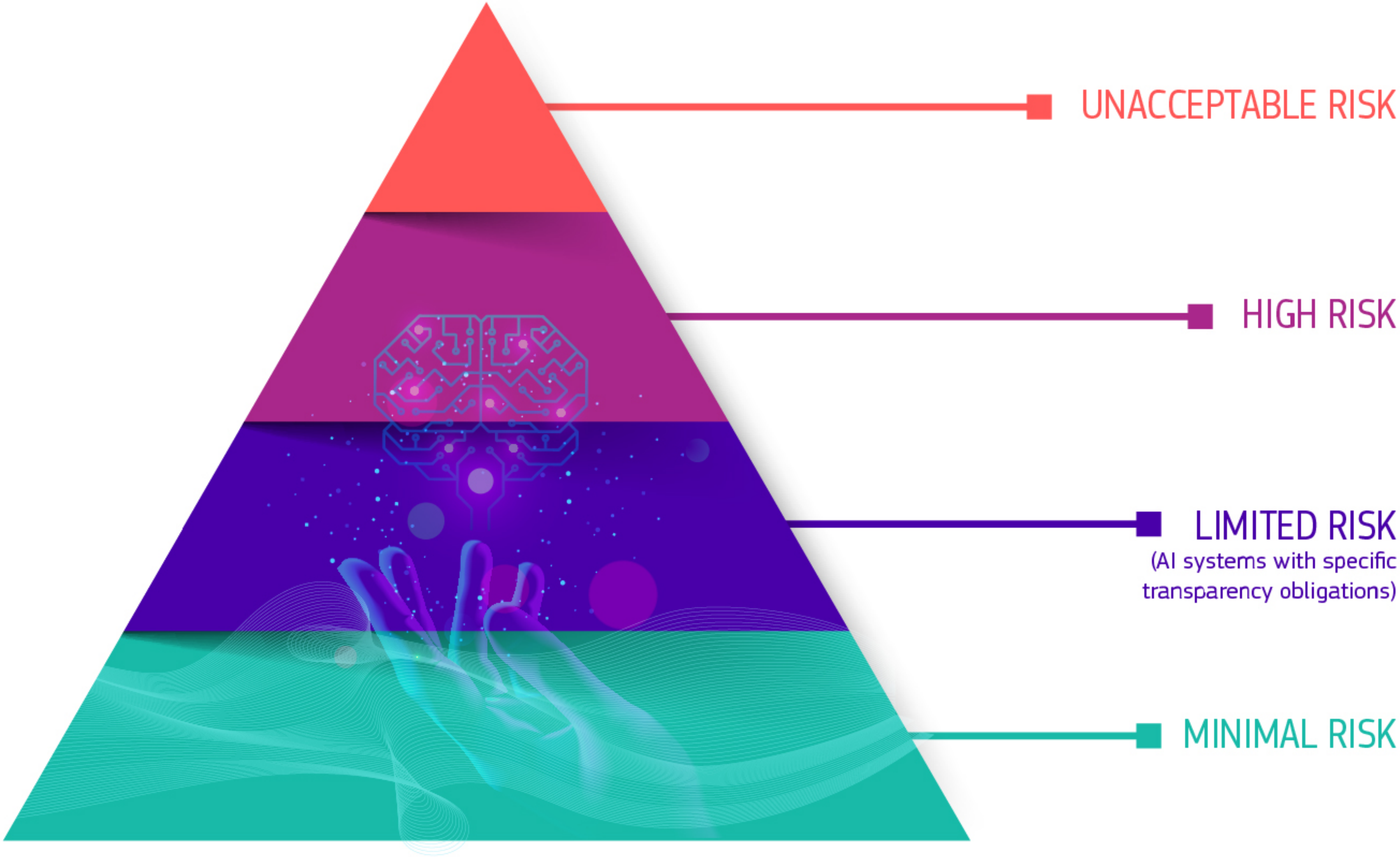
The EU Follows a Risk-adapted Regulatory Approach

- The German Data Ethics Commission recommended adopting a **risk-adapted regulatory approach** to algorithmic systems. The principle underlying this approach should be as follows: **the greater the potential for harm, the more stringent the requirements and the more far-reaching the intervention by means of regulatory instruments.**

- When assessing this potential for harm, the **sociotechnical system as a whole** must be considered, or in other words all the components of an algorithmic application, including all the people involved, from the development phase – for example the training data used – right through to its implementation in an application environment and any evaluation and adjustment measures.

Risk-adapted regulatory system

- **HIGH RISK** complete or partial ban of an algorithmic system
- additional measures such as live interface for “always on“ oversight by supervisory institutions
- additional measures such as ex-ante approval procedures
- measures such as formal and substantive requirements (e. g. transparency obligations, publication of a risk assessment) or monitoring procedures (e.g. disclosure obligations towards supervisory bodies, ex-post controls, audit procedures)
- **LOW RISK** no special measures



UNACCEPTABLE RISK

HIGH RISK

LIMITED RISK
(AI systems with specific
transparency obligations)

MINIMAL RISK

EU: AI systems identified as **high-risk** include AI technology used in

- **Critical infrastructures** (e.g. transport), that could put the life and health of citizens at risk;
- **Educational or vocational training**, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);
- **Safety components of products** (e.g. AI application in robot-assisted surgery);
- **Employment, workers management and access to self-employment** (e.g. CV-sorting software for recruitment procedures);
- **Essential private and public services** (e.g. credit scoring denying citizens opportunity to obtain a loan);
- **Law enforcement** that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);
- **Migration, asylum and border control management** (e.g. verification of authenticity of travel documents);
- **Administration of justice** and **democratic processes** (e.g. applying the law to a concrete set of facts).

High-risk AI systems will be subject to **strict obligations** before they can be put on the market:

- **Adequate risk assessment and mitigation systems;**
- **High quality of the datasets** feeding the system to minimise risks and discriminatory outcomes;
- **Logging of activity to ensure traceability of results**;
- **Detailed documentation** providing all information necessary on the system and its purpose for authorities to assess its compliance;
- **Clear and adequate information** to the user;
- **Appropriate human oversight** measures to minimise risk;
- High level of **robustness**, **security** and **accuracy**.

Remote biometric identification

- **All remote biometric identification** systems are considered high risk and subject to strict requirements.

Limited and Minimal Risks

- **Limited risk**, i.e. AI systems with specific transparency obligations: When using AI systems such as chatbots, users should be aware that they are interacting with a machine so they can take an informed decision to continue or step back.
- **Minimal risk:** The legal proposal allows the free use of applications such as AI-enabled video games or spam filters. The vast majority of AI systems fall into this category. The draft Regulation does not intervene here, as these AI systems represent only minimal or no risk for citizens' rights or safety.

AI governance

- In terms of governance, the Commission proposes that national competent market surveillance authorities supervise the new rules, while the creation of a **European Artificial Intelligence Board** will facilitate their implementation, as well as drive the development of standards for AI.
- Additionally, *voluntary codes of conduct* are proposed for non-high-risk AI, as well as *regulatory sandboxes* to facilitate responsible innovation.

Regulatory sandboxes

- **Regulatory sandboxes** enable in a real-life environment the testing of innovative technologies, products, services or approaches, *which are not fully compliant with the existing legal and regulatory framework*.
- They are operated for *a limited time and in a limited part* of a sector or area.
- The purpose of regulatory sandboxes is *to learn about the opportunities and risk*s that a particular innovation carries and to develop the right regulatory environment to accommodate it.
- Experimentation clauses are often the legal basis for regulatory sandboxes.

Source: https://www.bmwi.de/Redaktion/EN/Dossier/regulatory-sandboxes.html

Conformity assessments

- According to the EU proposal, for **healthcare products** (generally medical devices) **conformity assessment** would see the involvement of professional certification bodies already designated under the Regulations on medical devices. They are called "**Notified Bodies**".

Conformity assessments

- For **healthcare-related AI systems, which are not medical devices, but are in the "high-risk"** list of Annex III (e.g. AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid)**, for now the EU foresee a conformity assessment done directly by the manufacturer** (who certainly can foresee an audit mechanism, even external, but which remains entirely under **the manufacturer's responsibility**).

4. Performing a Self Assessment using The ALTAI web-tool

Getting started - A step by guide

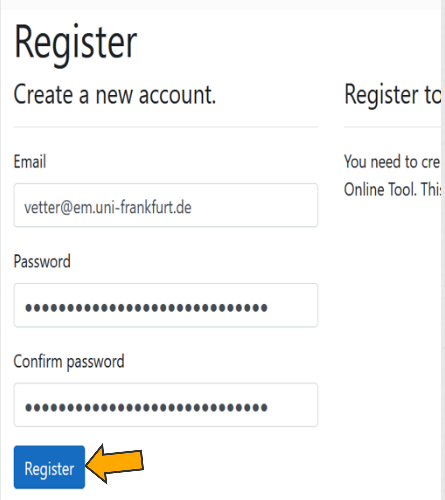
ALTAI TRUSTWORTHY AI ASSESSMENT LIST

Measure
if your organisation's AI is
trustworthy

ALTAI – Assessment List for Trustworthy Artifical Intelligence

Step 1: Sign up

- Go to https://altai.insight-centre.org/Identity/Account/Register to create an account
- Activate the account by clicking the link in the activation email
- Known problems:
  - If you don't receive an email within ~5 minutes, check spam, otherwise something went wrong
  - Emails attached to a google account seem to work every time



Register

Create a new account.

Register to

You need to cre
Online Tool. Thi

Email

vetter@em.uni-frankfurt.de

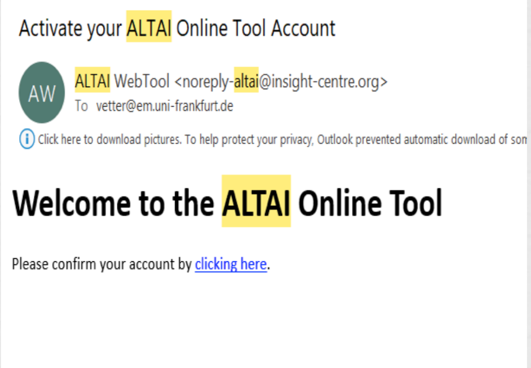
Password

•••••••••••••••••••••••••••••

Confirm password

•••••••••••••••••••••••••••••

Register



Activate your ALTAI Online Tool Account

AW ALTAI WebTool <noreply-altai@insight-centre.org>
To vetter@em.uni-frankfurt.de

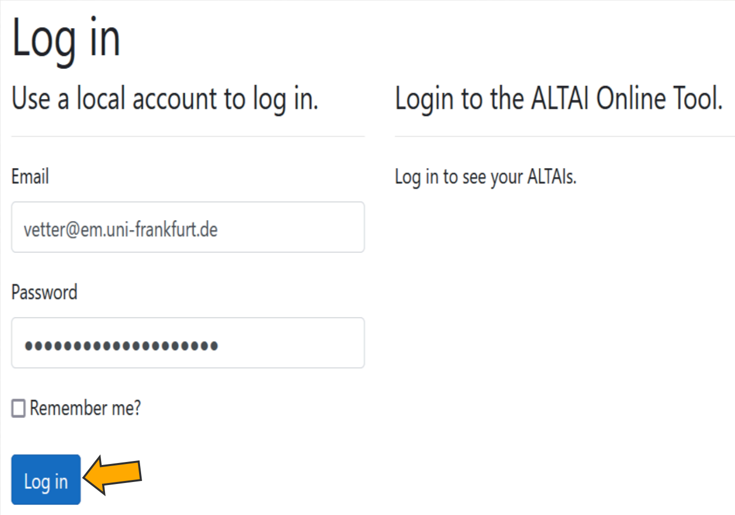
Click here to download pictures. To help protect your privacy, Outlook prevented automatic download of son

Welcome to the ALTAI Online Tool

Please confirm your account by clicking here.

Step 2: Login

- Go to https://altai.insight-centre.org/Identity/Account/Login
- Use your credentials from Step 1 to log in
- After successful login, the top bar changes



Log in

Use a local account to log in.

Email

vetter@em.uni-frankfurt.de

Password

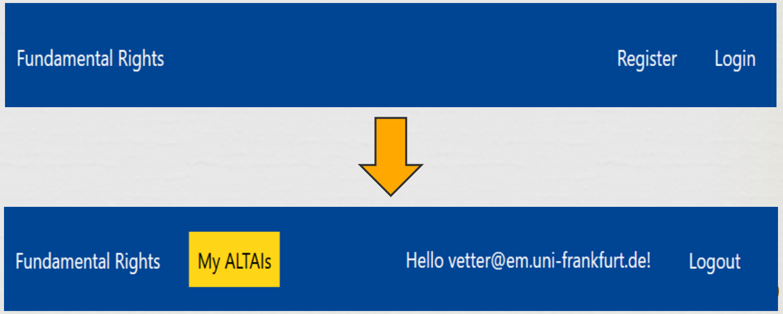
••••••••••••••••••••

□ Remember me?

Log in

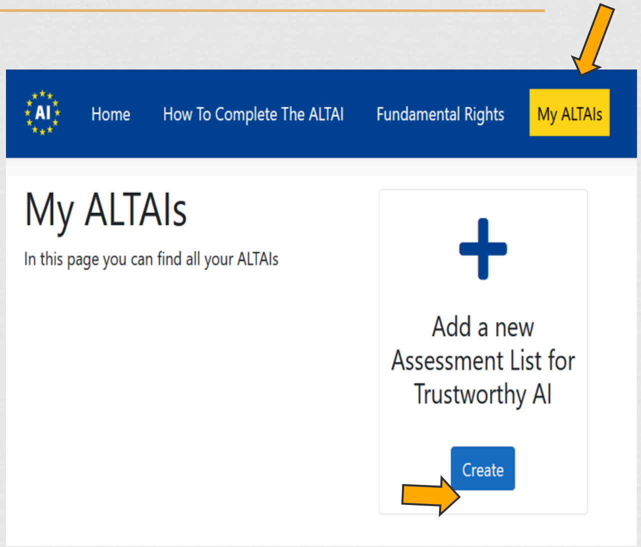
Login to the ALTAI Online Tool.

Log in to see your ALTAIs.



Step 3: Create a new Assessment List

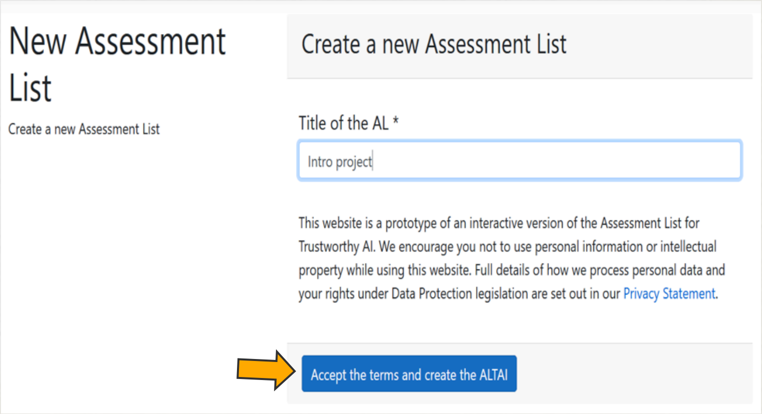
- Click the big yellow button ("My ALTAIs") or go to https://altai.insight-centre.org/Assessment to start a new Assessment List
- Click on create
- Give it a title
- Accept the terms and create



My ALTAIs

In this page you can find all your ALTAIs

Add a new Assessment List for Trustworthy AI



New Assessment List

Create a new Assessment List

Create a new Assessment List

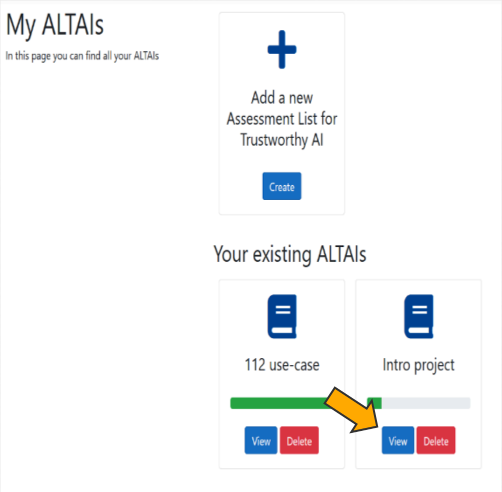
Title of the AL *

Intro project

This website is a prototype of an interactive version of the Assessment List for Trustworthy AI. We encourage you not to use personal information or intellectual property while using this website. Full details of how we process personal data and your rights under Data Protection legislation are set out in our Privacy Statement.

Step 4: Open the new Assessment List

- The new List appears at the end of the “Your existing ALTAIs” section at https://altai.insight-centre.org/Assessment
- Click on “View”
- Add notes to the List for better recollection of the project
- Click on “Submit” to go start with the actual assessment

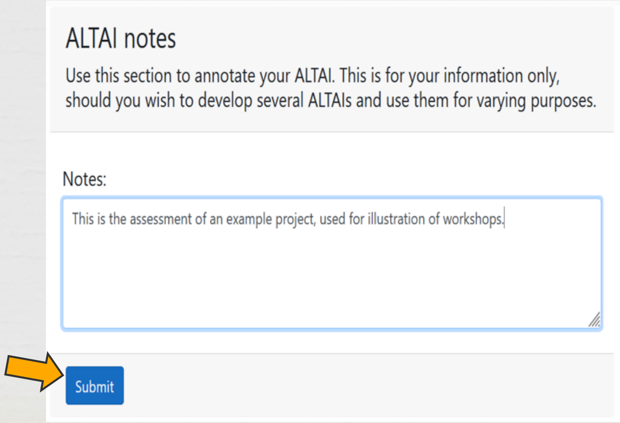


My ALTAIs

In this page you can find all your ALTAIs

Add a new Assessment List for Trustworthy AI

Your existing ALTAIs



ALTAI notes

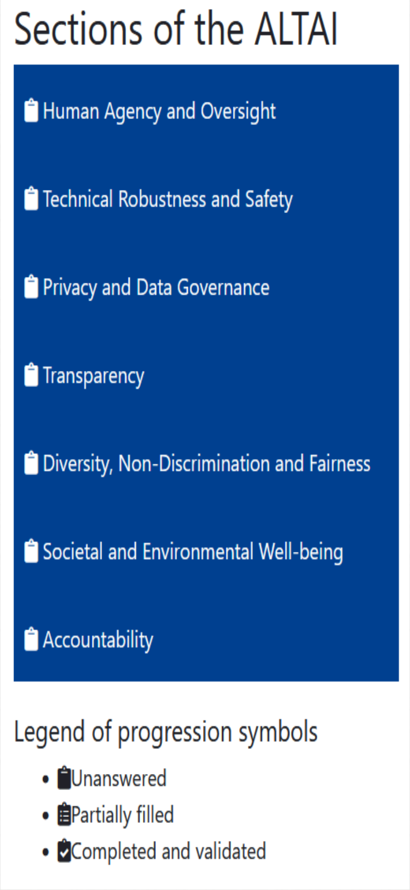
Use this section to annotate your ALTAI. This is for your information only, should you wish to develop several ALTAIs and use them for varying purposes.

Notes:

This is the assessment of an example project, used for illustration of workshops.

Step 5: Familiarize with the ALTAI structure

- There is a separate list for each of the seven key requirements, the sidebar shows you which you already processed
- Each list is separated into sections, each section corresponding to one of the sub-requirements
- Each list and section provides a short summary of the topic and why it is important for trustworthy AI

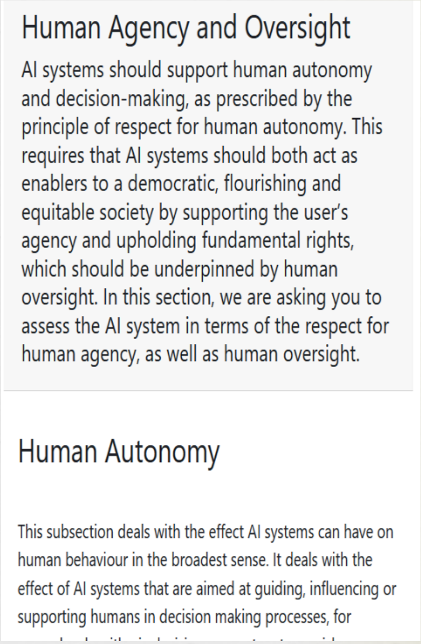


Sections of the ALTAI

- 📋 Human Agency and Oversight
- 📋 Technical Robustness and Safety
- 📋 Privacy and Data Governance
- 📋 Transparency
- 📋 Diversity, Non-Discrimination and Fairness
- 📋 Societal and Environmental Well-being
- 📋 Accountability

Legend of progression symbols

- 📋Unanswered
- 📋Partially filled
- ✅Completed and validated



Human Agency and Oversight

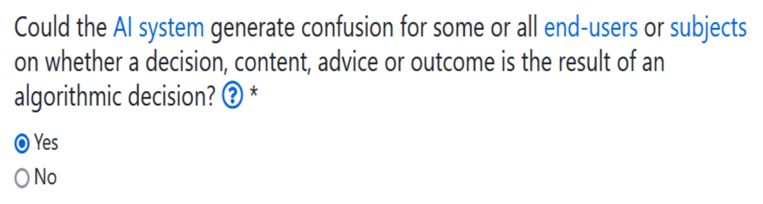
AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and upholding fundamental rights, which should be underpinned by human oversight. In this section, we are asking you to assess the AI system in terms of the respect for human agency, as well as human oversight.

Human Autonomy

This subsection deals with the effect AI systems can have on human behaviour in the broadest sense. It deals with the effect of AI systems that are aimed at guiding, influencing or supporting humans in decision making processes, for

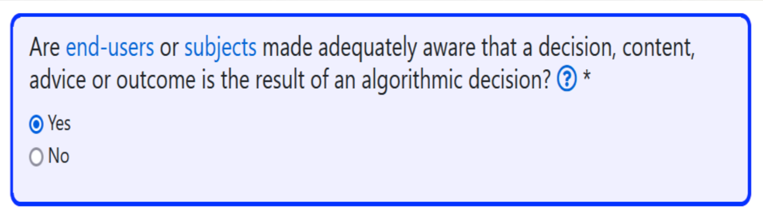
Step 6: Familiarize with the question types

- 3 types of questions
- White background => describe features
- Blue background => contributes to recommendations
- Red background => self-assessment based on the answers to the previous questions



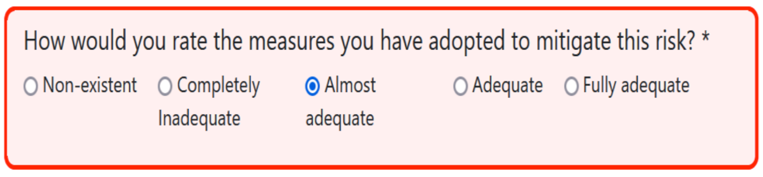
Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision? ? *

● Yes
○ No



Are end-users or subjects made adequately aware that a decision, content, advice or outcome is the result of an algorithmic decision? ? *

● Yes
○ No

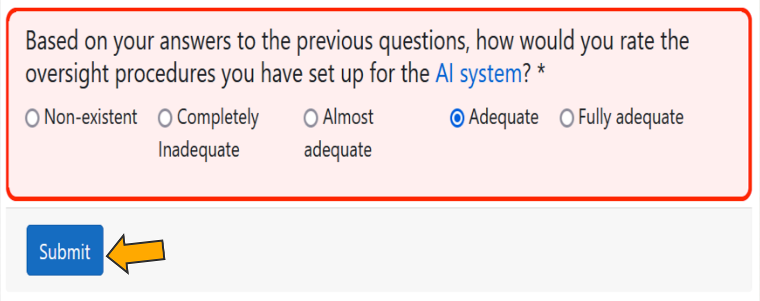


How would you rate the measures you have adopted to mitigate this risk? *

○ Non-existent ○ Completely Inadequate ● Almost adequate ○ Adequate ○ Fully adequate

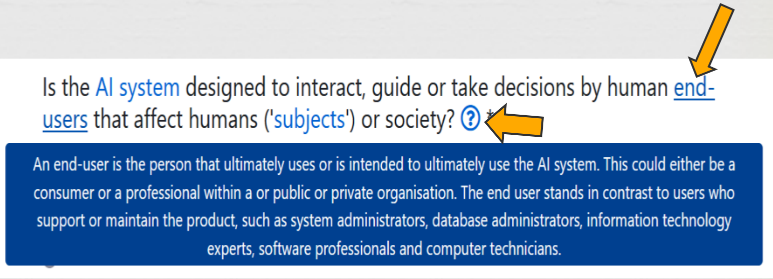
Step 7: Answer the questions

- Answering one question can make follow up questions appear
- Answer all questions until no new ones appear
- Questions might not always be applicable, in such cases just select something
- Click submit to complete the assessment for this key requirement
- Getting help:
  - Pointing the mouse on terms in blue to get a detailed definition of the term
  - Clicking on the (?) symbol at the end of a question gives more information why this question is important



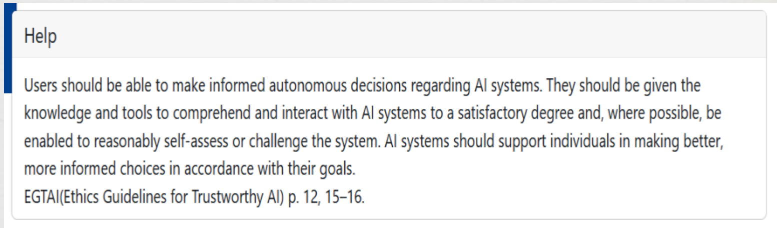
Based on your answers to the previous questions, how would you rate the oversight procedures you have set up for the AI system? *

○ Non-existent ○ Completely Inadequate ○ Almost adequate ● Adequate ○ Fully adequate



Is the AI system designed to interact, guide or take decisions by human end-users that affect humans ('subjects') or society? (?) *

An end-user is the person that ultimately uses or is intended to ultimately use the AI system. This could either be a consumer or a professional within a or public or private organisation. The end user stands in contrast to users who support or maintain the product, such as system administrators, database administrators, information technology experts, software professionals and computer technicians.

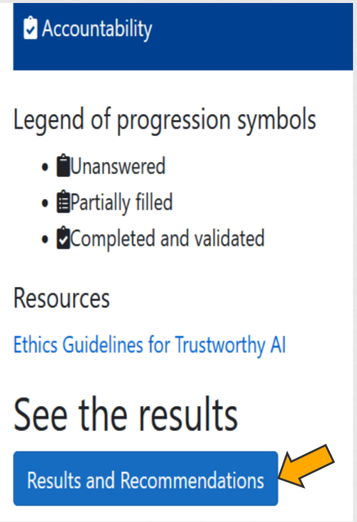


Help

Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. AI systems should support individuals in making better, more informed choices in accordance with their goals.
EGTAI(Ethics Guidelines for Trustworthy AI) p. 12, 15–16.

Step 8: Assessment Results

- After working through all the lists, go to the “Results and Recommendations” section
(in the sidebar on the left)
- The “spiderweb” shows the results of the self-assessment
  - 0 = non-existent measures to fulfill requirements
  - 5 = fully adequate measures to fulfill requirements



Accountability

Legend of progression symbols

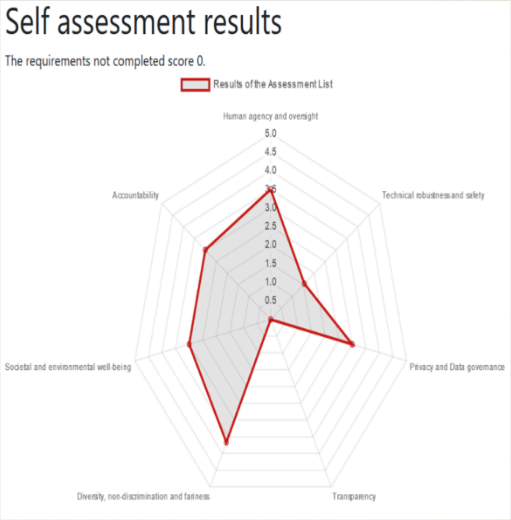
- Unanswered
- Partially filled
- ☑Completed and validated

Resources

Ethics Guidelines for Trustworthy AI

See the results

Results and Recommendations

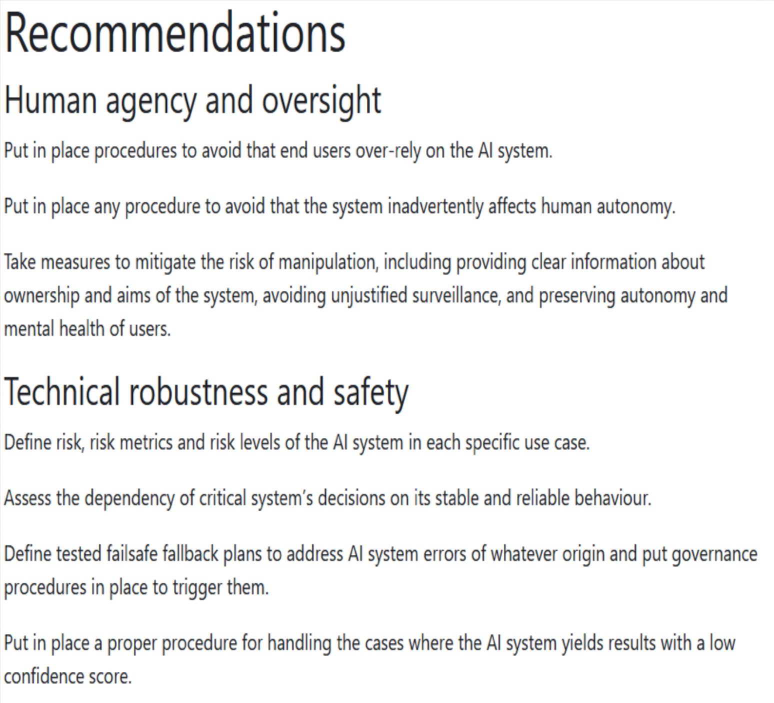


Self assessment results

The requirements not completed score 0.

Step 9: Recommendations

- On the same page as the assessment results
- Based on answers to blue questions
- Suggestions on how to improve the system so it better implements the key requirements for trustworthy AI
- Recommendations might be too general, check which ones might be useful



Recommendations

Human agency and oversight

Put in place procedures to avoid that end users over-rely on the AI system.

Put in place any procedure to avoid that the system inadvertently affects human autonomy.

Take measures to mitigate the risk of manipulation, including providing clear information about ownership and aims of the system, avoiding unjustified surveillance, and preserving autonomy and mental health of users.

Technical robustness and safety

Define risk, risk metrics and risk levels of the AI system in each specific use case.

Assess the dependency of critical system's decisions on its stable and reliable behaviour.

Define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them.

Put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score.

Resources

http://z-inspection.org