Towards An Inspection process to assess Ethical AI
A case study in health care.

Roberto V. Zicari
With contributions from: Irmhild van Halem, Matthew Eric Bassett, Karsten Tolle, Timo Eichhorn, Todor Ivanov.

Frankfurt Big Data Lab
www.bigdata.uni-frankfurt.de

JST-ISSIP Workshop - AI and physical, mental, and societal health
August 26, 2019
IBM Almaden, 650 Harry Rd, San Jose, CA 95120

Z-inspection

An Inspection process to assess Ethical AI

Photo: RVZ

Why doing an AI Ethical Inspection?

There are several reasons to do an AI Ethical Inspection:

- *Minimize Risks* associated with AI
- *Help establishing "TRUST"* in AI
- *Improve the AI*
- *Foster ethical values and ethical actions* (stimulate new kinds of innovation)

Help contribute to closing the gap between "*principles*" (the "what" of AI ethics) and "*practices*" (the "how").

Two ways to use Z-Inspection

1. As part of an *AI Ethics by Design* process,

and/or

2. if the *AI has already been designed,* it can be used to do an *AI Ethical sanity check,* so that a certain AI Ethical standard of care is achieved.

It can be used by a variety of AI stakeholders.

The *context* for the inspection *Ecosystems*

- The Rise of (Digital) Ecosystems paving the way to disruption.(*)

- Different Countries, Different Approaches, Cultures, Political Systems, and Values (e.g. China, the United States, Russia, Europe,…)

Ecosystems are part of the *context* for the inspection.

(*) Source: Digital Hospitality, Metro AG-personal communication.

What is the output of this investigation?

- *The output of this investigation is a degree of confidence that the AI analyzed* -taking into account the context (e.g. ecosystems), people, data and processes- *is ethical with respect to a scale of confidence*.

What to do with the output of this investigation?

- Based upon the score obtained, the process continues (when possible):
  - providing feedback to the AI designers (when available) who could change/improve the AI model/the data/ the training and/or the deployment of the AI in the context.
  - giving recommendations on how and when to use (or not) the AI, given certain constraints, requirements, and ethical reasoning (*Trade-off* concept).

Additional Positive Scoring Scale: Foster Ethical Values

In addition we could provide a score that identifies and defines AIs that in particular have been designed and result in production in *Fostering Ethical values and Ethical actions (FE)*

There is no negative score.

Precondition: Agree on selected principles for measuring the FE score.

Goal: reward and stimulate new kinds of Ethical innovation

Core Ethical Principle: *Beneficence. ("well-being", "common good"…)*

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 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

Model and Data Accessibility Levels

Level A++: AI in design, access to model, training and test data, input data, AI designers, business/government executives, and domain experts;

Level A+: AI designed (deployed), access to model, training and test data, input data, AI designers, business/government executives, and domain experts;

Level A- : AI designed (deployed), access to ONLY PART of the model (e.g. no specific details of the features used) , training and test data, input data,

Level **B**: AI designed (deployed), “black box”, NO access to model, training and test data, input data, AI designers, (business/government executives, and domain experts);

Pre-conditions

1. Agreement on *Context-specific ethical values*

2. Agreement on the *Areas of Investigation*

Example: Clinical Medical Ethics in the context of Ecosystems

The four classical principles of *Western* clinical medical ethics

- Justice
- Autonomy
- Beneficence
- Nonmaleficence

Where *Western* define a set of *ecosystems*.

Z-Inspection: *Areas of investigations*

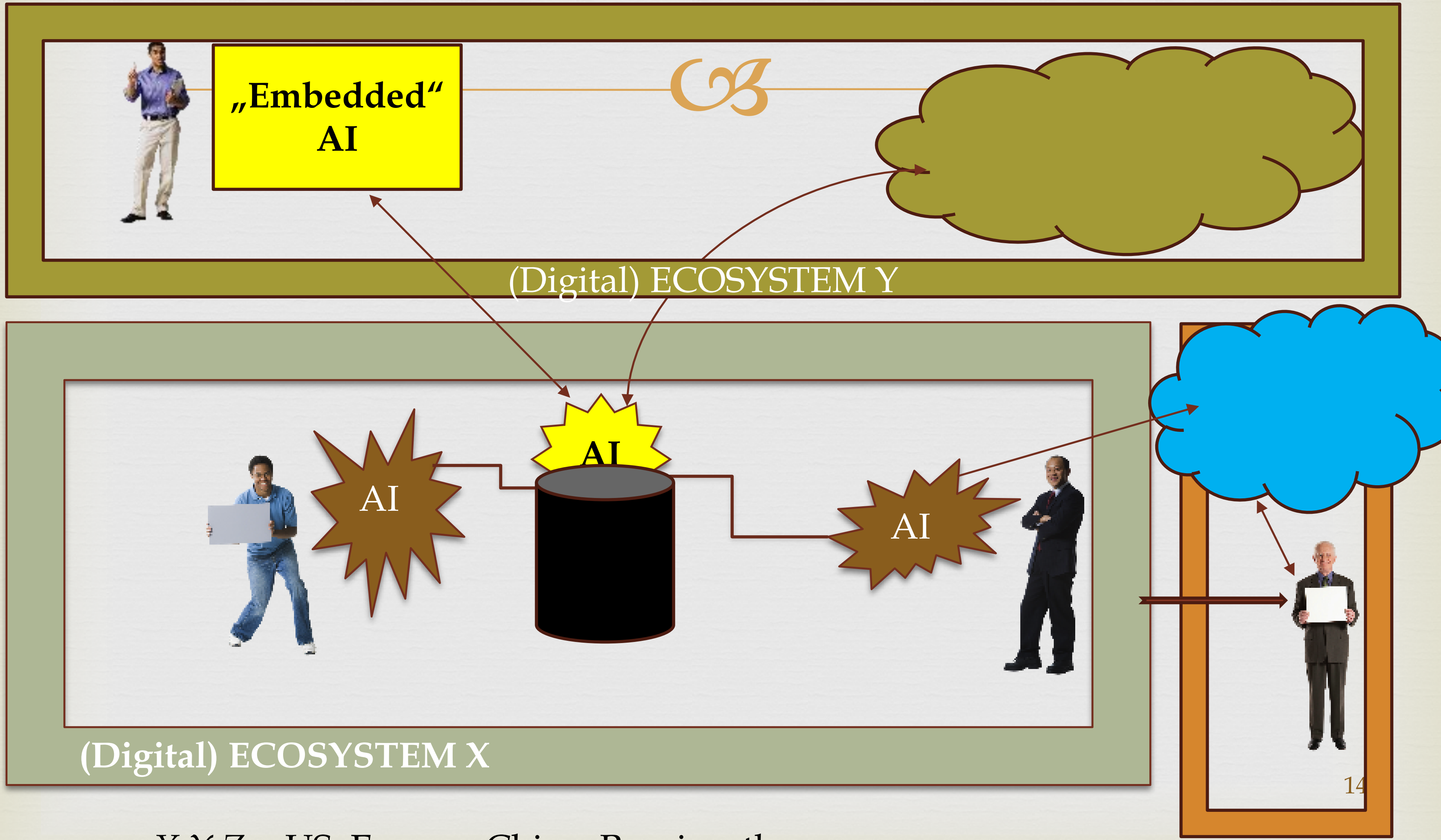
We use *Conceptual clusters* of:

- Bias/*Fairness*/discrimination
- Transparencies/*Explainability*/ intelligibility/interpretability
- Privacy/ responsibility/*Accountability*

and

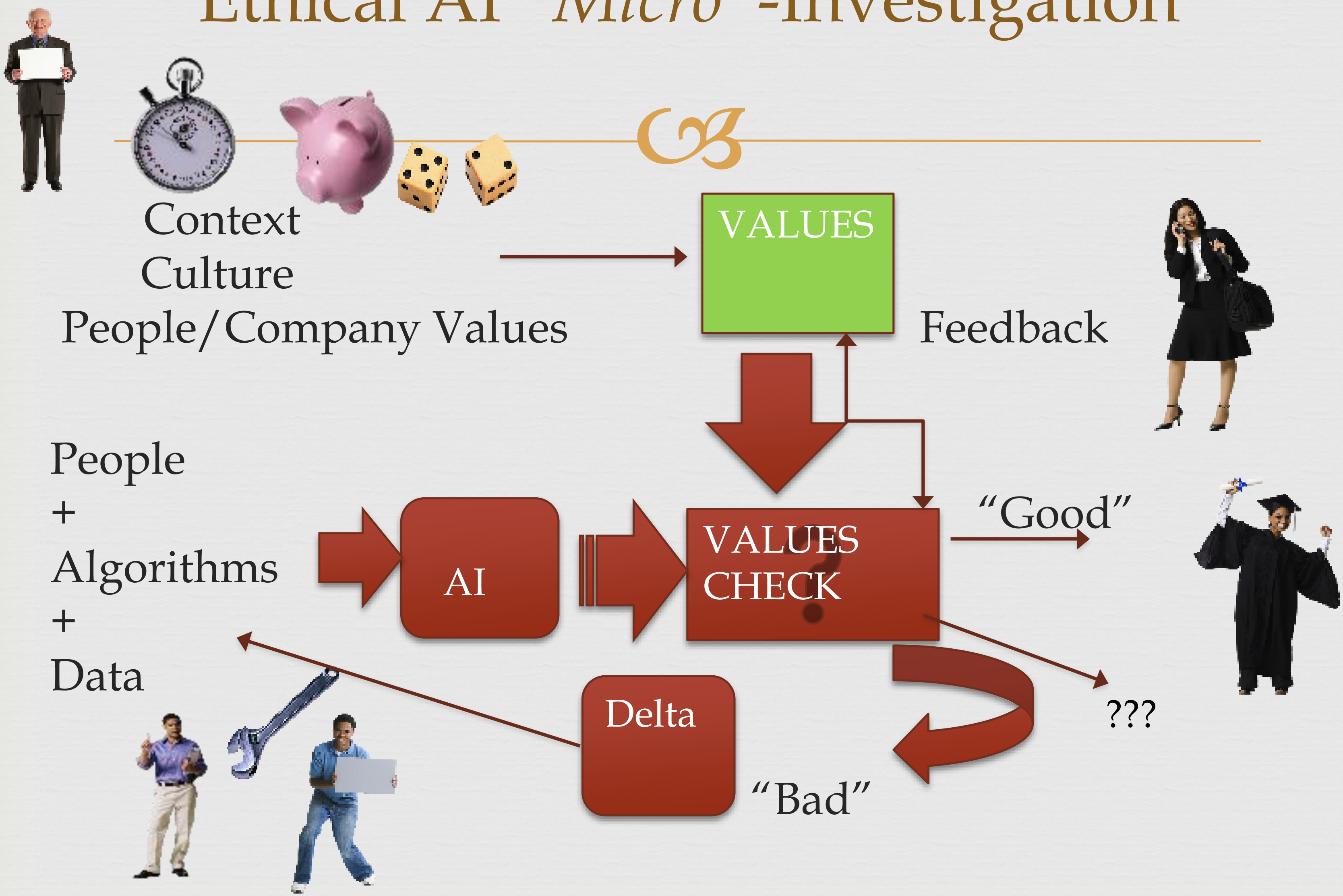
- Safety
- Human-AI
- Other (for example chosen from this list):
  - uphold human rights and values;
  - promote collaboration;
  - acknowledge legal and policy implications;
  - avoid concentrations of power,
  - contemplate implications for employment.

Ethical AI “*Macro*”-Investigation



X,*Y*,*Z* = US, Europe, China, Russia, others…

Ethical AI "*Micro*"-Investigation



Context
Culture
People/Company Values

VALUES

Feedback

People
+
Algorithms
+
Data

AI

VALUES CHECK

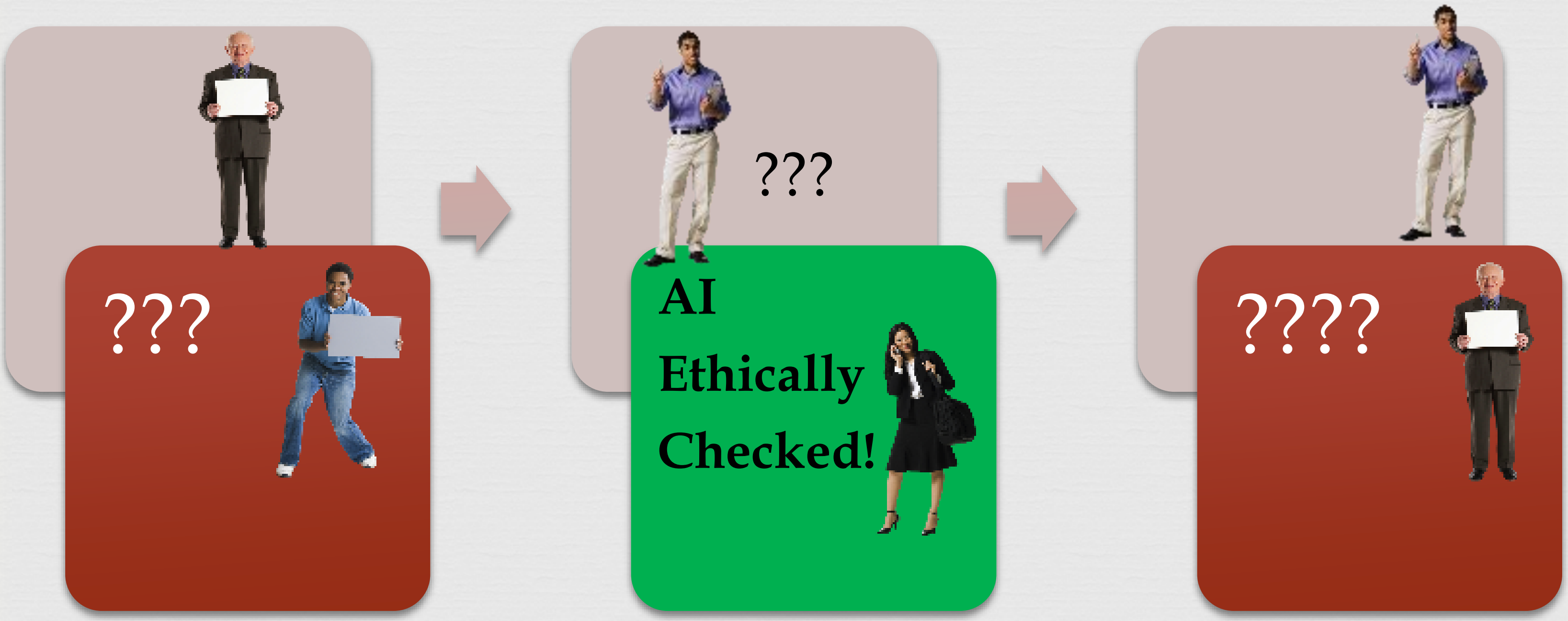
"Good"

???

Delta

"Bad"

Micro-validation does not imply *Macro*-validation



???

???

AI Ethically Checked!

????

Z-Inspection Process

1. **Define an holistic Methodology**

Extend Existing Validation Frameworks and Practices to assess and mitigate risks and undesired "un-ethical side effects", support Ethical best practices.

- Define Scenarios (Data/ Process/ People / Ecosystems),

- Use/ Develop new Tools, Use/ Extend existing Toolkits,

- Use/Define new ML Metrics,

- Define Ethics AI benchmarks

2. Create a Team of inspectors

3. Involve relevant Stakeholders

4. **Apply,Test/Refine the Methodology to Real Use Cases (in different domains)**

5. Manage Risks/ Remedies (when possible)

6. Feedback: Learn from the experience

7. Iterate: Refine Methodology / Develop Tools

We are testing Z-inspection with a use case in Health Care

Assessing

CARDISIO
Your heart counts

"The first highly accurate and non-invasive test to determine a risk factor for coronary heart disease.

Easy to use. Anytime. Anywhere." (*)

(*) Source: https://cardis.io

Preliminaries

- The start up company (with offices in Germany and representatives in the Bay Area, CA) agreed to work with us and work the process together.
- We have NO conflict of interests with them (direct or indirect) nor with tools vendors
- We initially set up a scenario which corresponds to our classification A-/B. i.e. No NDA signed (meaning no access to the ML model, training and test data), but access to all people in the company involved in the AI design/AI deployment/domain experts (e.g. cardiologists)/ business/sales/communications
- They agree to have regular meetings with us to review the process.
- They agree that we publish the result of the assessment.
- They agree to take the results of our assessment into account to improve their AI and their communication to the external world.

Cardisio: Socio-technical scenario: *The Domain*

- *Coronary angiography* is the reference standard for the detection of **stable coronary artery disease** (CAD) at rest (invasive diagnostic 100% accurate)

- **Conventional non-invasive diagnostic** modalities for the detection of stable coronary artery disease (CAD) at rest are subject to significant limitations: low sensitivity, local availability and personal expertise.

- Latest experience demonstrated that **modified vector analysis** possesses the potential to overcome the limitations of conventional diagnostic modalities in the screening of stable CAD.

Source Cardisio

Cardisio Socio-technical scenario: *Cardisiography*

- ***Cardisiography* (CSG)** is a denovo development in the field of applied vectorcardiography (introduced by Sanz, et al. in 1983) using Machine Learning algorithms.

- **Design:** By applying standard electrodes to the chest and connecting them to the Cardisiograph, CSG recording can be achieved.

- **Hypothesis**: „By utilizing computer-assisted analysis of the electrical forces that are generated by the heart by means of a continuous series of vectors, abnormalities resulting from impaired repolarization of the heart due to impaired myocardial perfusion, it is hypothesized that CSG is an user-friendly screening tool for the detection of stable coronary artery disease (CAD)."

Source: Cardisio

Cardisio Socio-technical scenario:
Clinical Screening for Coronary Heart Disease

Classification: Level A- No NDA, IP protected. No Code review

Usage situation: screen people for coronary heart disease as part of the general check-up with a primary care physician.

Design goals:
(1) enable broad screening for coronary heart disease, even if symptom-free;
(2) reduce the number of first-time heart attacks;
(3) avoid unnecessary loss of life and compromised quality of life; and
(4) reduce the financial burden on the health system.
(5) educate people (especially 40 years and older) about a completely new way of screening for coronary heart disease;
(6) constantly monitor the usage of the algorithm and learn from false or dubious diagnoses.

Stakeholders:

Test subject,
Primary care physician,
Cardiologist
Sales agents
Distributors
Resellers

Environment: a society where news about people suffering heart attacks and loss of life are an almost daily occurrence.

Cardisio Socio-technical scenario: *Actions taken based on model`s prediction*

- Patients received “Green” score (*continuous prediction*). Doctor agree. Patient does nothing;
- Patients received “Green” (*continuous prediction*). He and/or Doctor do not trust, asked for further invasive test;
- Patient received “Red” (*continuous prediction*). Doctor agree. Patient does nothing;
- Patient received “Red” (*continuous prediction*). Doctor agree. Patient asks for further invasive test;
- ….

In any of the above cases, Patient and/or Doctor may ask for an *explanation*.

Cardisio Socio-technical scenario: *Go-to-market ecosystem*

- **Go-to-market ecosystem**: Cardisio markets and sells its service directly and via a multi-tiered distribution model.
- Direct sales: Cardisio's network on full-time and contracted sales agent **(largely in Germany, Austria, Switzerland, the Netherlands**) directly approach two types of end users: **Cardiologists,** who will give preferential treatment to individuals whose Cardisiography tested positively; **general care physician**, who are beginning to integrate Cardisiography into their standard tests. People with a positive test result will be referred to a Cardiologist.
- Indirect sales: Cardisio has executed distribution agreements and a joint venture (**covering southern Africa**) with distributors that purchase Cardisiographs and test licenses in bulk, and distribute them to their own regional network of resellers, which in turn target primary care physicians and cardiologists.
- Customer support is conducted centralized by Cardisio via an outsourcing partner.

Source : Cardisio

Cardisio: Socio-technical scenario: *Legal*

- The *Cardisiograph* has CE clearance to be sold as an electronics product in the European Union.
- The company decided that registering it as a medical device was not required. The device itself records and transmits data.
- Medical analysis is being conducted by the Cardisio Cloud algorithm, which has been registered as such with the appropriate EU institution.

Source Cardisio

Cardisio Socio-technical scenario: *Operational model*

Step1. Measurements, Data Collection (Data acquisition, Signal processing)

Step 2 Automated Annotation, feature extraction, statistical pooling, features selection

Step 3. Neural Network classifier training

An ensemble of 25 Feedforward neural networks. Each neural network has two hidden layers of 20 and 22 neurons. Each neural network has an input of 27 features. One output: Cardisio Index (range -1 to 1)

Step 4. Actions taken based on the model´s prediction

Source: Cardisio

Cardisio Socio-technical scenario: Neural Network classifier: *Data*

All clinical data to train and test the Classifier was received from experiments conducted in 3 hospitals in Germany, all of three near to each other (Duisburg area). The data has been supplied to the technical team by Prof. med Gero Tenderich (Cardiologist and shareholder of Cardisio).

The data contains **600 patient records, of which 250 women and 350 man** (all from the 3 hospitals). Due to regulation, no information of the background of the patients is given.

Previously the data sets was under-representing young people and represents mainly older people. With the current data set (600 people) this has been mitigated.

- From April 2017 to February 2019 cardisiographic results were obtained from **546** unselected adult patients **(male: 340, female: 206**) of three centers (Evangelisches Krankenhaus Duisburg-Nord, Herzzentrum Duisburg, St. Bernhard Hospital Kamp-Lintfort) who had **undergone coronary angiography** and then retrospectively correlated blindly by an independent reader to their angiographic findings.

Source Cardisio

Cardisio: Socio-technical scenario Neural Network classifier: *Training and Output*

The net is trained by a back propagation algorithm and is optimized for *Sensitivity, Specificity, Positive predictive value, Negative predictive value, AUC.* With 1.5-weighted sensitivity.

The output of the network is the Cardisio Index (range -1 to 1), a scalar function dependent on the input measurement, classifying impaired myocardial perfusion.

Source: Cardisio

Assessing ***fairness*** (Bias/Discrimination)

"Clarifying what kind of algorithmic "fairness" is most important is an important first step towards deciding if this is achievable by technical means" (*)

Identify Gaps/Mapping conceptual concepts between:

1. *Context-relevant Ethical values,*

↓

2. *Domain-specific metrics,*

↓

3. *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.

Context-relevant Ethical values

No uniform consensus within philosophy on the "*exact*" definition of "fairness". (e.g. *utilitarianism, egalitarianism, minimax*).

Different focus on *individual*, or the *collective*.

Highly dependent on the *context* (Ecosystems)

Navigating disagreements may require *political solutions*.

(*) *Source:* Whittlestone, J et al (2019)

ML and *Fairness* criteria in healthcare

(domain specific)

Distributive justice **(from philosophy and social sciences) options for machine learning**

Possible Mitigation
(*Fairness* criteria)

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/

ML Bias in healthcare

(domain specific)

- **Biases in model design**
  - *Labels bias, Cohort bias*

- **Biases in training data**
  - *Minority bias*
  - *Missing Data bias*
  - *Informativeness bias*
  - *Training-serving skew*

- **Biases in interactions with clinicians** *(domain specific)*
  - *Automation bias*
  - *Feedback Lops*
  - *Dismissal bias*
  - *Allocation discrepancy*

- **Biases in interactions with patients** *(domain specific)*
  - *Privilege bias*
  - *Informed mistrust*
  - *Agency bias*

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/

From Domain Specific to ML metrics

- Different interpretations/definitions of *fairness* pose different requirements and challenges to Machine Learning (metrics) !

Mapping Domain specific "Fairness" to Machine Learning metrics

Several Approaches: **Individual fairness , Group fairness, Calibration, Multiple sensitive attributes, casuality**.(*).
In Models : Adversarial training, constrained optimization. regularization techniques,….(*)

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

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)

ML and *Fairness* criteria in healthcare

(domain specific)

- **Does the Model produces Equal Outcomes?**
  - Do both the protected group and non protected group benefit similarly from the model (**equal benefit**)?
  - Is there any outcome disparity lessened (**equalized outcomes**)?

- **Does the Model produces Equal Performance?**
  - Is the model equally accurate for patients in the protected and non protected groups?
    - 1**. equal sensitivity (equal opportunity**)

      A higher false-positive rate may be harmful leading to unnecessary invasive interventions (angiography
    - **2. equal sensitivity and specificity** (**equalized odds**)

      Lower positive predictive value in the protected group than in the non protected group, may lead to clinicians to consider such predictions less informative for them and act on them less (**alert fatigue**)
    - **3. equal positive predictive value (predictive parity**)

- **Does the Model produces Equal Allocation (demographic parity)?**
  - Are resources proportionally allocated to patients in the protected group?

Source. Alvin Rajkomar et al. Ensuring, Fairness in Machine Learning to Advance Health, Equity, Annals of Internal Medicine (2018). DOI: 10.7326/M18-1990

Trade Offs (Incompatible types of fairness)

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/

Cardisio: Socio-technical scenario Developing an evidence base

When interviewed, Prof. Dr. med Gero Tenderich (heart surgeon and co-founder of Cardisio) confirmed that there are significant differences in the physicality of the human cardiovascular system.

Gero Tenderich said this is also well documented in the medical literature.

Gero Tenderich also said that there is conclusive scientific evidence that the electricity of the human heart does not vary by ethnicity or other qualifiers. This was first published by **Aschoff-Tawara** (1906)

Cardisio: Socio-technical scenario *Discover potential ethical issues*

*Overall, from **an ethical point of view** the chances that more people with an undetected serious CAD problem will be diagnosed in an early stage need to be weighted against the risks and cost of using the CSG app.*

Trust vs. Human Perception

When asked two of the AI developers of Cardisio if they both used Cardisio themselves, they both said yes.

But one said that there were some hesitation and resistance:

"I really don't want to know about it".

Trust is not all. There are human feelings/fear, not necessarily based on technical information based on *fairness*, *accuracy* and *explanability* of the AI….

Assessment of Bias/fairness/discrimination: Remedies

The AI (ML) model is already deployed.

AI is being sold.

- Possible Remedies
  - Monitor the performance of the model and outcomes measurements.
  - Perform formal clinical trial design.
  - Improve the model over time by collecting more representative data (FLAG!) .

Cardisio: Assessing Evidence base Verify Tension: *Accuracy vs. Fairness*

If we consider **Bias/Fairness/Discrimination**, the next step is to decide how deep we want to go.

Assumptions:

1. Significant differences in the physicality of the human cardiovascular system;

2. Electricity of the human heart does not vary by ethnicity or other qualifiers;

3. CSG does measure the electrical forces that are generated by the heart:

All clinical data to train and test the classified come form three hospitals in Germany

We have no access to the Model, the training data and the 27 Features

Cardisio: Deeper Assessment of Bias/fairness/discrimination

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

i) Signing an NDA makes it easier to go deeper.

ii) The alternative is to audit the output only.

Lessons learned so far:

We decided to go for an open development and incremental improvement to establish our process and brand ("*Z Inspected*").

This requires a constant flow of communication and discussion with the company so that we can mutually agree on what to present publically during the assessment process, without harming the company, and without affecting the soundness of the assessment process.

Photo RVZ

Cardisio: Further levels of inspection

- Bias/*Fairness*/discrimination

- Transparencies/*Explainability*/ intelligibility/interpretability

- Privacy/ responsibility/*Accountability*

- Safety

- Human-AI

- Legal

AI Ethical Assessment: Questions, Metrics, Tools, Limitations

- How much of the inspection is questioning, negotiating?
- How much of the inspection can be carried out using software tools? Which tools for what?
- How much of the inspection is simply not possible at present state of affairs?

Which Tools to Use for what?

Open Source Tools
(non-exhaustive list)

Tool	*Purpose*	*Map to Ethical Values*	*Limitations*
AI Fairness 360 (IBM)			
***What-if Tool, Facets, Model and Data Cards* (Google)**			
***Aequita*s *(Univ. Chicago)* https://dsapp.uchicago.edu/projects/aequitas/			
Lime *(Univ. Washington)* https://github.com/marcotcr/lime			
FairML https://github.com/adebayoj/fairml			
SHAP https://github.com/slundberg/shap			
DotEveryone Consequence Scanning Event https://doteveryone.org.uk/project/consequence-scanning/			
Themis https://github.com/LASER-UMASS/Themis	testing *discrimination* (group discrimination and causal discrimination.)		
Mltest **https://github.com/Thenerdstation/mltest**	writing simply ML unit test		
Torchtest https://github.com/suriyadeepan/torchtest	writing test for pytorch-based ML systems		
CleverHans https://github.com/tensorflow/cleverhans	benchmark for ML testing		
FalsifyNN https://github.com/shromonag/FalsifyNN	detects *blind spo*ts or *corner cases* (autonomous driving scenario)		

Word of caution

- Scenarios, parts of the Inspection, and the whole Inspection, can be misused.

"expert´s statements on the technological future, can also be used to legitimize and justify the role of a new, not-yet established technology or application and thus have a strategic role in welcoming the technology and convincing an audience" (*)

- The risk of such a check quickly be obsolete, as the AI system evolves and adapts to changing environments.

- There is a need of a continuous *ethical maintenance.*

- (*) source: Ethical Framework for Designing Autonomous Intelligent Systems. J Leikas et al. J. of Open Innovation, 2019, 5, 1

Possible unwanted *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..

Chances and Risks

The case study shows how important interdisciplinary cooperation is in designing and deploying AI.

There is no perfect solution but chances and risks of new technologies have to be weighted.