Z-inspection

Towards a process to assess Ethical AI

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

October 22**, 2019**

Capco Institute

Frankfurt, Germany

The Ethics of Artificial Intelligence

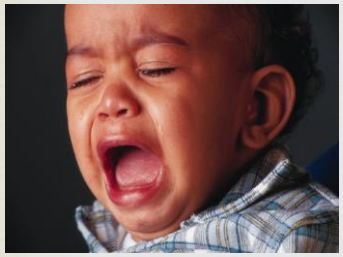
"Who will decide what is the impact of AI on Society?"

The Ethics of Artificial Intelligence

- AI is becoming a sophisticated tool in the hands of a variety of stakeholders, including political leaders.

- Some AI applications may raise new **ethical** and **legal** questions, and in general have a significant impact on **society** (for the good or for the bad or for both).

- **People motivation** plays a key role here.



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?

- This poses an ethical and societal problem.

The Ethics of Artificial Intelligence

- **With AI the important question is how to avoid that it goes out of control, and how to understand how decisions are made and what are the consequences for society at large.**

Policy Makers and AI

*"**Citizens** and **businesses** alike need to be able to **trust** the technology they interact with, and have effective safeguards protecting fundamental rights and freedoms.*

*In order to increase **transparency** and **minimise the risk of bias**, AI systems should be developed and deployed in a manner that allows humans to **understand** the basis of their actions.*

***Explainable AI** is an essential factor in the process of strengthening people's trust in such systems."* (*)

-- Roberto Viola ***Director General of DG CONNECT (Directorate General of Communication Networks, Content and Technology) at the European Commission.***

(*) Source On the Future of AI in Europe. Interview with Roberto Viola, ODBMS Industry Watch, 2018-10-09

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/deployed*, 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.

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

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 have been designed and result in production in *Fostering Ethical values and Ethical actions (FE)*

There is no negative score.

Goal: reward and stimulate new kinds of Ethical innovation.

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

Core Ethical Principle: *Beneficence. ("well-being", "common good"…)*
The Problem: Debatable even in the Western World…

Closing the Gap

"Most of the principles proposed for AI ethics are not specific enough to be action-guiding."

"The real challenge is recognizing and navigating the tension between principles that will arise in practice."

" Putting principles into practice and resolving tensions will require us to identify the underlying assumptions and fill knowledge gaps around technological capabilities, the impact of technology on society and public opinion". (*)

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

Formulating universal AI principles?

" Given different cultural traditions, philosophers could spend many lifetimes debating a set of universal AI principles"

-- John Thornhill. (*)

(*) *Formulating AI values is hard when human fail to agree*, John Thornhill, Financial Times, July 22, 2019

What Practitioners Need

Need for ethical frameworks and case studies

- " Several interviewees suggested it would be helpful to have access to domain-specific resources, such as **ethical frameworks and case studies**, to guide their teams´ ongoing efforts around **fairness**"
- 55% of survey respondents indicated that having access to such resources would be at least "Very" useful (*)

- (*) **Based on 35 semi-structured interviews and an anonymous survey of 267 ML practitioners in USA**. Source: Improving Fairness in Machine Learning Systems: What Practitioners Need? K. Holstein et al. CHI 2019; May 4-0, 2019

Need for More Holistic Auditing Methods

- "Interviewers working on applications involving richer, complex interaction between the user and the system bought up needs for more ***holistic***, system-level **auditing methods**." (*)

- (*) source: Improving Fairness in Machine Learning Systems: What Practitioners Need? K. Holstein et al. CHI 2019; May 4-0, 2019

Need for Metrics, Processes and Tools

- "Given that *fairness* can be highly context and application dependent, there is an **urgent need for domain-specific educational resources, metrics, processes and tools** to help practitioners navigate the unique challenges that can arise in their specific application domains" (*)

- (*) source: Improving Fairness in Machine Learning Systems: What Practitioners Need? K. Holstein et al. CHI 2019; May 4-0, 2019

Z-inspection

A process to assess Ethical AI

Photo: RVZ

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

Why?

- *Who* requested the inspection?
  - Recommended vs required (mandatory inspection)
- *Why*?
- For *whom* is the inspection relevant?
- How to use the results of the Inspection?
  - Verification, Certification, Sanctions (if illegal),
  - Share (Public), Keep Private (*Why keeping it private?)*

What do we wish to investigate?

- AI is not in isolation.

It is part of one or more (digital) ecosystems
It is part of Processes, Products, Services, etc.
It is related to People, Data, Ethical Values.

AI is not a single element
Made up of various components, e.g. deep neural network architectures: neural networks building blocks.

Pre-conditions

1. Agreement on *Context-specific ethical values*

2. Agreement on the *Areas of Investigation*

Z-Inspection: *Areas of investigations*

We use *Conceptual clusters* of:

Bias /*Fairness*/ Discrimination

Transparencies /*Explainability*/ Intelligibility/Interpretability

Privacy/ Responsibility/*Accountability*

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.

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.

AI, Ethics, Democracy

Do we want to assess if the *Ecosystem(s)* where the AI has been designed/produced/used is *Democratic*?

Is it Ethical?

Is it part of an AI Ethical Inspection or not?

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);

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)*

Focus of Z-inspection

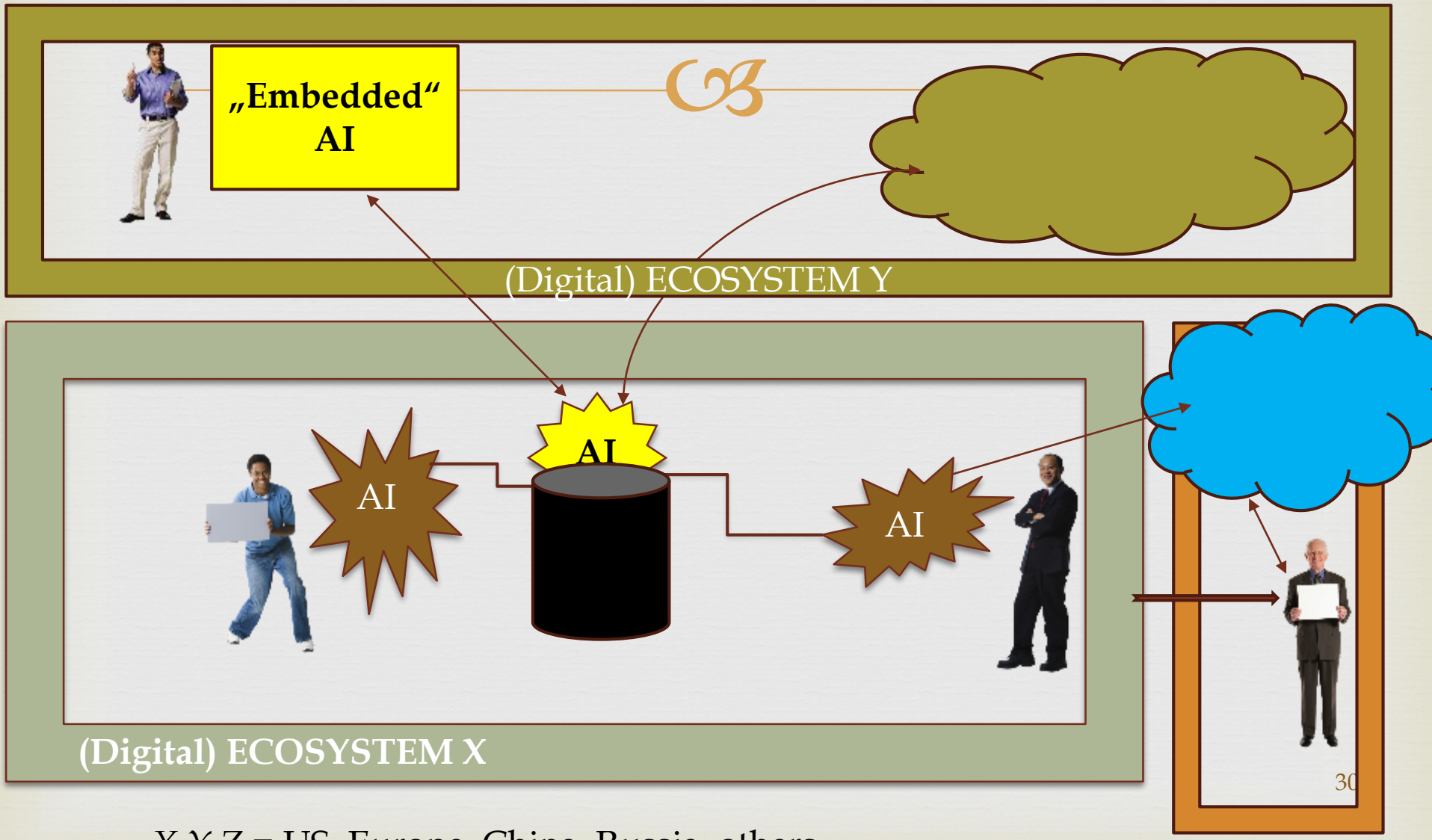
- Ethical
- Technical
- Legal

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.

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"

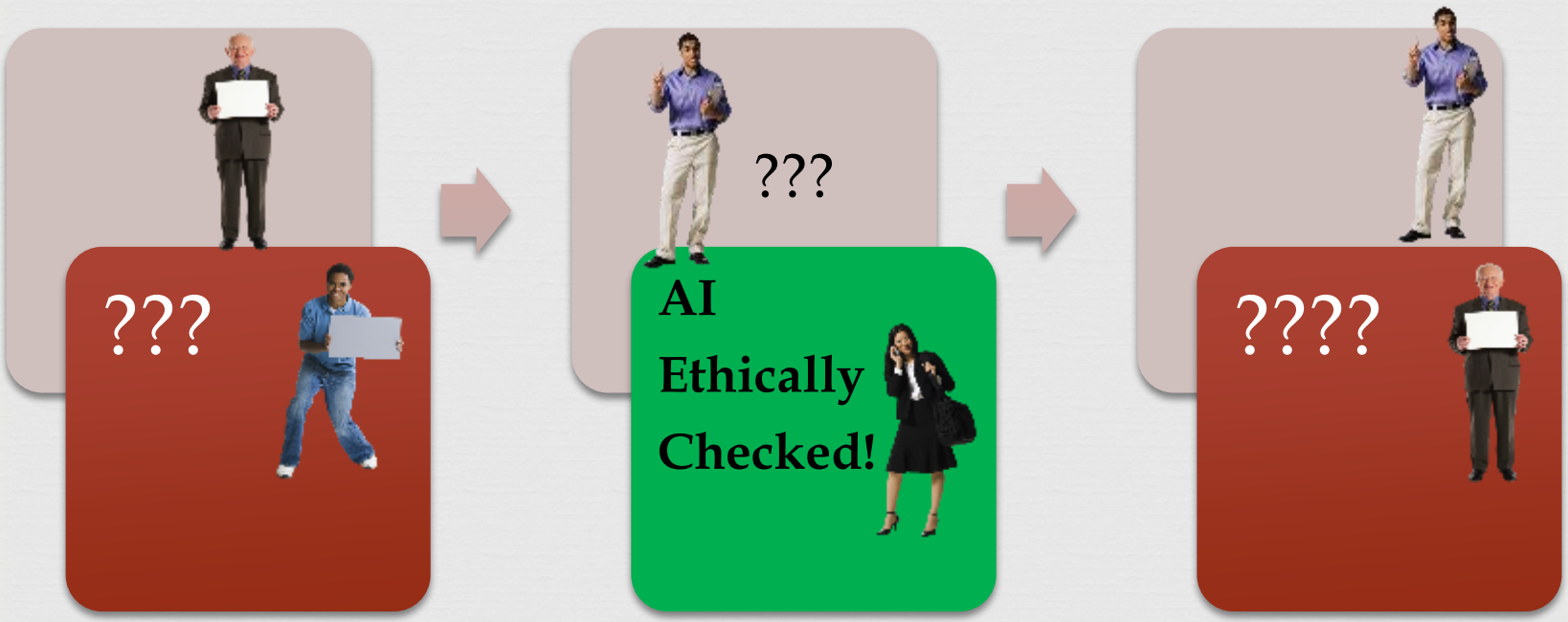
???

Delta

"Bad"



Micro-validation does not imply *Macro*-validation



Discover potential ethical issues

- We use Socio-technical scenarios to describe the *aim of the s*ystem, the *actors and their expectations*, the *goals of actors´ action*, the *technology* and the *context*. (*)

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

Concept Building

As suggested by Whittlestone, J et al (2019), we do *Concept Building*:

- *Mapping and clarifying ambiguities*
- *Bridging disciplines, sectors, publics and cultures*
- *Building consensus and managing disagreements*

Developing an evidence base

- 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)

Identify *Tensions*

Identifying Tensions
(*different ways in which values can be in conflict*)

Accuracy vs. fairness

e.g. *An algorithm which is most accurate on average may systematically discriminate against a specific minority.*

Using algorithms to make decisions and predictions more accurate versus ensuring fair and equal treatment

Accuracy vs explainability

Accurate algorithm (e.g. deep learning) but not explainable (degree of 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**

Source: Whittlestone, J et al (2019)

Address, Resolve *Tensions*

- ***Resolving Tensions*** (Trade-offs)
  - ***True ethical dilemma*** - the conflict is inherent in the very nature of the values in question and hence cannot be avoided by clever practical solutions.
  - ***Dilemma in practice***- the tension exists not inherently, but due to our current technological capabilities and constraints, including the time and resources we have available for finding a solution.
  - ***False dilemma*** - situations where there exists a third set of options beyond having to choose between two important values.

- ***Trade-offs***: How should trade-off be made?

Source: Whittlestone, J et al (2019)

List of potential ethical issues

- The outcome of the analysis is a list of potential ethical issues, which need to be further deliberated when assessing the design and the system`s goal and outcomes. (*)

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

Z-inspection verification concepts (subset)

Verify Purpose

Questioning the AI Design

Verify Hyperparameters

Verify How Learning is done

Verify Source(s) of Learning

Verify Feature engineering

Verify Interpretability

Verify Production readiness

Verify Dynamic model calibration

Feedback

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
- 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
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 and interpreted by an expert

Source: Cardisio

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

- Patients received "Green" score (*continuous prediction: dark to light Green*). Doctor agree. Patient does nothing;
- Patients received "Green" (*continuous prediction*). Patient and/or Doctor do not trust, asked for further invasive test;
- Patient received "Red" (*continuous prediction: dark to light Red*). 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
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.*

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

Diagnostic Trust and Competence – ethical issues:

- When CSG is being used in screening un-symptomatic patients who are "*notified*" by Cardisio with a "minor" CAD problem that might not impact their lives, **they might get worried- change their lifestyles after the *notification* even though this would not be necessary**

- If due to the CSG test more patients with minor CAD problems are being "notified" and sent to cardiologists, **this might result in significant increase of health care costs, due to further diagnostics tests.**

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

Diagnostic Trust and Competence – ethical issues:

- Using a black-box algorithm **might impair the trust of the doctor** in the diagnostic app, especially if the functioning of the app / algorithm has not been verified by independent studies.
- Using an AI assisted diagnostic **app could in the long-term impair the diagnostic competence of the medical personal and also the quality of the diagnostic process** when more "physician assistance" instead of medical doctors do the diagnostic "ground work".
- **The doctor's diagnostic decision might become biased** by the assumed "competence" of AI – especially when the doctor's and the AI's diagnosis differ.
- **How high is the risk that an application /diagnostic error** happens with the traditional diagnostic instruments compared to using the CSG app?

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

Safety/ Use of Data

- Will the CSG app patient data stay with the medical doctor and be linked to the patients records?
- How secure is the Cloud data?

Transparencies/Explainability/ Intelligibility/ Interpretability

- Which risk factors (features) contribute most to the result of the classification?

Z-inspection: 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**

What if the Z-inspection happens to be false or inaccurate?

- There is a danger that a *false* or *inaccurate* inspection will create natural skepticism by the recipient, or even harm them and, eventually, backfire on the inspection method.

- This is a well-known problem for all quality processes. It could be alleviated by an open development and incremental improvement to establish a process and brand (like "*Z Inspected*").

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.

assessment process.

Photo RVZ

"Z Inspected": *Certify AI?*

As part of the output of the Z-Inspection perhaps we can "*certify*" AIs by the number of testing with synthetics data sets and extreme scenario they went through- before allowing AIs to be deployed (similar to what happens to airplane pilots).

Somebody would need to define when *good is enough*. And this may be tricky…

How often AI should be inspected?

- Need to define a set of *checkpoints* that need to be monitored over time
- For *minimal* inspection and *full* inspection.
- Regularly monitor and inspect as part of an ongoing *ethical maintenance*.
- How to cope with *changes over time* (Ecosystems, Ethical values, technological progress, research results, politics, etc.)

Responsibility

AI system designers, their managers do have ethical responsibilities.

and

Other stakeholders (e.g. policy makers, politicians, opinion leaders, educators) do have ethical responsibilities.

What about Citizens?

What is the implication for them of the AI Ethical Inspection?

Shall we involve them as well? How?

e.g. consultations and public deliberations

(see *Democracy*)

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..

Approaching Ethical Boundaries

"But if we just let machines learn ethics by observing and emulating us, they will learn to do lots of unethical things.

So maybe AI will force us to confront what we really mean by ethics before we can decide how we want AIs to be ethical." *(*)*

--Pedro Domingos (*Professor at University of Washington*)

- (*) Source: **On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos**, ODBMS Industry Watch, June 18, 2018

Acknowledgements

Many thanks to

Kathy Baxter, Jörg Besier, Stefano Bertolo, Vint Cerf, Virginia Dignum, Yvonne Hofstetter, Alan Kay, Graham Kemp, Stephen Kwan, Abhijit Ogale, Jeffrey S. Saltz, Miroslaw Staron, Dragutin Petkovic, Michael Puntschuh, Lucy Suchman and Clemens Szyperski

for proving valuable comments and feedback.