Z-Inspection: A holistic and analytic process to assess Ethical AI

Roberto V. Zicari
Frankfurt Big Data Lab
www.bigdata.uni-frankfurt.de

Lecture May 27, 2020

Z-Inspection: Team Members

Roberto V. Zicari, Irmhild van Halem (1), Matthew Eric Bassett (1), Gemma Roig (1), Pedro Kringen (1), Karsten Tolle (1), Todor Ivanov (1), Timo Eichhorn (1), Naveed Mushtaq (1), Melissa McCullough (1), Jesmin Jahan Tithi (2), Romeo Kienzler (4), Georgios Kararigas (3), Marijana Tadic (5), John Brodersen (6), Magnus Westerlund (7), Boris Düdder (8), Florian Möslein (9), Norman Stürtz (1), Rebecca C. Ruehle (10), James Brusseau (11).

(1) Frankfurt Big Data Lab, Goethe University Frankfurt, Germany
(2) *Intel Labs, Santa Clara, CA, USA*
(3) *German Centre for Cardiovascular Research, Charité University Hospital, Berlin, Germany*
(4) *IBM Center for Open Source Data and AI Technologies, San Francisco, CA, USA*
(5) *Cardiology Department, Charité University Hospital, Berlin, Germany*
(6) *Department of Public Health, Faculty of Health Sciences, University of Copenhagen, Danemark*
(7) *Arcada University of Applied Sciences, Helsinki, Finland*
(8) *Department of Computer Science (DIKU), University of Copenhagen (UCPH), Denmark.*
(9) *Institute of the Law and Regulation of Digitalization, Philipps-University Marburg, Germany*
(10) *School of Business and Economics, Vrije Universiteit Amsterdam, The Netherlands*
(11) *Philosophy Department, Pace University, New York, USA*

Ethics and the View of the World

Contemporary Western European democracy.

Fundamental values

The essence of a modern democracy is based on respect for others, expressed through support for fundamental human rights.

- Source:

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

Z-Inspection

A holistic and analytic process to assess Ethical AI

Photo: RVZ

Motivation

- "Ethical impact evaluation involves evaluating the ethical impacts of a technology's use, not just on its users, but often, also on those indirectly affected, such as their friends and families, communities, society as a whole, and the planet."

Source: Dorian Peters, et. al, Responsible AI- Two Frameworks for Ethical Design Practice. IEEE Transactions on Technology and Society, Vol. 1, No. 1, March 2020

Z-Inspection Methodology

Photo RVZ

Z-Inspection methodology

Z-Inspection is designed by integrating and complementing two approaches:

- A **holistic** approach, to try grasping the *whole* without consideration of the various parts;

and

- An **analytic** approach, to consider *each part* of the problem domain.

Combining a holistic and analytic approach to assess Ethical AI in practice

- Z-Inspection is a general inspection process for Ethical AI which can be applied to a variety of domains such as business, healthcare, public sector, etc. It uses applied ethics. To the best of our knowledge, Z-Inspection is the first process that combines a holistic and analytic approach to assess Ethical AI in practice.

- Our approach is learning by doing. We have started to use Z-Inspection to assess a real use case in the area of *AI-based medical devices for enhancing decision-making*.

Orchestration Process

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

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

Why doing an AI Ethical Inspection?

We developed the *Z-Inspection process* with the following goals in mind:

- To help the decision-making process to assess if the use AI in a given context is appropriate;
- To help minimize risks vs. identifying chances associated with an AI in a given context;
- To help establish trust in AI;
- To help improve the design of the AI from a socio-legal-technical viewpoint;
- To help foster ethical values and ethical actions (i.e. stimulate new kinds of innovation).

AI stakeholders

- The assessment process proposed can be used by a variety of AI stakeholders (e.g. from [1]: Designers and engineers, Organisations and corporate bodies, Policymakers and regulators, Researchers, NGOs and civil society, Users/general public, Marginalised groups, Journalists and communicators).

AI Ethics Design and Ethical Maintenance

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

and/or

2. For "*Ethical Maintenance*": If the AI has already been designed/deployed, it can be used to do an AI Ethical sanity check over time, so that a certain AI Ethical standard of care is achieved.

Mindful Use of AI

- We believe we are all responsible, and that the individual and the collective conscience is the existential place where the most significant things happen.
- With Z-Inspection we want to help to establish what we call a *Mindful Use of AI* (#MUAI).

Z-Inspection: Pre-conditions

The following are important questions that need to be addressed and answered before the Z-Inspection assessment process starts:

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

Z-Inspection: Pre-conditions

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

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

Z-Inspection: 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

AI and the Context

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

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

Boundaries of the inspection: Ecosystems

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

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

[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Socio-technical systems

- *"The assessment depends on the entire socio-technical system, i.e. all components of an algorithmic application including all human actors, from the development phase (e.g. with regard to the training data used) to implementation in an application environment and the phase of evaluation and correction."*

  -- German Data Ethics Commission (DEK)

AI, Context, Trust

- *Trust* between humans and AI is not monolithic and the *context* is vital.

- There often exist *varying degrees of trust*, and the *level of trust* sufficient to deploy AI in different *contexts* is therefore an important question for future exploration.

AI, Context, Trust, Ethics, Democracy

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

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

-- German Data Ethics Commission (DEK)

What if the Ecosystems are not Democratic?

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

Do we want to assess if the ecosystem(s) -where the AI has been designed/produced/used- is democratic?

Should this be part of an AI Ethical assessment or not?
We think the answer is yes.

Political and institutional contexts

- We therefore recommend that the decision-making process as to whether and where AI-based products/ services should be used must include, as an integral part, the political assessment of the "democracy" of the ecosystems that define the context.
- The responsible use of AI (processes and procedures, protocols and mechanisms and institutions to achieve it) inherit properties from the wider political and institutional contexts.

We understand that this could be a debatable point.

AI could consolidate the concentration of power

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

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

-- German Data Ethics Commission (DEK)

Should this be part of the assessment?

We think the answer is yes.

“Embedded” Ethics into AI.

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

Example of “Embedded” Ethics into AI: Medical Diagnosis

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

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

-- Thomas Grote , Philipp Berens

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

"Embedded" Ethics into AI: Medical Diagnosis

"In the philosophy of medicine, the status of concepts such as '**health**' and '**disease**' is heavily contested. Here, we can mostly distinguish between two camps, '**naturalists**' and '**normativists**'.

Naturalists assume that these concepts are value-free representations of the world. For instance, according to Christopher Boorse's influential account, disease might be conceived as a biological dysfunction.

Normativists assume that disease is a value-laden concept, mostly employed by practical purposes, such as deciding who should get medical treatment."

-- Thomas Grote , Philipp Berens

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

AI Ethics Scores (Labeling)

We use the word *scoring* to denote the assignment of a numerical value (a score) to an AI-based software for the purpose of evaluating certain areas of our investigation.

Scoring may have different meaning, depending *when* and *why* they are used:

- Before deployment, as part of the AI product-services delivered. In this case scores are static.
- After deployment, as part of a post-ante ethical inspection. In this case scores evolve over time.

Design of the scoring system

The following are design questions that need to be addressed when defining the scoring system:

- Which *areas of investigation* (indicators) should be included in scores, and which should be excluded?
- Which *quality* criteria should scores meet?
- Which elements of scores should be *known*, which should be made *transparent and comprehensible, and which should not*?
- What is the time frame for the scoring system (*static, dynamic*)?

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…*

What to do with the output of this investigation?

- Provide feedback to the AI designers/developers (when available) to help them change/improve the AI model/the data/ the training and/or the deployment of the AI in the context;

- Give feedback to decision makers to help them to decide how and when to use (or not) the AI (*Trade-off* concept) - given certain constraints, requirements, and ethical reasoning.

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.

Ethical Tensions

- We use the term ‘tension’ as defined in [1] to refer to different ways in which values can be in conflict, Specifically „tensions between the pursuit of different values in technological applications rather than an abstract tension between the values themselves.“

Z-Inspection:
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)

Implication of IP on the Investigation

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

- A recent published letter by a number of medical scientists [*] mentions "*Google published a paper in Nature claiming their artificial intelligence system outperforms human radiologists. Unfortunately they do not provide the code or data that backs up this claim, arguing that doing so is "not feasible" and claiming that others can reproduce it without code*".

- The letter continues describing "*the problems for science brought about by failure to make computational methods transparent and reproducible.* "

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

Focus of Z-Inspection

- Ethical
- Technical
- Legal

Note1: *Illegal and unethical are not the same thing*.

Note2: *Legal and Ethics depend on the context*

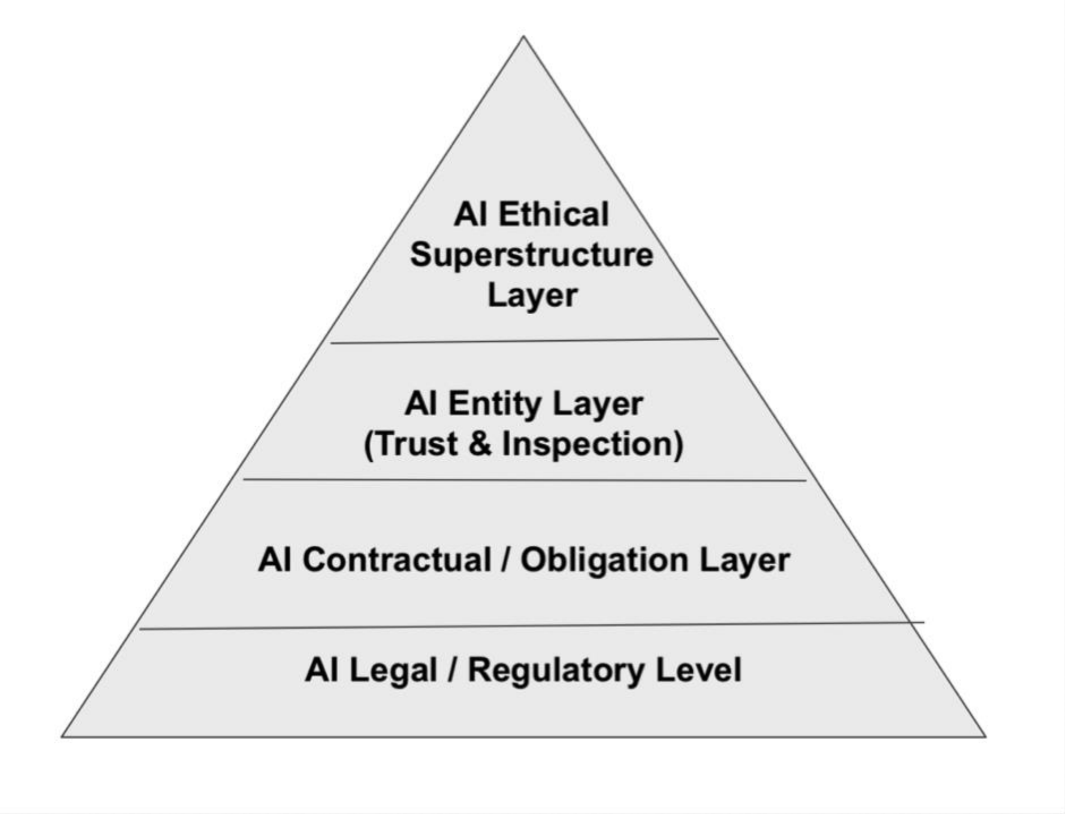
Note 3*:* Relevant/accepted for the ecosystem(s) of the AI use case.

Trustworthy artificial intelligence

EU High-Level Expert Group on AI presented their ethics guidelines for trustworthy artificial intelligence:

- (1) **lawful** - respecting all applicable laws and regulations
- (2) **ethical** - respecting ethical principles and values
- (3) **robust** - both from a technical perspective while taking into account its social environment

Z-Inspection Layered Model


AI Ethical Superstructure Layer
AI Entity Layer (Trust & Inspection)
AI Contractual / Obligation Layer
AI Legal / Regulatory Level

Z-Inspection Layered Model

I. AI Legal/Regulatory Must Layer

This layer refers to actions or elements of Z-Inspection that are either proposed by law or could fullfill the purpose of the law. This layer could be referred to, for example, by actions necessary to fullfill GDPR.

II. AI Contractual Obligation Layer

This layer represents all obligations, duties and rights from a contract that a given entity using and developing AI solutions enters with their counterparts, either by contractual negotiation or also by documented, auditable consent.

Z-Inspection Layered Model

III. AI Entity Layer

This layer reflects all paths and their resulting actions that are voluntarily done by and inside the entity employing AI solutions has established in order to inspect an AI object and document results on that inspection. This layer is legally or regulatory not a must and should logically not be part of the contractual obligation, respectively any contractual obligation should refer to it separately.

IV. AI Ethical Superstructure

This layer goes into the realms of not defined actions and processes that should take place in order to cater for overarching higher principles, e.g. data rights, human dignity, civil liberty, which are not covered by scientific, societal or political discussion. It is a layer which is not mandatory but where ethical principles are discussed and as such implicitly made a goal post to orient at.

Ethical Principles in the Context of AI Systems

EU **four ethical principles**, rooted in fundamental rights

(i) Respect for human autonomy

(ii) Prevention of harm

(iii) Fairness

(iv) Explicability

- **Tensions between the principles**

Requirements of Trustworthy AI

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

1 Human agency and oversight

Including fundamental rights, human agency and human oversight

2 Technical robustness and safety

Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility

3 Privacy and data governance

Including respect for privacy, quality and integrity of data, and access to data

4 Transparency

Including traceability, explainability and communication

Requirements of Trustworthy AI

5 Diversity, non-discrimination and fairness

Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

6 Societal and environmental wellbeing

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

7 Accountability

Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

Z-Inspection: *Areas of investigations*

We use ***Conceptual clusters:***

- 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.

Use Socio-technical scenarios

- 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*

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

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Concept Building

An important obstacle to progress on the ethical and societal issues raised by AI-based systems is the *ambiguity* of many central *concepts* currently used to identify salient issues:

- **Terminological overlaps**
- **Differences between disciplines**
- **Differences across cultures and publics**

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Developing an evidence base

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

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

Source: Whittlestone, J et al (2019)

Identify, Classify and Describe Tensions

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

Examples of Tensions:

- **Accuracy vs. fairness**
- **Accuracy vs explainability**
- **Privacy vs. Transparency**
- **Quality of services vs. Privacy**
- **Personalisation vs. Solidarity**
- **Convenience vs. Dignity**
- **Efficiency vs. Safety and Sustainability**
- **Satisfaction of Preferences vs. Equality**

Source: Whittlestone, J et al (2019)

Address, Resolve *Tensions*

Describe and Classify Trade-offs- Iterative process:

- ***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.

Source: Whittlestone, J et al (2019)

Ethical Issues and Flags

The outcome of this part of the investigation is a list of *Ethical issues*, E1….Ei, and of *Flags*, F1...Fj which need to be further investigated.

- An *Ethical issue* or tension refers to different ways in which values can be in conflict.
- A *Flag* is an issue that needs to be assessed further. It could be a potential ethical tension, and/or policy issue, and/or a technical issue, and/or a legal issue.
- This is the result of an iterative process, based on the common understanding of the scenarios by whom is analysing them;

Mappings from Ethical issues and Flags to the Areas of Investigation

This is a process per se.

It may require more than one iteration between the team members in charge. *The choice of who is in charge has an ethical and a practical implication*.

It may require once more the application of Concept building, to help mapping for example, how an ethical issue Ei is assigned to a conceptual cluster of area, e.g. Bias/Fairness/Discrimination, and to arrive to a "consensus".

There are several Steps for this Mapping

Macro vs Micro Investigation

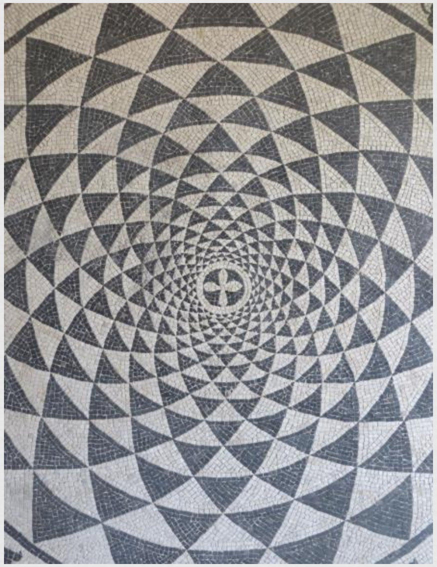
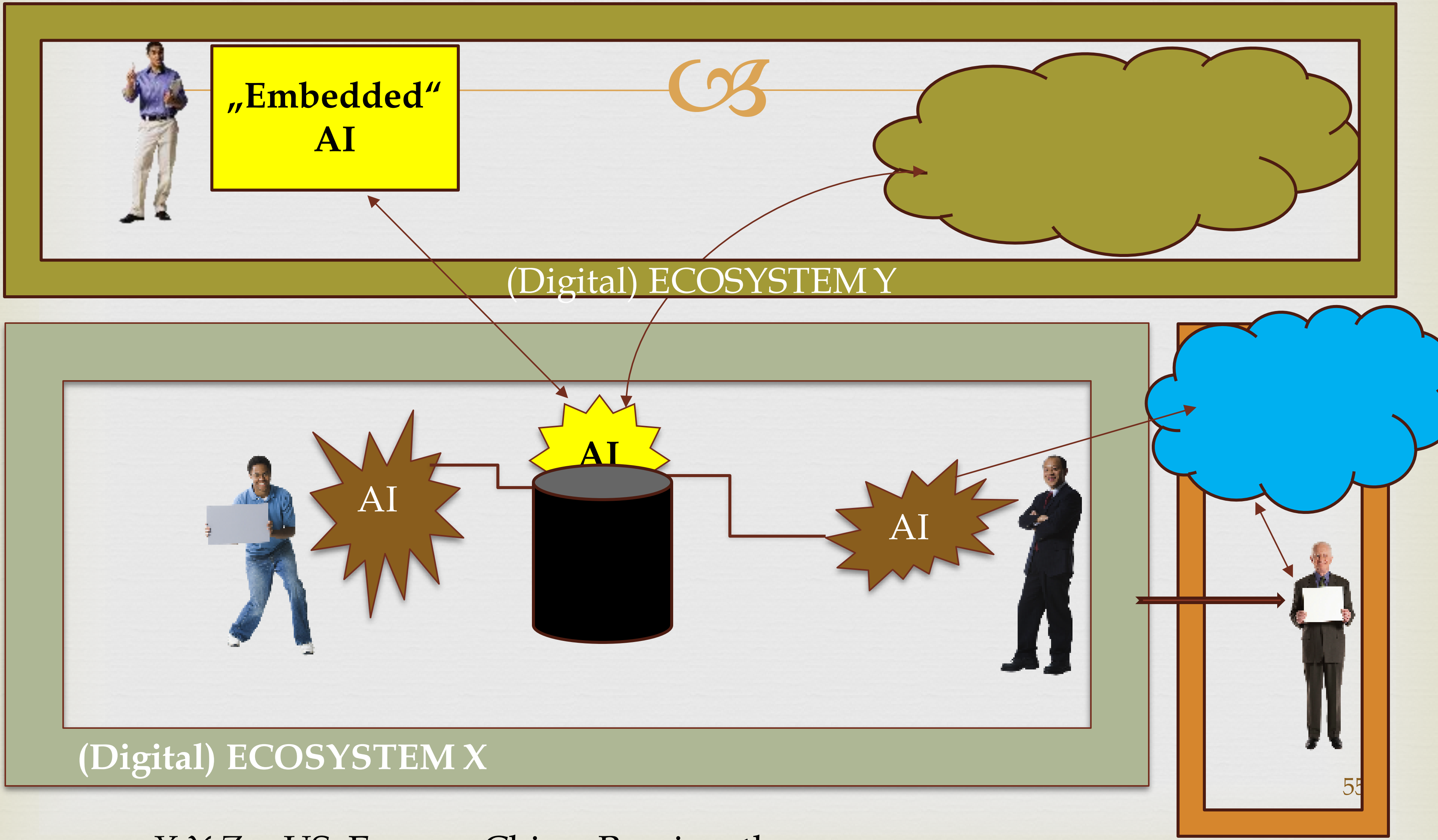


Photo RVZ

Ethical AI “*Macro*” -Investigation



„Embedded“ AI

(Digital) ECOSYSTEM Y

AI

AI

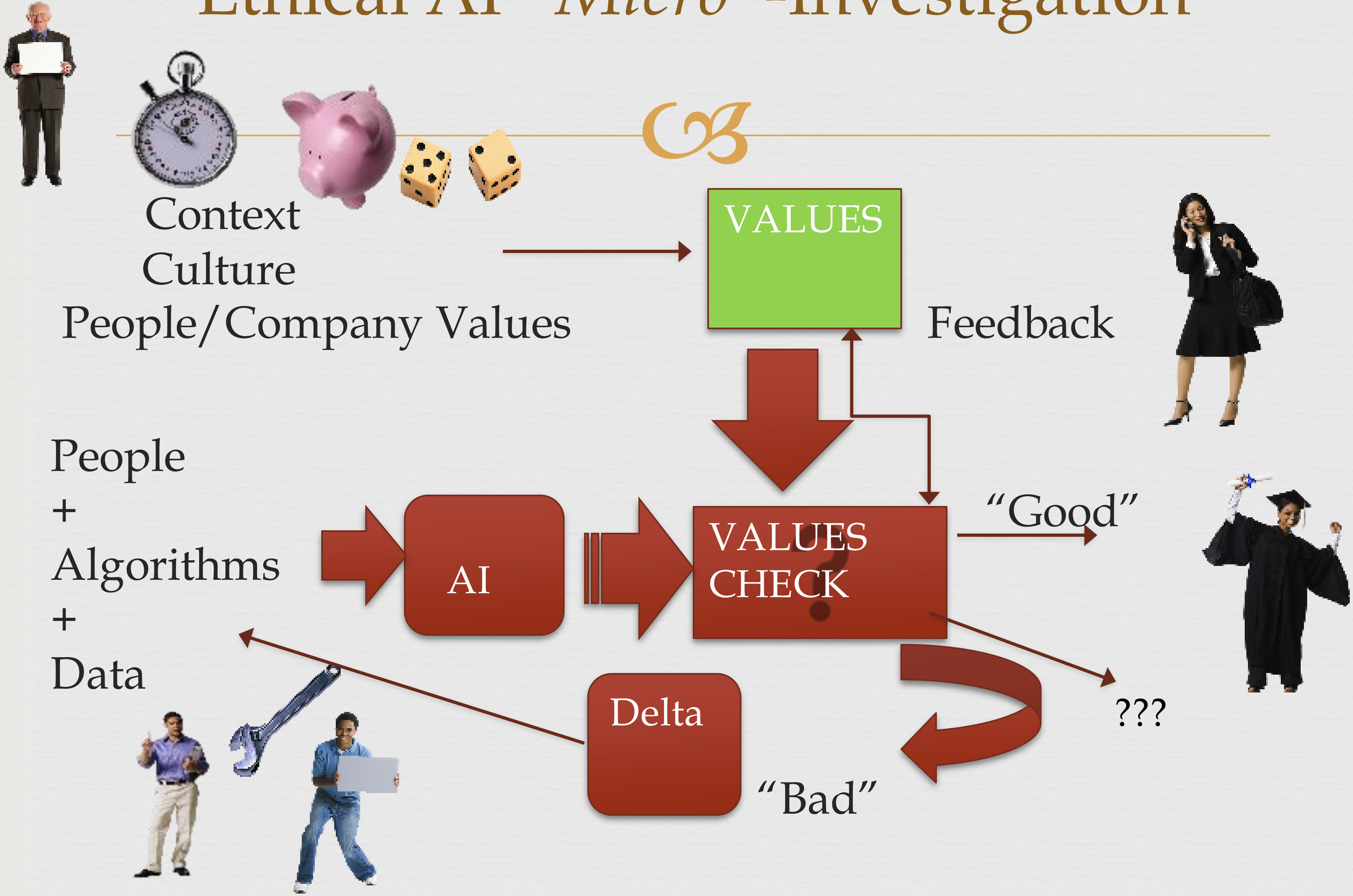
AI

AI

(Digital) ECOSYSTEM X

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

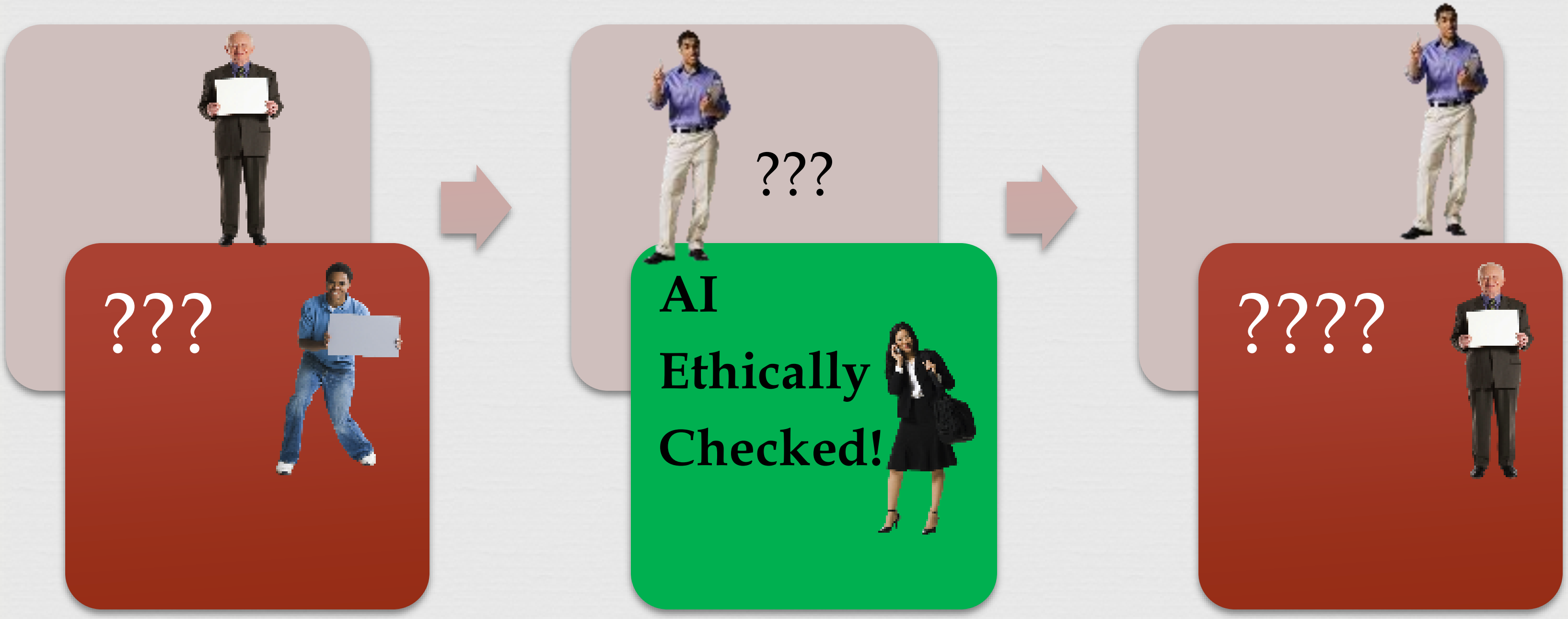
"Good"

???

Delta

"Bad"

Micro-validation does not imply *Macro*-validation



???

???

AI Ethically Checked!

????

Layers

- A layer is a subset of the boundaries of the inspection considered at a certain level of abstraction (Macro vs. Micro). Each level of abstraction is a layer.
- A number of layers may be created for the given boundaries.

Paths

- A *Path* P addressing E_i and F_j associated to a cluster area C (e.g Bias/Fairness/Discrimination) can be composed of a number of steps to assess a set of Ethical tensions E_i and Flags F_j.

- *Execution of a Path* corresponds to the execution of the corresponding steps; steps of a path are performed by team members. A step of a path is executed in the context of one or more layers. Execution is performed in a variety of ways, e.g. via workshops, interviews, checking and running questionnaires and checklists, applying software tools, measuring values, etc.

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)

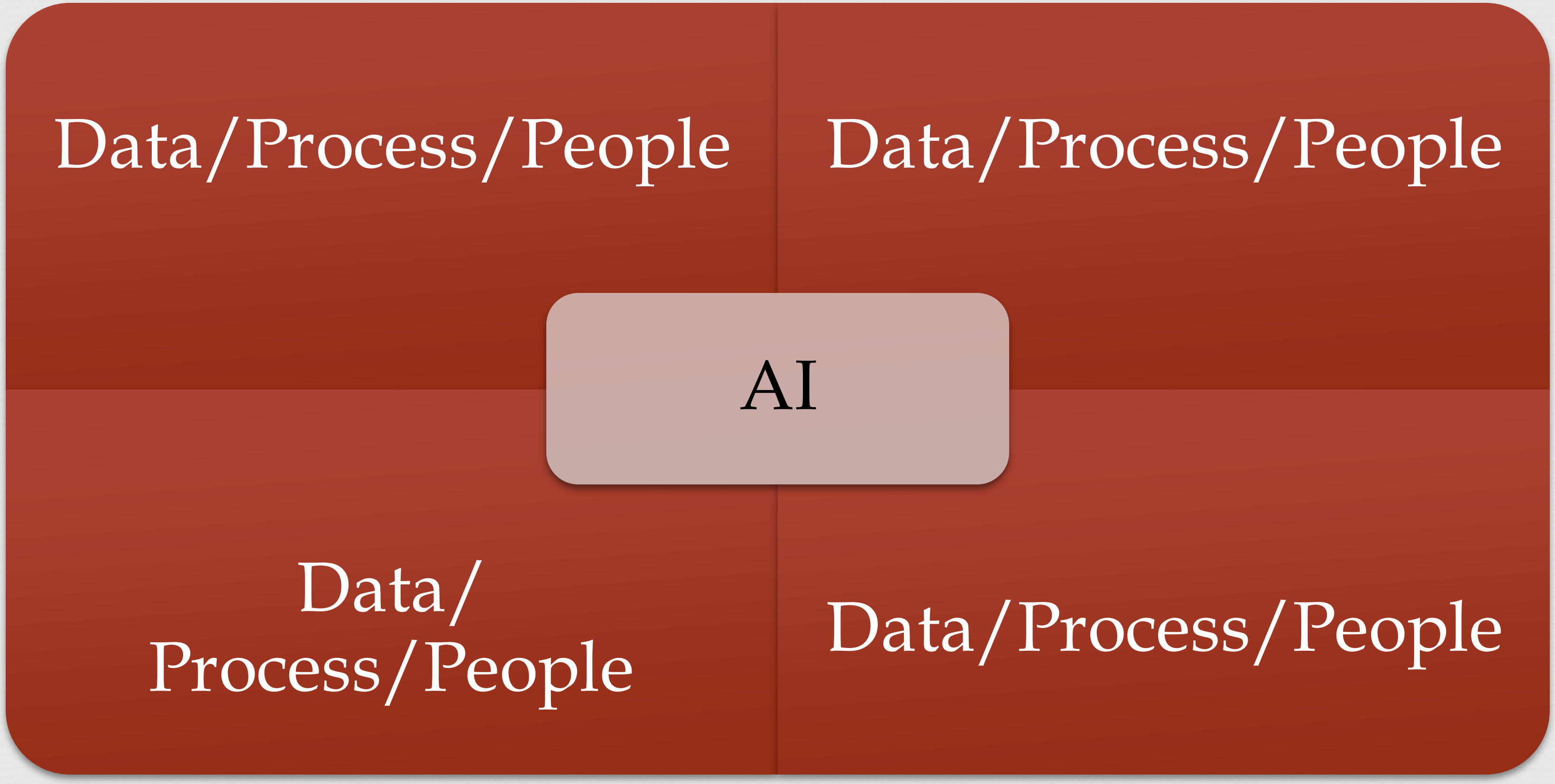
Choosing an Inspection Methodology

- 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

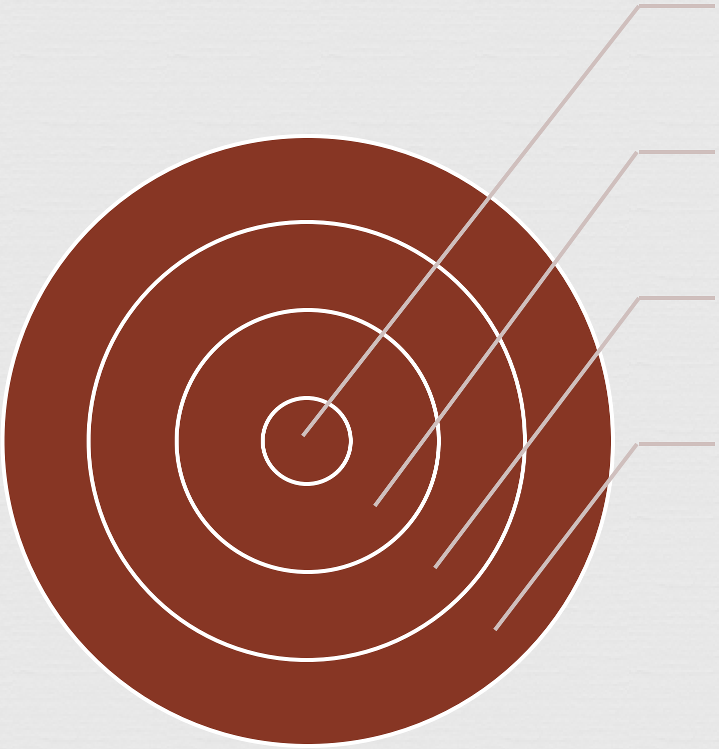
Data/Process/People

AI

Data/
Process/People

Data/Process/People

Iterative Inside Out Approach



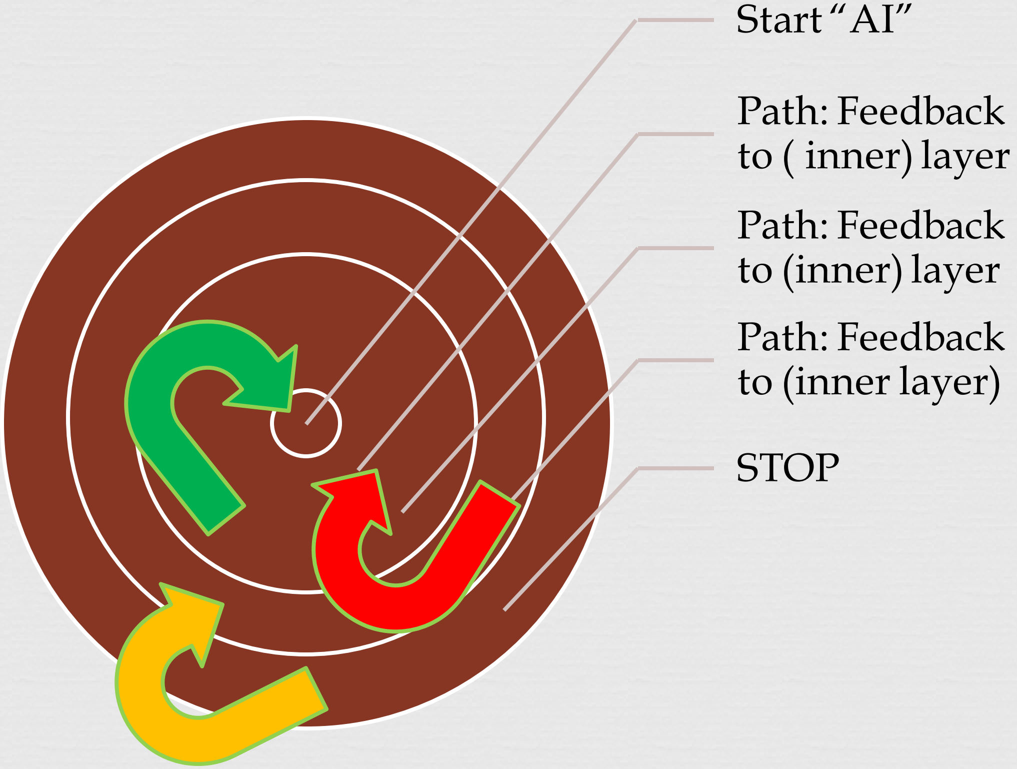
Start with AI. Iterate 5 phases: Explanability, Fairness, Safety, Human-AI, Liability

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

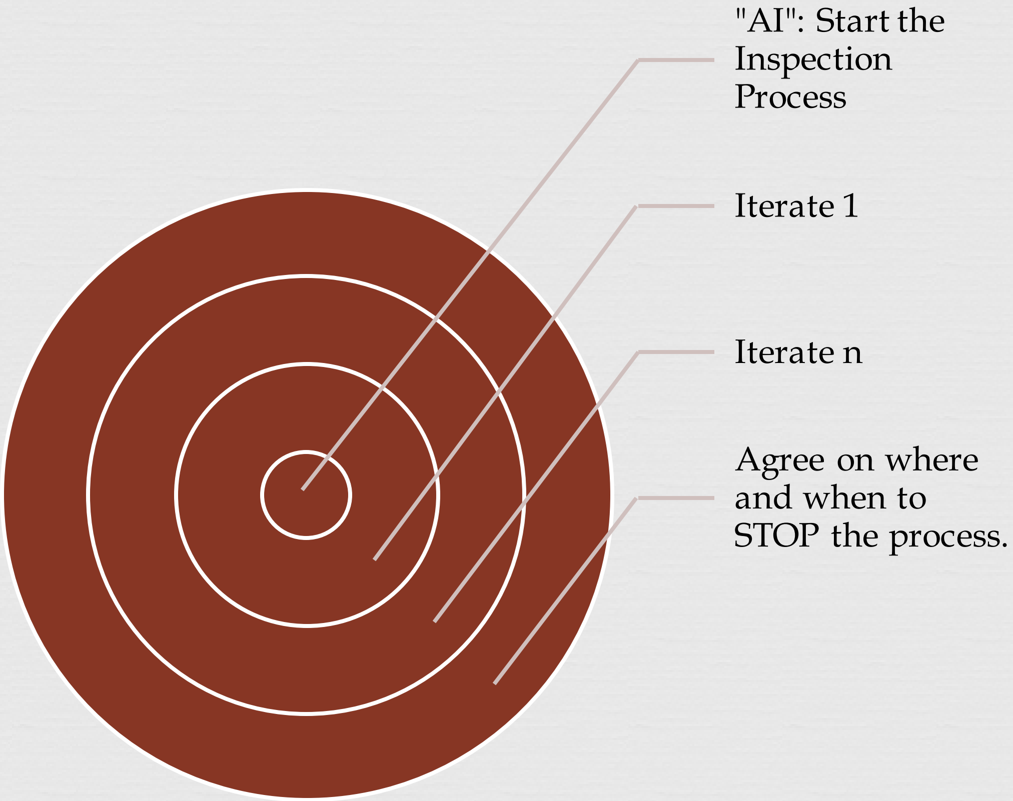
Augument Explanability++, Fairness++, Safety++, Human-AI++, Liability++

Iterate taking into account the big picture(Macro/Ecosys tems)

Interactive Inside Out Approach Paths and Feedback mechanism


Start "AI"
Path: Feedback to (inner) layer
Path: Feedback to (inner) layer
Path: Feedback to (inner layer)
STOP

Agree on when and where to STOP the inspection



"AI": Start the Inspection Process

Iterate 1

Iterate n

Agree on where and when to STOP the process.

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

Use Case

Assessing an AI-based Medical Device for Enhancing Decision-making (Cardiology)

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.

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.

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

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 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.**

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

- Patients received "Green". Doctor agree. Patient does nothing;
- Patients received "Green". Patient and/or Doctor do not trust, asked for further invasive test;
- Patient received "Red/ Yellow". Doctor agree. Patient does nothing;
- Patient received "Red/Yellow" - Patient asks for further invasive test;

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

Building an Evidence Base

- Grote and Berens [58] argue that "*deploying machine learning algorithms in healthcare entails trade-offs at the epistemic and the normative level, with the risk of potentially undermining the epistemic authority of clinicians, and possible pitfalls with respect to paternalism, moral responsibility and fairness*".

[58] Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare, *J Med Ethics* 2019;0:1–7. doi:10.1136/medethics-2019-105586.

Building an Evidence base

- They also noticed how the deployment of machine learning algorithms might shift the evidentiary norms of medical diagnosis.

- For example, a patient may come to harm if the prediction is not accurate and if no explanation for the result is possible, while gaining truly informed consent from the patient might not be possible.

Building an Evidence Base

As stated in [58]:

- *"As the patient is not provided with sufficient information concerning the confidence of a given diagnosis or the rationale of a treatment prediction, she might not be well equipped to give her consent to treatment decisions."*

[58] Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare, *J Med Ethics* 2019;*0*:1–7. doi:10.1136/medethics-2019-105586.

Building an Evidence Base: Accountability

- Even an accurate prediction of a deteriorating patient state may be problematic as the clinicians may lack a sufficient understanding of the algorithm output to perform an evidence-based treatment.

- This has a severe ethical implication because clinicians are being held accountable for their decisions.

Building an Evidence Base: “low risk” AI-based Medical Devices

- As indicated by [51], existing AI systems in healthcare are not as rigorously tested as other medical devices, and this could raise risks.

[51] *Artificial Intelligence Is Rushing Into Patient Care - And Could Raise Risks*Liz Szabo, Scientific America, December 24, 2019

Building an Evidence Base: Ethical Maintenance

- Our practical experience in assessing the Ethical implications of AI systems in medicine calls for what we define the need of an "Ethical Maintenance".

- Especially when the AI is not fixed once deployed and evolve over time via model updates/continual interaction. When the AI model is constantly updated/ improved using new training/test data, it becomes impossible to compare predictions for the same patients, which have been produced with different versions of the AI model.

Building an Evidence Base: Ethical Maintenance

- In this case, even peer-reviewed medical evidence published based on a specific AI model trained and validated with a specific data set, may not hold true with respect to the new upgraded version of the AI product/service based on the new AI model.

- But after deployment it gets more tricky, how is it possible to show continued evidence?

Choosing Context-related Ethics

For our use case, we consider *Western clinical medical ethics.*

Four classical principles of (*)

- Justice
- Autonomy
- Beneficence
- Nonmaleficence

Where *"Western"* define a set of implicit *ecosystems*…

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

Analysis of Socio-technical scenario *Discover potential ethical issues*

Main findings:

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.*

Analysis of Socio-technical scenario *Discover potential ethical issues*

Diagnostic Trust and Competence – ethical issues:

- When CSG is being used in screening asymptomatic patients who are "*notified*" by the AI 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**

Analysis of Socio-technical scenario *Discover potential ethical issues*

Diagnostic Trust and Competence – ethical issues:

- 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**.

Analysis of 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.

Analysis of Socio-technical scenario *Discover potential ethical issue*

Diagnostic Trust and Competence – ethical issues:

- 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".

Analysis of Socio-technical scenario *Discover potential ethical issue*

Diagnostic Trust and Competence – ethical issues:

- **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.

Evidence Base
Machine Learning Bias in healthcare

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

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

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/

Evidence Base
Machine Learning Bias in healthcare

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

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/

Evidence Base
Machine Learning Bias in healthcare

- **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/

Analysis of Socio-technical scenario *Discover potential ethical issue*

- **How high is the risk that an application/diagnostic error happens** with the traditional diagnostic instruments compared to using the CSG app?

Identify, Classify and Verify Tensions

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

- Need to Develop a sound (medical) evidence base

- Decide how deep we want to go with the investigation (taking into account IP)

- Create and Execute a Path

Execution of Path
*Assessing **fairness***

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

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

a. *Context-relevant Ethical values,*

↓

b. Domain-specific metrics,

↓

c. Machine Learning fairness metrics.

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

a. *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)

Fairness in Healthcare: *Different definitions*

For our use case, suppose we are concerned with whether the AI-based device 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)

Choosing *Fairness* criteria

(domain specific)

Step 3. For *healthcare* one possible approach is to use ***Distributive justice*** (from philosophy and social sciences) **options for machine learning** (*)

Step 4. 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/

Choosing *Fairness* criteria

(domain specific)

BUT, could we use another 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.*

A consensus need to be reached….

Applying ML and *Fairness* criteria

in healthcare (domain specific)

Step 4. 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**)?

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

Applying ML and *Fairness* criteria

in healthcare (domain specific)

- **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)**

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

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

- **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

Applying ML and *Fairness* criteria

in healthcare (domain specific)

Equal Outcomes

Equal Performance

Equal Allocation

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

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 ?

From Domain Specific to ML metrics

Several Approaches:

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

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

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

Mapping Domain specific “Fairness” to Machine Learning metrics

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

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

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

Trust in Machine Learning "Fairness" metrics

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

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

Predictions in conjunction with other "signals"

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

→ *Trust in the ecosystem*

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

Known Trade Offs
(Incompatible types of fairness)

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

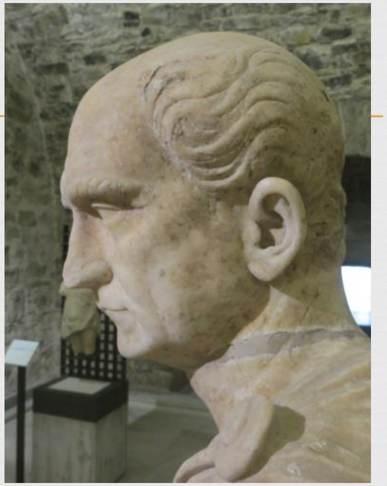
- Equal positive and negative predictive value vs. equalized odds
- Equalized odds vs. equal allocation
- Equal allocation vs. equal positive and negative prediction value

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/

Reflection Moment



At the beginning of the process we re-assessed our team, and we realized that having a *independent medical public health experts and cardiologists* in the team would improve our inspection process for this use case and help us assessing the relevant medical *evidence base…*

- Photo RVZ

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?

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-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**

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

“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*…

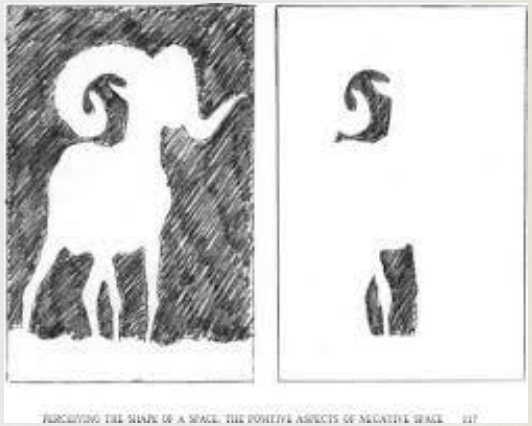
- I mean, if we really insist on *AI Transparency*, perhaps this would force us to reveal our real *motives*…
- But, we do not always wish to make our motives visible to the outside world, e.g. we do not wish transparency….
- But with no transparency, there is a lack of trust.

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?


PERCEIVING THE SHAPE OF A SPACE: THE POSITIVE ASPECTS OF NEGATIVE SPACE 117

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 ?

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..

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, Clemens Szyperski and Shizuka Uchida

for proving valuable comments and feedback.