

# 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, Jesmin Jahan Tithi (\*)

Frankfurt Big Data Lab

[www.bigdata.uni-frankfurt.de](http://www.bigdata.uni-frankfurt.de)

(\*) Intel Labs

**CSIG TALK** October 21 , 2019

© 2019 by Roberto V. Zicari and his colleagues

The content of this presentation is open access distributed under the terms and conditions of the

Creative Commons (**Attribution-NonCommercial-ShareAlike**  
CC BY-NC-SA) license (<https://creativecommons.org/licenses/by-nc-sa/4.0/>)

# The Ethics of Artificial Intelligence

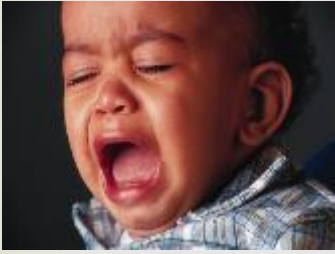


*Who will decide what is the impact of AI on  
Society?*

# 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.

# Another kind of Harm



## *"Big Nudging"*

*He who has large amounts of data can manipulate people in subtle ways.*

*But even benevolent decision-makers may do more wrong than right. (\*)*

(\*) Source: *Will Democracy Survive Big Data and Artificial Intelligence?*. Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A.. (2017). *Scientific American* (February 25, 2017).

# 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

# Mindful Use of AI



*We are all responsible.*

*The individual and collective  
conscience is the existential place  
where the most significant things  
happen.*

# 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 an AI Ethical 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.

# 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



# 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?*)

# The Politics of AI *Ecosystems*



- ∞ The Rise of (Digital) Ecosystems paving the way to disruption.<sup>(\*)</sup>
- ∞ 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.**

<sup>(\*)</sup> Source: Digital Hospitality, Metro AG-personal communication.

# What do we wish to investigate?



∞ AI is not a single element

∞ 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.**

# 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?



# Z-inspection: Pre-conditions



1. Agreement on *Context-specific ethical values*
2. Agreement on the *Areas of Investigation*

# 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 propertyDefine the implications if any of the above conditions are not satisfied.

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

# Focus of the AI Ethics 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.

# Z-inspection: *Areas of investigations*



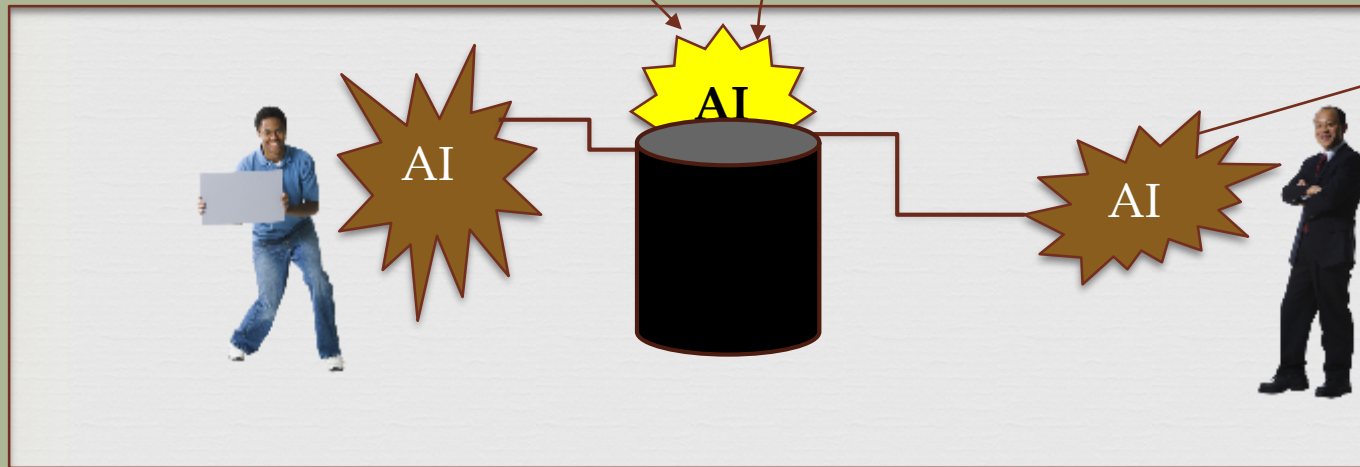
We use *Conceptual clusters* of:

- 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.

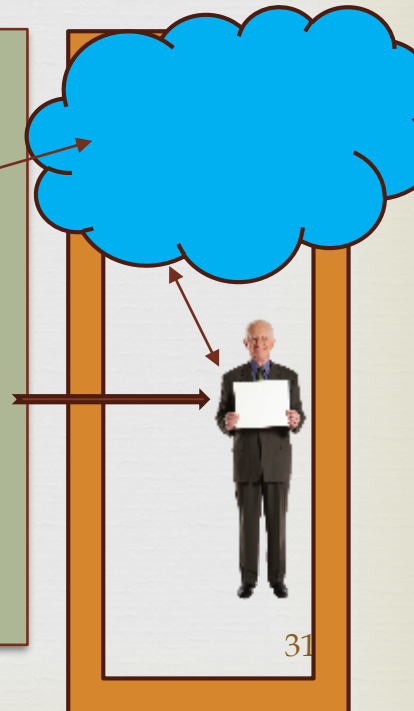
# Macro vs Micro Investigation



# Ethical AI "Macro"-Investigation



(Digital) ECOSYSTEM X



X, Y, Z = US, Europe, China, Russia, others...

# Ethical AI "Micro"-Investigation



Context  
Culture

People/Company Values



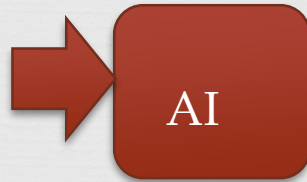
Feedback



People

+  
Algorithms

+  
Data



"Good"



???

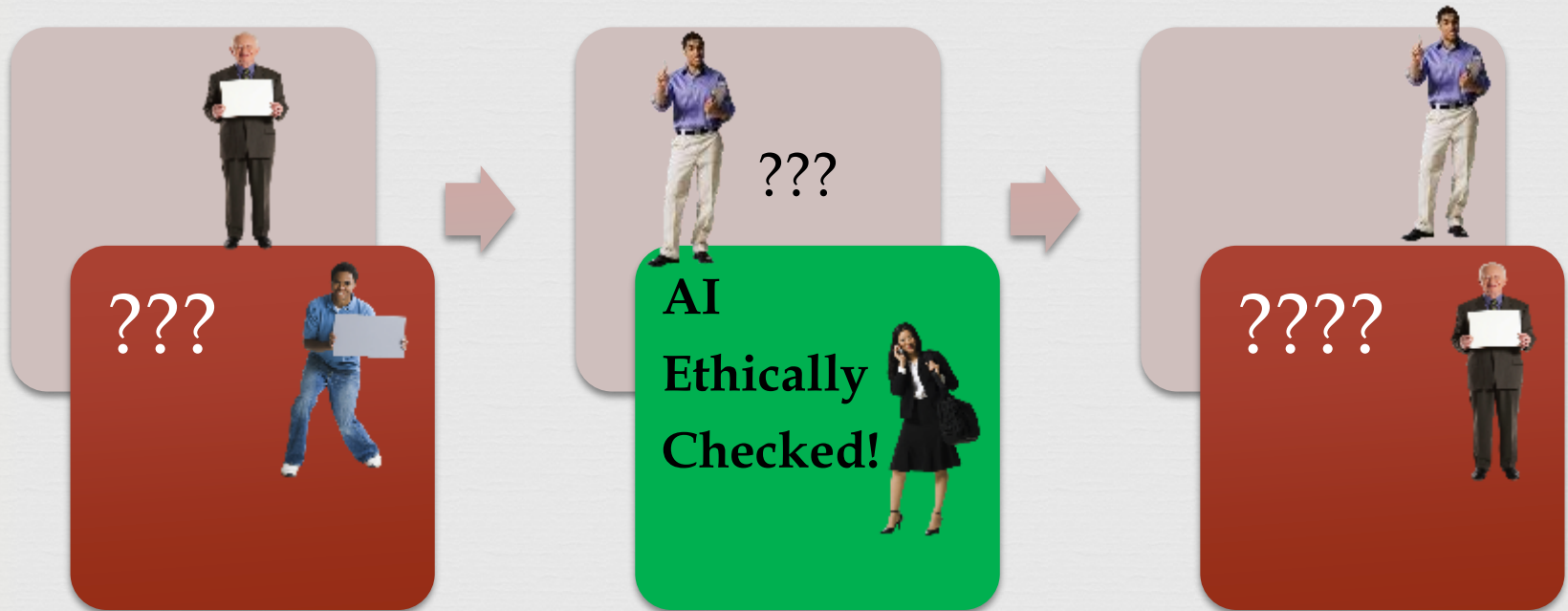


"Bad"





# Micro-validation does not imply Macro-validation



# Z-inspection Methodology



*Photo RVZ*

# Discover potential ethical issues



- ❧ We use **Socio-technical scenarios** to describe the *aim of the system, the actors and their expectations, the goals of actors' action, the technology and the context.* (\*)
  - ❧ What kind of **ethical challenges** the deployment of the AI in the **life of people** raises;
  - ❧ Which **ethical principles** are appropriate to follows;
  - ❧ What kind of **context-specific values and design principles** should be embedded in the design outcomes.
- ❧ We mark possible ethical issues as **FLAGS!**
- ❧ **Socio-technical scenarios and the list of FLAGS! are constantly revised and updated.**

❧ (\*) 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*), e.g.

❧ **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** *e.g 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**

# 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



# Definition of the Inspection Methodology



- ❧ Bottom-up (from Micro to Macro Inspection)
- ❧ Top Down (from Macro to Micro Inspection)
- ❧ Inside-Out (horizontal inspection via layers)
- ❧ Mix : Inside Out, Bottom Up and Top Down

# How to start



- ❧ One possible strategy is start with a *Micro-Investigation* and then if needed progressively extend it in an incremental fashion to include a *Macro-Investigation* (using an *Inside-Out Methodology*)

# Layer of Inside Out



Data/Process/People

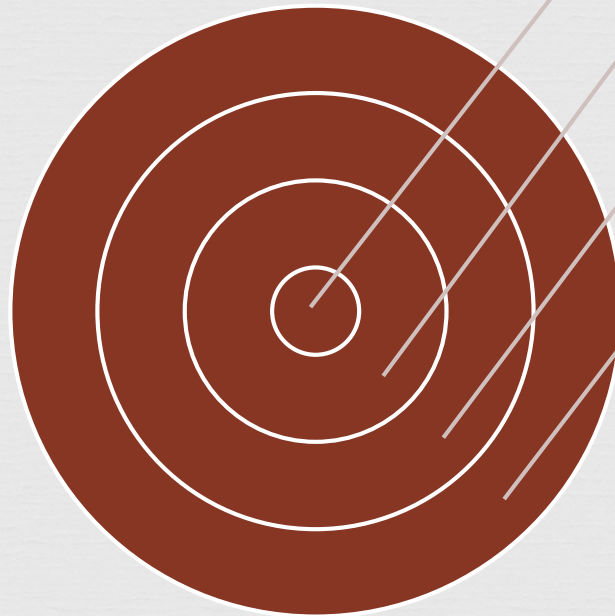
Data/Process/People

AI

Data/  
Process/People

Data/Process/People

# Iterative Inside Out Approach



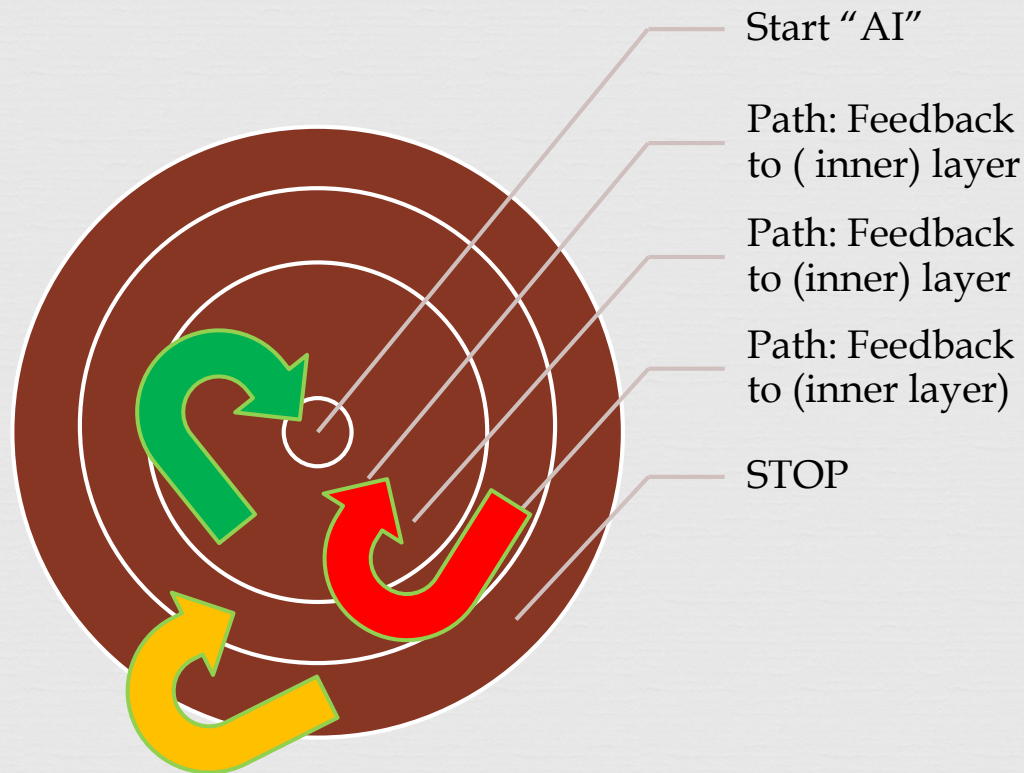
Start with AI. Iterate 5 phases: Explainability, Fairness, Safety, Human-AI, Liability

Each iteration corresponds to a *layer* in an *inside-out methodology*

Augment Explainability++, Fairness++, Safety++, Human-AI++, Liability++

Iterate taking into account the big picture(Macro/Ecosystems)

# Interactive Inside Out Approach Paths and Feedback mechanism



# What is a Path?



- ❧ A *path* describes the dynamic of the inspection
- ❧ It is different case by case
- ❧ By following Paths the inspection can then be traced and reproduced
- ❧ Parts of a Path can be executed by different teams of inspectors with special expertise.

## Example

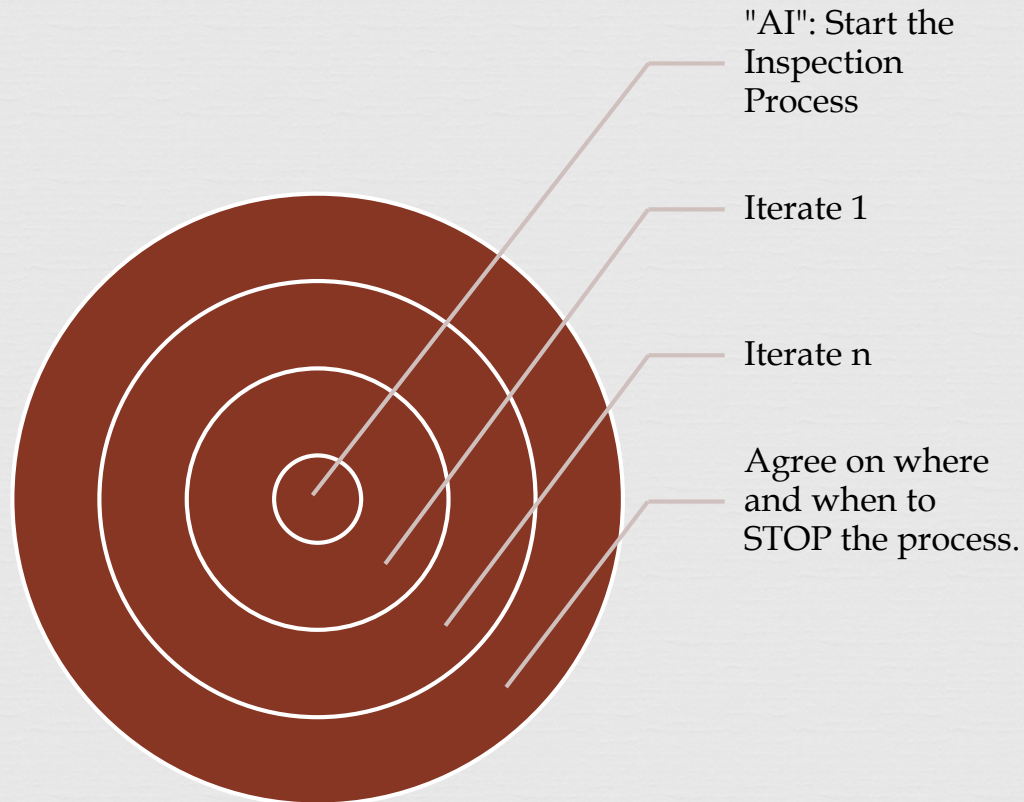
**Path:** from **Fairness**: *training data not trusted, Negative legacy, Labels unbiased (Human raters)* TO **Security** → **Feedback** To **Fairness** TO **Explainability**

# Looking for Paths



- ⌘ Like water finds its way (case by case)
- ⌘ One can start with a predefined set of paths and then follow the flows
- ⌘ Or just start random
- ⌘ Discover the missing parts (what has not been done)

# Agree on when and where to STOP the inspection





# 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



*“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*



- ❧ We conducted a number of interviews with key people from Cardisio (Business, Communication, Domain experts, ML-software developers) to define a socio-technical scenario and a medical evidence base.
- ❧ The resulting socio-technical scenario has been preliminary discussed by our team.
- ❧ We have in our team members with expertise in Ethics, Moral values, Technology (ML, Big Data), Business, Health care, PR/Communication and Marketing.

# 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.

# 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).**”

# Cardisio: Socio-technical scenario

## *Operational model*



**Step 1. 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 and discussed with the person.**

# Cardisio: Socio-technical scenario

## *Actions taken based on model`s prediction*



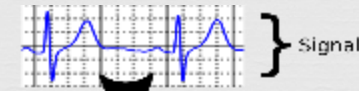
- ❧ Patients received “Green” score (*continuous prediction: dark to light Green*). Doctor agree. Patient does nothing;
- ❧ Patients received “Green” (*continuous prediction*). Patient and/or Doctor do not trust, asked for further invasive test;
- ❧ Patient received “Red” (*continuous prediction: dark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



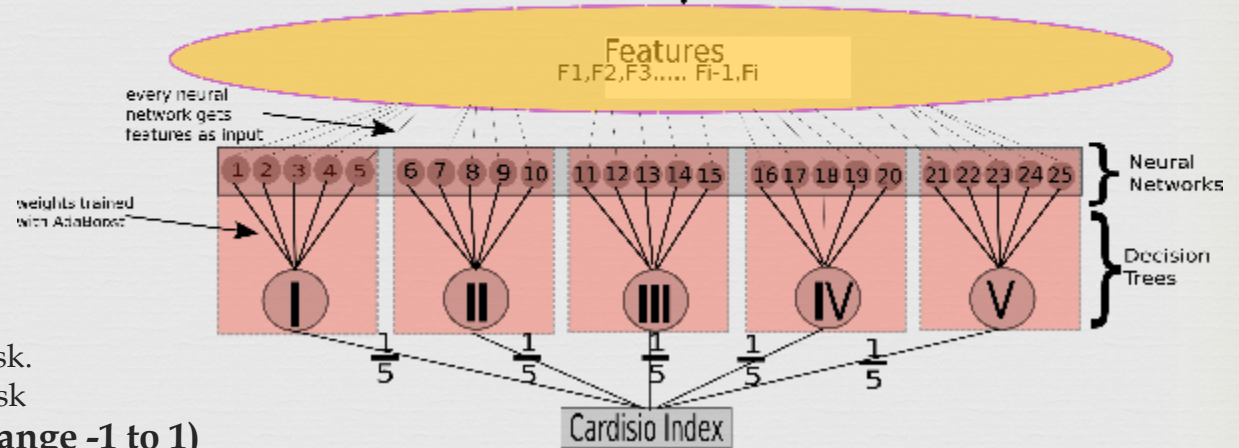
# Cardisio: Socio-technical scenario

## Neural Network classifier



Signal processing

A Neural Network classifier (supervised learning)



Two labels used

Yes-coronary heart disease risk.

NO-coronary heart disease risk

Output: **Cardisio Index (range -1 to 1)**

**An ensemble of 25 Feedforward neural networks.** Each neural network has two hidden layers of 20 and 22 neurons. Each has an input of 27 features. One output.

Selected 27 features, out of 2,600 features calculated (including separation, filtering, correlation). The 27 selected features now do not contain personal information, except for the feature sex. In previous version of the system personal info were used.

# Cardisio: Socio-technical scenario

## *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) **FLAG!****, a scalar function dependent on the input measurement, classifying impaired myocardial perfusion.

Source: Cardisio

 **A FLAG!** identifies potential critical issues.

# Cardisio: Socio-technical scenario

## *Training and Test Data*



All clinical data to train and test the Classifier was received from 3 hospitals in Germany, all of them near to each other (Duisburg area). **FLAG!**

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.

# Cardisio: Socio-technical scenario

## *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.

# Cardisio: Socio-technical scenario

*Legal*



- ☞ The algorithm (Cloud service) has been approved as a Class 1 medical device in the EU.

Source: Caridisio

# 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*



### *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*



### *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?

# Cardisio: Identify and Verify Tension



## **Verify Tension: *Accuracy vs. Fairness***

- Need to Develop a sound (medical) evidence base
- Decide how deep we want to go with the investigation.

# Reflection Moment



At this point we re-assessed our team, and we realized that having an *independent medical expert/ cardiologist* in the team would improve our inspection process for this use case and help us assessing the relevant medical *evidence base*

# 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*”).

# Setting the Boundaries of Ecosystems and Choosing Context-related Ethics



For our use case, if restrict our scope to *Western* clinical medical ethics, we have four classical principles of (\*)

- ❧ Justice
- ❧ Autonomy
- ❧ Beneficence
- ❧ Nonmaleficence

Where “*Western*” define a set of implicit *ecosystems*...

(\*) Source. Alvin Rajkomar et al. (2018)



# 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.

# *Fairness: Different definitions*



For our use case, suppose we are concerned with whether the cardisio-AI used to make healthcare decision is *fair* to all patients.

*Different definitions, e.g.*

- ❧ *Egalitarian concept of fairness: assess if the algorithm produces equal outcomes for all users (or all “relevant” subgroups)*
- ❧ *Minimax concept of fairness: ensure the algorithm results in the best outcomes for the worst off user group.*

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

# Context-relevant Ethical values: Fairness



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)



# Choosing *Fairness* criteria (domain specific)



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

**Possible Mitigation**  
(*Fairness* criteria)



*Equal Outcomes*  
*Equal Performance*  
*Equal Allocation*

BUT, could we use other fairness criteria?

e.g **Kaldor-Hicks criterion**

*This criterion is used in welfare economics and managerial economics to argue that it is justifiable for society as a whole to make some worse off if this means a greater gain for others.*

(\*) 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/>

# Applying ML and *Fairness* criteria in healthcare (domain specific)



Do we have protected groups? If yes:

## ☞ 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?

# Known Trade Offs

## (Incompatible types of fairness)



### **Known Trade Offs (Incompatible types of fairness)**

Equal positive and negative predictive value vs. equalized odds

Equalized odds vs. equal allocation

Equal allocation vs. equal positive and negative prediction value

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

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

Source. Alvin Rajkomar et al. Ensuring, Fairness in Machine Learning to Advance Health, Equity, Annals of Internal Medicine (2018).

DOI: 10.7326/M18-1990

Link: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594166/>

# Example: Fairness / Bias



- ❧ AI Technically correct does not necessarily mean Ethical AI
- ❧ E.g. A dataset which is “unbiased” (in the **statistical** sense) may nonetheless encode common biases (in the **social sense**) towards certain individuals or social groups (\*)

**Q. Is it “fair” to use a feature in a given decision making scenario?**  
*Fairness Disagreements*

(\*) source: Whittlestone J (2019)

# 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) !
- ❧ Engineers like to measure.
- ❧ But, can we really *measure* what “fairness” is for an AI-based decision ?

# Mapping Domain specific “Fairness” to Machine Learning metrics

Several Approaches: Individual fairness, Group fairness, Calibration, Multiple sensitive attributes, causality.\*.

In Models : Adversarial training, constrained optimization. regularization techniques,...(\*)

## Resulting Metrics

## Formal “non-discrimination” criteria

- |   |              |
|---|--------------|
| Statistical parity  | Independence |
| Demographic parity (DemParity)<br>(average prediction for each group should be equal) | Independence |
| Equal coverage  | Separation   |
| No loss benefits  |              |
| Accurate coverage   |              |
| No worse off  |              |
| Equal of opportunity (EqOpt)<br>(comparing the false positive rate from each group)   | Separation   |
| Equality of odds<br>(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)



# Cardisio: *Fairness/ Remedies*



The AI (ML) model is already deployed.

AI is being sold.

Current Remedies in place:

- ❧ 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!**)

# *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.

# 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 AI Explainability 360 Open Source Toolkit (IBM)**

**What-if Tool, Facets, Model and Data Cards (Google)**

**Aequitas** (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** testing *discrimination* (group discrimination and causal discrimination.)

<https://github.com/LASER-UMASS/Themis>

**Mltest** writing simply ML unit test

<https://github.com/TheNerdStation/mltest>

**Torchtest** writing test for pytorch-based ML systems

<https://github.com/suriyadeepan/torchtest>

**CleverHans** benchmark for ML testing

<https://github.com/tensorflow/cleverhans>

**FalsifyNN** detects *blind spots* or *corner cases* (autonomous driving scenario)

<https://github.com/shromonag/FalsifyNN>

# Collaborations



- ❧ We are working together with Fiddler Labs and plan to use their beta version of the *Fiddler AI engine* (proprietary software) for assessing the *explainability* of cardisio.
- ❧ **The goal is to Understand the AI predictions** and bring a human in the loop to audit the predictions and ensure they are correct.
- ❧ GO: We have no conflict of interests with Fiddler Labs

# 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**

# 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, Mirosław Staron, Dragutin Petkovic, Michael  
Puntschuh, Lucy Suchman, Clemens Szyperski  
and Shizuka Uchida

for proving valuable comments and feedback.



# Open Questions



# Levels of Z-inspection



How to define what is a *minimal-but sufficient*-level of inspection?

Need to define what are the *sufficient* conditions

Need to define what are the *necessary* conditions

# Who is qualified to conduct a Z-inspection?



- ❧ Manual Inspection (meaning conducted by human)
- ❧ Who validates and how to validate the Ethical values of the *controller*?

## “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.)

# AI and The Paradox of Transparency



- ❧ I do not mean *cognitive biases*...
- ❧ I mean, if we really insist on *AI Transparency*, perhaps this would force us to reveal our real *motives*...
- ❧ But, we do not always wish to make our motives visible to the outside world, e.g. we do not wish transparency....
- ❧ But with no transparency, there is a lack of trust.

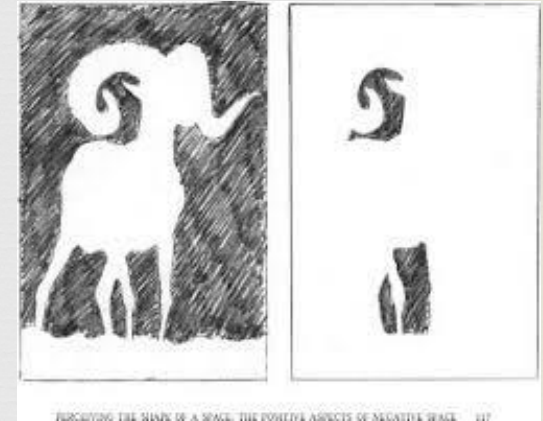
# Negative spaces



- Two terms traditionally used in art (\*):
    - Negative spaces
    - Positive forms
- Skill: the perception of negative spaces

Is this useful skill for an AI Ethical Inspection?

If we look at **bias** as a *negative space*  
then **discrimination** may becomes visible?



(\*) Source: The New Drawing on the Right Side of the Brain. Betty Edwards, 1999, Tarder Putman.

# *AutoML for Ethics?*



- ❧ *Can AI validate the Ethical level of another AI (sort of an AutoML for Ethics)?*
- ❧ *Can we apply reinforcement learning to train the controller of what is Ethical and what is not Ethical ? (sort of using policy gradient to define Ethical rewards. E.g. The controller will give higher probabilities to architectures that receive high Ethical accuracy)*
- ❧ *If this is possible? If yes then who validates the AI controller ?*



# Determining whether a system is making ethical decisions or not



- ❧ “As a layperson looking at this particular field of ethical systems, I see some parallels between determining whether a system has intelligence and whether a system is making ethical decisions or not. In both cases, we are faced with a **kind of Turing test scenario** where we find it difficult to articulate what we mean by intelligence or ethics, and can only probe a system in a Turing test manner to determine that it is indistinguishable from a model human being.
- ❧ The trouble with this approach though is that we are assuming that if the system passes the test, it shares the same or similar internal representations as the human tester, and it is likely that its intelligence or ethical behavior generalizes well to new situations. We do the same to assess whether another human is ethical or not.
- ❧ **This is a great difficulty, because we currently know that our artificial ML systems learn and generalize differently than humans do, so this kind of approach is unlikely to guarantee generally intelligent or ethical behavior.**
- ❧ **I think the best we can currently do is to explicitly engineer/bound and rigorously test the system against a battery of diverse scenarios to check its decisions and reduce the likelihood of undesirable behavior.**
- ❧ **The number of tests needs to be large and include long-tail scenarios because deep learning systems don't have as large a generalization horizon as human learning, as evidenced by their need of a mountain of training data. “**

--- [Abhijit Ogale](#)

*Disclaimer: personal viewpoint as a ML researcher, not in his role at Waymo.*

# AI Team Diversity



“If AI/ML teams are too homogeneous, the likelihood of group-think and one-dimensional perspectives rises – thereby increasing the risk of leaving the whole AI/ML project vulnerable to **inherent biases and unwanted discrimination.**”

-- Nicolai Pogadl (\*)

*How to assess if and when the team is biased and what are the implications?*

(\*) Source: personal communication.

# Is *trustworthy AI* the right approach for assessing AI?



- ❧ Trust is not equal to Ethical
- ❧ Trust is not equal to Technically Correctness
- ❧ Trust is not equal to Compliance to Law

In practice The key question for TRUST is:

will **YOU** use it?

# Unduly harm



## ❧ How can we ensure any such inspection process does not unduly harm small firms at the benefit of large firms?

It is already a critical situation in that large firms often have all the data. If data is key for developing innovative algorithms, you can think of them as the "*means of production*". So the data = "*means of production*" belong to a few, any smaller firms are left out.

But this critical situation could be compounded if an expensive and time consuming ethics process was mandated. Only large companies could afford to carry it out. It could easily become a tool that keeps data locked in large corporate silos for their own interests.

(and on the other side of this coin, **you have the issue that the lack of clear ethical guidelines and sensible regulation around data and privacy would prevent any broader sharing.**)

# 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 (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..

# 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.