Mindful Use of AI
Z-inspection®
A process to assess Trustworthy AI

Roberto V. Zicari
Frankfurt Big Data Lab
http://z-inspection.org

AI4EU Workshop. Nov. 13, 2020

Z-inspection® : Core Team Members

Roberto V. Zicari *(1)*, John Brodersen *(4)(9)*, James Brusseau *(8)*, Boris Düdder *(6)*, Timo Eichhorn *(1)*, Todor Ivanov *(1)*, Georgios Kararigas *(3)*, Pedro Kringen *(1)*, Melissa McCullough *(1)*, Florian Möslein *(7)*, Naveed Mushtaq *(1)*, Gemma Roig *(1)*, Norman Stürtz *(1)*, Karsten Tolle *(1)*, Jesmin Jahan Tithi *(2)*, Irmhild van Halem *(1)*, Magnus Westerlund *(5)*.

(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) Section of General Practice and Research Unit for General Practice, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Danemark; (5)Arcada University of Applied Sciences, Helsinki, Finland; (6) Department of Computer Science (DIKU), University of Copenhagen (UCPH), Denmark; (7) Institute of the Law and Regulation of Digitalization, Philipps-University Marburg, Germany; (8) Philosophy Department, Pace University, New York, USA; (9) Primary Health Care Research Unit, Region Zealand, Denmark

Trustworthy AI Framework

Photo RVZ

European Commission. Independent High-Level Experts Group on AI.

Four ethical principles, rooted in fundamental rights

(i) Respect for human autonomy

(ii) Prevention of harm

(iii) Fairness

(iv) Explicability

- Tensions between the principles

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

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

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

Requirements of Trustworthy AI

1 Human agency and oversight
Including fundamental rights, human agency and human oversight

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

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

4 Transparency
Including traceability, explainability and communication

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

Requirements of Trustworthy AI

5 Diversity, non-discrimination and fairness

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

6 Societal and environmental wellbeing

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

7 Accountability

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

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

How do we know what are the Benefits vs. Risks of an AI system?

- Our approach is inspired by both theory and practices (" learning by doing").

photo CZ

Best Practices

- Assessing Trustworthy AI. Best Practice: AI for Predicting Cardiovascular Risks (Jan. 2019-August 2020)
- Assessing Trustworthy AI. Best Practice: Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. (September 2020-March 2021)
- Assessing Trustworthy AI. Best Practice: Deep Learning based Skin Lesion Classifiers. (November 2020-March 2021)

http://z-inspection.org/best-practices/

Our Team

Roberto V. Zicari (1), Nikita Aggarwal (23),Vegard Antun (26),Valentina Beretta (22), John Brodersen (4)(9), James Brusseau (8), Herbert Chase (12), Megan Coffee (18), Maria Chiara Demartini (22), Boris Düdder (6), Alessio Gallucci (28), Marianna Ganapini (21), Thomas Gilbert (15), Emmanuel Goffi (16), Philippe Gottfrois (33), Christoffer Bjerre Haase (34), Thilo Hagendorff (29), Elisabeth Hildt (17), Ludwig Christian Hinske (24), Sune Holm (25), Todor Ivanov (1), Georgios Kararigas (3), Pedro Kringen (1), Ulrich Kühne (32), Vince Madai (27), Melissa McCullough (1), Carl-Maria Mörch (12), Florian Möslein (7), Naveed Mushtaq (1), Davi Ottenheimer (20), Matiss Ozols (14), Laura Palazzani (10), Anne Riechert (30),Eberhard Schnebel (1), Andy Spezzati (11), Gemma Roig (1), Rita Singh (31), Norman Stürtz (1), Karin Tafur, Jim Tørresen (19), Karsten Tolle (1), Jesmin Jahan Tithi (2), Irmhild van Halem *(1)*, Dennis Vetter (1), Magnus Westerlund *(5)*.

Our Team

- (1) Frankfurt Big Data Lab, Goethe University Frankfurt, Germany;
(2) Intel Labs, Santa Clara, CA, USA;
(3) Department of Physiology, Faculty of Medicine, University of Iceland, Reykjavik, Iceland;
(4) Section of General Practice and Research Unit for General Practice, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Danemark;
(5) Arcada University of Applied Sciences, Helsinki, Finland;
- (6) Department of Computer Science (DIKU), University of Copenhagen (UCPH), Denmark;
(7) Institute of the Law and Regulation of Digitalization, Philipps-University Marburg, Germany;
(8) Philosophy Department, Pace University, New York, USA;
(9) Primary Health Care Research Unit, Region Zealand, Denmark;
(10) Philosophy of Law, LUMSA University, Rome, Italy;
(11) Computer Science Department, UC Berkeley, USA
(12) Clinical Medicine, Columbia University Medical Center, USA
(13) Université de Montréal and Mila, Canada
(14) Division of Cell Matrix Biology and Regenerative Medicine, The University of Manchester, UK
(15) Center for Human-Compatible AI, University of California, Berkeley, USA
(16) Observatoire Ethique & Intelligence Artificielle de l'Institut Sapiens, Paris and aivancity, School for Technology, Business and Society, Paris-Cachan, France
(17) Center for the Study of Ethics in the Professions, Illinois Institute of Technology Chicago, USA
(18) Department of Medicine and Division of Infectious Diseases and Immunology,NYU Grossman School of Medicine, New York, USA
(19) Department of Informatics, University of Oslo, Norway
(20) Inrupt, San Francisco, USA
(21) Philosophy Department, Union College, NY, USA
(22) Department of Economics and Management, Università degli studi di Pavia, Italy
(23) Faculty of Law and Oxford Internet Institute, University of Oxford, UK
(24) Klinik für Anaesthesiologie, LMU Klinikum. Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie, Ludwig-Maximilians-Universität München, Germany
(25) Department of Food and Resource Economics, Faculty of Science, University of Copenhagen, DK
(26) Department of Mathematics, University of Oslo, Norway
(27) Charité Lab for AI in Medicine, , Charité Universitätsmedizin Berlin, Germany
(28) Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands.
(29) Cluster of Excellence "Machine Learning: New Perspectives for Science" – Ethics & Philosophy Lab University of Tuebingen , Germany
(30) Data Protection Law and Law in the Information Processing, Frankfurt University of Applied Sciences, Germany
(31) Language Technologies Institute, School of Computer Science, Carnegie Mellon University, USA
(32) "Hautmedizin Bad Soden", Germany
(33) Department of Biomedical Engineering, Basel University, Switzerland
(34) Section for Health Service Research and Section for General Practice, Department of Public Health, University of Copenhagen, Denmark. Centre for Research in Assessment and Digital Learning, Deakin University, Melbourne, Australia

Focus of Z-inspection®

Z-inspection® covers the following:

- Ethical and Societal implications;
- Technical robustness;
- Legal/Contractual implications.

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

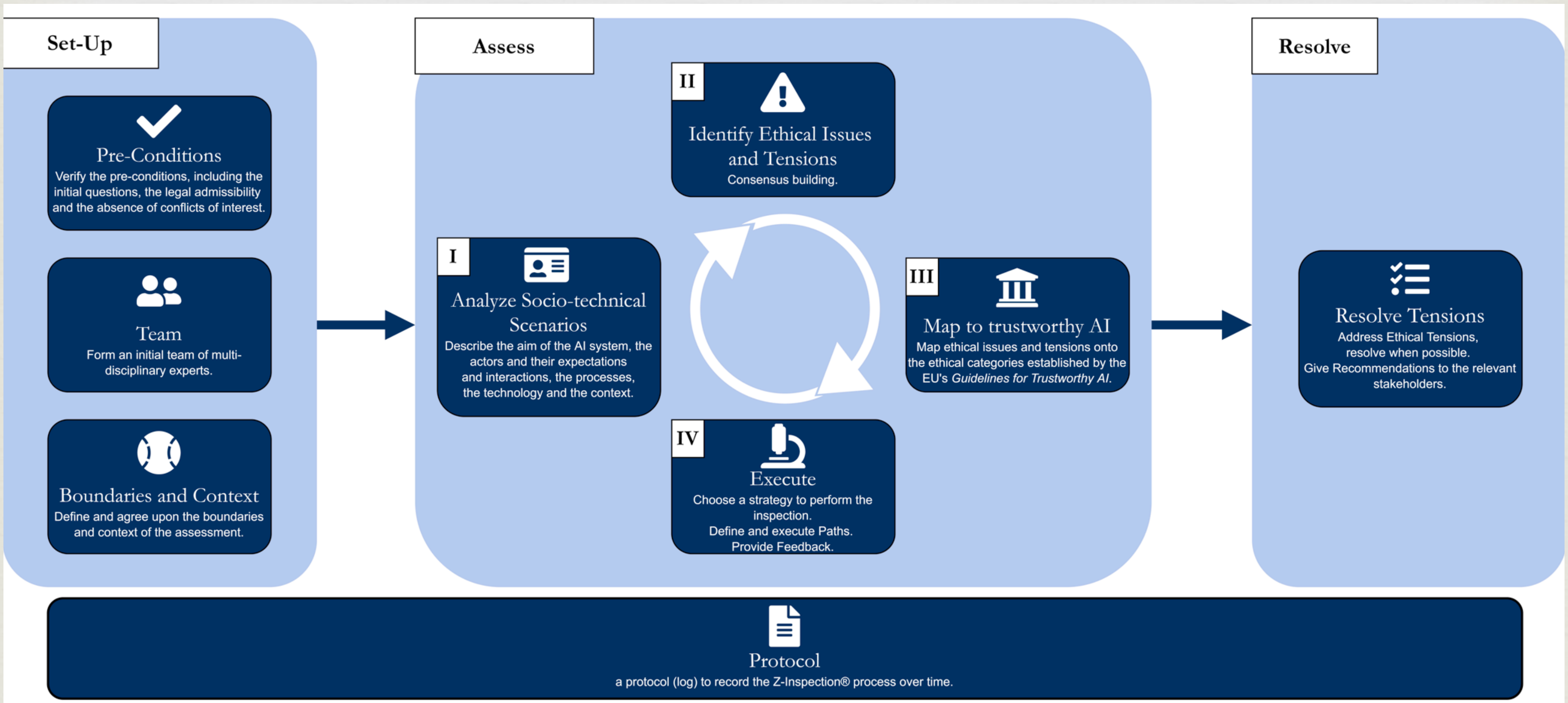
Note2: *Legal and Ethics depend on the context*

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

Orchestration Process

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

Z-inspection® Process in a Nutshell



Set Up

Who? Why? For Whom?

We defined a catalogue of questions to help clarify the expectation between stakeholders, 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 analysed?
- 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).

What to do with the assessment?

- 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.

No conflict of interests: 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/platforms 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

Responsible use of AI

- 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**.

AI, Context, Trust, Ethics, Democracy

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

There is a "*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:

What if the Ecosystems are not Democratic?

Political and institutional contexts

- We 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.

We understand that this could be a debatable point.

What if the AI consolidates 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.

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 or not disclosing them at all.

Build a Team

A team of multi-disciplinary experts is formed. The composition of the team is a dynamic process. Experts with different skills and background can be added at any time of the process.

The choice of experts have an ethical implication!

Create a Log

- A protocol (log) of the process is created that contains over time several information, e.g. information on the teams of experts, the actions performed as part of each investigation, the steps done in data preparation and analyses and the steps to perform use case evaluation with tools.
- *The protocol can be shared to relevant stakeholders at any time to ensure transparency of the process and the possibility to re-do actions;*

Define the Boundaries and Context of the inspection

- 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.

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.

Define the time-frame of the assessment.

We need to decide which time-scale, we want to consider when assessing Ethical issues related to AI.

A useful framework that can be used for making a decision, is defined in [1], formulating three different time-scales:

- Present challenges: "*What are the risks we are already aware of and already facing today?*"
- Near-future challenges: "*What risks might we face in the near future, assuming current technology?*"
- Long-run challenges: "*What the risks and challenges might we face in the longer-run, as technology becomes more advanced?*"

The choice of which time-scale to consider does have an impact on our definition of an "Ethical maintenance"

[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.

Assess

Socio-technical Scenarios

- Socio-technical scenarios are created (or given to) by the team of experts to represent possible scenarios of use of the AI. This is a process per se, that involves several iterations among the experts, including using *Concept Building*.

Socio-technical Scenarios

By collecting relevant resources, socio-technical scenarios are created and analyzed by the team of experts:

to **describe the aim of the AI systems**,

the actors and their expectations and interactions,

the process where the AI systems are used,

the **technology and the context**.

Identification of Ethical issues and tensions.

- An appropriate *consensus building* process is chosen that involves several iterations among the experts of different disciplines and backgrounds and result in identifying ethical issues and ethical tensions.

Ethical Tensions

- We use the term 'tension' as defined in [1]
  „tensions between the pursuit of different values in technological applications rather than an abstract tension between the values themselves."

[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.

“Embedded” Ethics into AI.

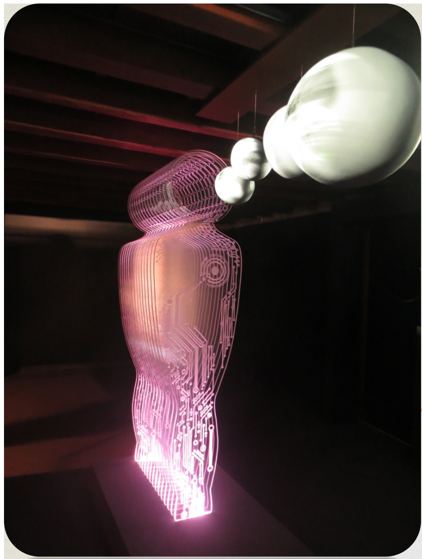
- When designing, training and testing an AI-system (e.g. Machine-Learning algorithm) we do “embed” into the system notions such as “good”, “bad”, “healthy”, “disease”, etc. mostly not in an explicit way.

“Embedded” Ethics 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.”

-- 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



Identify Ethical Issues and Tensions, and Flags

- As a result of the analysis of the scenarios, **Ethical issues** and **Flags** are identified .

- 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 technical, legal, ethical issue)

Describe Ethical issues and Tensions

- *Confirm*, *describe* and *classify* if such Ethical Issues represent ethical tensions and if yes, describe them.
- This is done by a selected number of members of the inspection team, who are experts on ethics and/or the specific domain.
- Goal is to reach a "consensus" among the experts (when possible) and agree on a common definition of Ethical tensions to be further investigated in the Z-Inspection process.

Describe Ethical issues and Tensions

- A method we have been using consists of reviewing the applied ethical frameworks relevant for the domain, asking the experts to classify the ethical issues discovered with respect to
  - **a pre-defined catalog of ethical tensions.**
  - **a classification of ethical tensions.**

Catalogue of Examples of Tensions

From (1):

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

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

Classification of ethical tensions

From [1]:

- **true dilemma**, i.e. "a conflict between two or more duties, obligations, or values, both of which an agent would ordinarily have reason to pursue but cannot";
- **dilemma in practice**, i.e. "the tension exists not inherently, but due to current technological capabilities and constraints, including the time and resources available for finding a solution";
- **false dilemmas**, i.e. "situations where there exists a third set of options beyond having to choose between two important values".

Mapping to Trustworthy AI.

- 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*.

Mapping to Trustworthy AI.

- Once the ethical issues and tensions have been agreed upon among the experts, the consensus building process among experts continue by asking them to map ethical issues and tensions onto

- **the four ethical categories,** and
- **the seven requirements established by the EU High Level Experts Guidelines for Trustworthy AI**

Case Study
AI for Predicting Cardiovascular Risks

- Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year. Over the past decade, several machine-learning techniques have been used for cardiovascular disease diagnosis and prediction. The potential of AI in cardiovascular medicine is high; however, ignorance of the challenges may overshadow its potential clinical impact

The AI System

- The product we assessed was a non-invasive AI medical device that used machine learning to analyze sensor data (i.e. electrical signals of the heart) of patients to predict the risk of cardiovascular heart disease.
- The company uses a traditional machine learning pipeline approach, which transforms raw data into features that better represent the predictive task. The features are interpretable and the role of machine learning is to map the representation to output. The mapping from input features to output prediction is done with a classifier based on several neural networks that are combined with a Ada boost ensemble classifier.
- The output of the network is an Index (range -1 to 1), a scalar function dependent on the input measurement, classifying impaired myocardial perfusion.

Machine Learning Pipeline

1. Measurements, Data Collection (Data acquisition, data annotation with the ground truth, Signal processing)

2. Feature extraction, features selection

3. Training of the Neural Network-based classifier using the annotated examples.

4. Once the model is trained (step 3), actions are taken for new data, based on the model's prediction and interpreted by an expert and discussed with the person.

Actors and Scenarios of use (simplified)

When the AI-system is used in a patient, the possible actions taken based on model's prediction are:

- The AI-systems predict a **"Green"** score for the patient. Doctor agrees. No further action taken, and the patient does nothing;
- AI-systems predict a **"Green"** score for the patient. The patient and/or Doctor do not trust the prediction. Patient is asked for further invasive test;
- The AI-systems predict a **"Red"** score for the patient. Doctor agrees. Nevertheless , no further action taken, and the patient does nothing;
- The AI-systems predicts a **"Red"** score for the patient; Doctor agrees. Patient is asked for further invasive test;
- In a later stage, the company introduced a third color, **"Yellow"**, to indicate a general non specified cardiovascular health issue.

Examples of mapping

ID Ethical "Issue": E7 Description: The data used to optimize the ML predictive model is from a limited geographical area, and no background information on difference of ethnicity is available. All clinical data to train and test the ML Classifier was received from three hospitals in all of them near to each other. There is a risk that the ML prediction be biased towards a certain population segment.

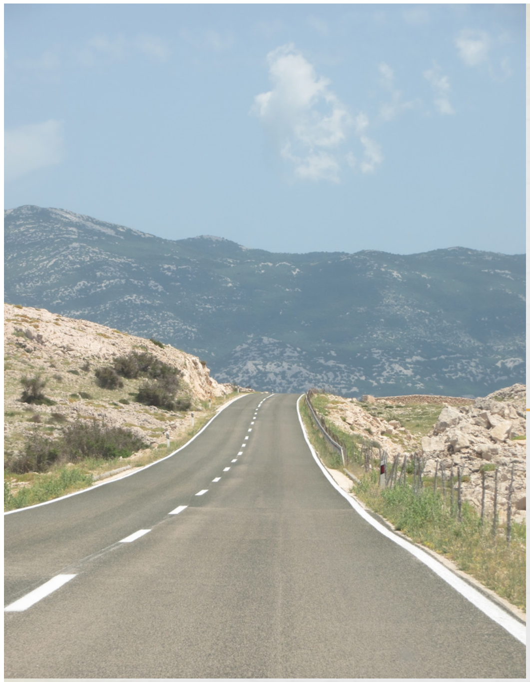
- Validating if the *accuracy* of the ML algorithm is worse with respect to certain subpopulations.

- MAP TO ETHICAL Pillars: **Fairness**
- MAP TO 7 trustworthy AI REQUIREMENTS **: Diversity, non-discrimination and fairness > Avoidance of unfair bias**

- IDENTIFY Ethical Tension: **Accuracy *versus* Fairness**
- Kinds of tension: **Practical dilemma**

Ethical Tension: **Accuracy** *versus* **Fairness**

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



Create Paths

- A *Path* P is created for investigating a subset of *Ethical Issues* and *Flags*
- *Ethical Issues* and *Flags* are associated areas of investigations (= 7 Trustworthy AI requirements)
- A Path can be composed of a number of steps

Run Paths

- *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 (using a log)
- Parts of a Path can be executed by different teams of inspectors with special expertise.

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)

Develop an evidence base

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

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

Source: Whittlestone, J et al (2019)

On Developing an evidence base

Our experience in practice (e.g. domain healthcare/ cardiology) suggests that this is a non obvious process.

For the same domain, there may be different point of views among “experts” of what constitutes a “neutral” and “not biased” evidence, and “who” is qualified to produce such evidence without being personally “biased”.

Do a Pre-Check

- At this point in some cases, it is already possible to come up with an initial ethical pre-assessment that considers the level of abstraction of the domain, with no need to go deeper into technical levels (i.e. considering the AI as a black box).
- This is a kind of pre-check, and depends on the domain.

Paths: verification (subset)

Verify Fairness
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

*Example: **Verify “fairness”***

Step 1. **Clarifying what kind of algorithmic “fairness” is most important for the domain** *(*)*

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.

Choosing Fairness criteria
(domain specific)

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

Define *Fairness* criteria, e.g.

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/

Fairness criteria and Machine Learning

- ***Equal patient outcomes*** refers to the assurance that protected groups have equal benefit in terms of patient outcomes from the deployment of machine-learning models
- ***Equal performance*** refers to the assurance that a model is equally accurate for patients in the protected and non protected groups.
- ***Equal allocation*** (also known as demographic parity), ensures that the resources are proportionately allocated to patients in the protected group.

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

From Domain Specific to ML metrics

Several Approaches in Machine Learning:

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) (average prediction for each group should be equal)	Independence
Equal coverage	Separation
No loss benefits	
Accurate coverage	
No worse off	
Equal of opportunity (EqOpt) (comparing the false positive rate from each group)	Separation
Equality of odds (comparing the false negative rate from each group)	Separation
Minimum accuracy	
Conditional equality,	Sufficiency
Maximum utility (MaxUtil)	

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

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)

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/

Use Case: Example of a Path

Path 1 Accuracy, Bias, Fairness, Discrimination

- This path mainly analysis accuracy, bias, fairness and discrimination. It also takes into account unfair bias avoidance, accessibility and universal design, stakeholder participation.

Final Execution Feedback:

- For the data sets used by the AI, a correlation with age was identified. The analysis of the data used for training, indicates that **there are more positive cases in certain age segments than others**, and this is probably the reason for a bias on age.

- **A higher accuracy prediction for male than female patients was identified**. The dataset is biased in having more male than female positive cases, and this could be the reason.

- **The size of the datasets for training and testing is small (below 1,000) and not well balanced** (wrt. gender, age, and with unknown ethnicity). This may increase the bias effects mentioned above.

- *Sensitivity* **was discovered to be lower than** *specificity*, i.e. not always detecting positive cases of heart attack risks.

"Explain" the feedback!

- The process continues by sharing the feedback of all Paths executed to the domain and ethics experts.

- Since the feedback of the execution of the various paths may be too technical-specific, it is useful to **"explain" the meaning** to the rest of the team (e.g. domain and ethical experts) who may not have prior knowledge of Machine Learning.

Re-asses Ethical Issues and Flags

- Execution of Paths may imply that Ethical issues and Flags are re-assessed and revised;
- The process reiterates from until a *stop* is reached.

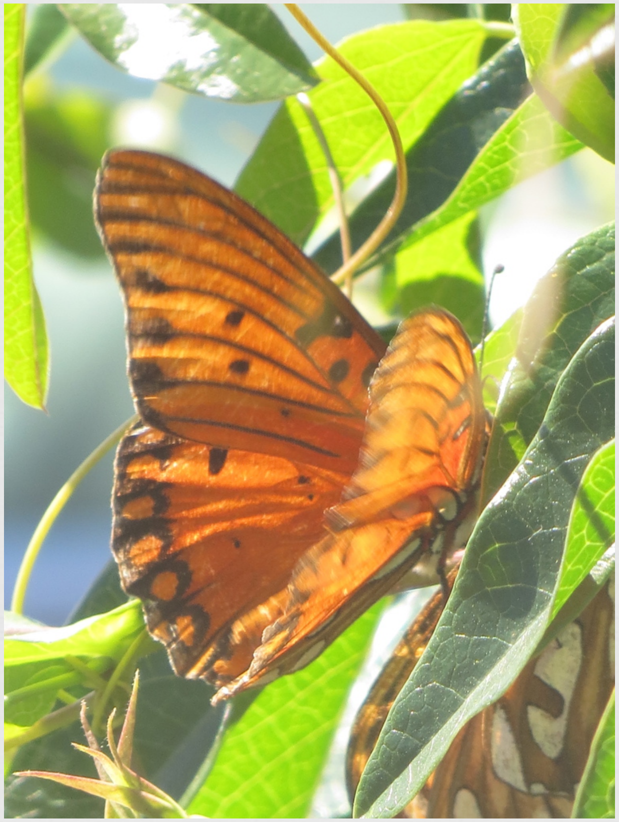
Classify Trade-offs

Iterative process.
A useful classification [1]:

- ***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.

[1] Source: Whittlestone, J et al (2019)

Resolve



Next Steps

- (Optional) Scores/Labels are defined;
- Address, Resolve Tensions;
- Recommendations are given;
- (Optional) Trade off decisions are made;
- (Optional) Ethical maintenance starts.

Use Case: Example of recommendations given to relevant stakeholders (simplified)

Accuracy, sensitivity and specificity deviate in part strongly from the published values and not sufficient medical evidence exists to support the claim that the device is accurate for all gender and ethnicity. This poses a risk of non-accurate prediction when using the device with patients of various ethnicities. There is no clear explanation on how the model is being medically validated when changed, and how the accuracy performance of the updated model compares to the previous model.

Example of Recommendations (cont.)

Recommendations :

- Continuously evaluate metrics with automated alerts.
- Consider a formal clinical trial design to assess patient outcomes.

Periodically collect feedback from clinicians and patients.

- An evaluation protocol should be established, and clearly explained to users.
- It is recommended that feature importance for decision making should be given, providing valuable feedback to the doctor to explain the reason of a decision to the model (healthy or not). At present, this is not provided, giving only the red/green/yellow flag with the confidence index.

Decide on Trade offs

- **Appropriate use:** Assess if the data and algorithm are appropriate to use for the purpose anticipated and perception of use.
  - Suppose we assess that the AI is technically *unbiased* and *fair* –this does not imply that it is acceptable to deploy it.

- **Remedies**: If risks are identified, define ways to mitigate risks (when possible)

- **Ability to redress**

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..

Resources

http://z-inspection.org