

# Ethical Implications of AI - series of lectures- Introduction



Course Coordinators:  
**Prof. Roberto V. Zicari,**  
**Prof. Heo, Seongwook,**

Assistants:  
**Jaewon Chang**  
**Dennis Vetter,**

*SNU Data Science, Seoul National University*

March 3, 2021

# Artificial Intelligence (AI)



*“Everything we love about civilization is a product of intelligence, so amplifying our human intelligence with artificial intelligence has the potential of helping civilization flourish like never before – as long as we manage to keep the technology beneficial.”*

**Max Tegmark**, President of the Future of Life Institute



# The Key question



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

# *Course Description*



## *i) Learn Basic Principles*

- ❧ Students will learn the **ethical implications of the use of Artificial Intelligence (AI)**. What are the consequences for society? For human beings / individuals? Does AI serve human kind?
- ❧ Discussion and debate of ethical issues is an essential part of professional development.



# *Course Description*



## *ii) Apply Principles in Practice*

- ❧ Students will learn how to assess an AI system in practice, identifying possible risks and “issues”.
- ❧ Students will work in small groups of two assessing real AI systems.

## Course Description (cont.)



∞ The course will offer a series of live and pre-recorded video lectures covering topics related to Ethics and AI:.

*Ethics, Moral Values, Humankind*

*Trust*

*Fairness/bias/discrimination;*

*Transparencies / Explainability*

*Privacy/ Responsibility/Accountability, Safety;*

*Human-in the loop;*

*Ethics AI in healthcare and other domains.*

**Live** Lessons: Legal Aspects (**Prof. Heo, Seongwook**)

# Ethical Implications of AI - series of lectures-



- ☞ Remote: Zoom video call
- ☞ Class schedule: **Mon. Wed. 18:00 ~ 19:15 Korean time (KST)**  
(10:00 am till 11:15 CET)
- ☞ Course starts: **March 3rd**, Ends: **Monday June 16.**  
**(16 weeks)** The language of the lectures is **English.**
- ☞ Course Web site:

<http://z-inspection.org/seul-national-university-ethical-implications-of-ai-series-of-lectures/>



# Course Web site



**<http://z-inspection.org/seul-national-university-ethical-implications-of-ai-series-of-lectures/>**

# Assignments



1. Create **teams of two students**;
  2. Each week ( See web site):
    - Watch 2 video lectures of the week;**
    - Read two papers of the week;** (or use some selected software recommended)
  4. Choose a **AI product/solution in healthcare**;
  5. Assess the AI system for “trustworthiness”
- ✎ -----> *trustworthy AI*, based on the EU “Framework for Trustworthy AI”, using the Z-inspection® process.

Z-inspection® is a registered trademark.

The content of this work 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/>)

# Assignments for this week (MARCH 3)



MARCH 3

✎ Intro Live Lesson 1 (Prof. Roberto V. Zicari)

Assignments of this week: **2 Papers to read**

**Paper 1: Whittlestone et al. (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research***

--

**Paper 2: Independent High-Level Expert Group on Artificial Intelligence (2019) – *Ethics Guidelines for Trustworthy AI***



# Assignments for next week (MARCH 8)



MARCH 8 Live Intro Lesson 2  
(Prof. Roberto V. Zicari)

Watch two pre-recorded video lectures:

MARCH 10 **The Ethics of Artificial Intelligence (AI)** (Prof. Roberto V. Zicari)

MARCH 15 **The Ethics of Artificial Intelligence (AI)** (Dr. Emmanuel Goffi)

Read 1 paper, Try a web tool:

**Paper: High-Level Expert Group on Artificial Intelligence (2020) – *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment***

---

**Web tool: ALTAI web tool**

# Mid Term and Final Reports



- ❧ Mid-term report due → May 10
- ❧ Min. 3 pages, Max. 5 pages
- ❧ Final Report
  - Continue the work of the Mid Term and create a Final Report
- ❧ Total MAX 10 pages (including the min 3 pages and max. 5 pages of the Mid term report)

# Mid-Term Report Requirements



- ❧ The goal of the mid-term report is to select an **AI system** (i.e. an AI-product and or an AI-based service) used in **healthcare**, and start with the evaluation process.
- ❧ In teams of **two students**:  
Choose a AI data-driven product/solution in healthcare, and assess for “trustworthiness” based on the EU Ethics Guidelines for Trustworthy AI, adapted to the healthcare domain and the Z-inspection® process.



# Z-inspection® Process in a Nutshell



## Mid Term Report

- ❧ Create an initial team of students;
- ❧ Boundaries and Context Define and agree upon the boundaries and context of the assessment;
- ❧ Analyze Socio-technical scenarios;
- ❧ Identify Ethical Issues and Tensions;

## Final Report

- ❧ Map to trustworthy AI;
- ❧ Execute (Verify Claims, Support Evidence);
- ❧ List Ethical Issues and Tensions (when possible resolve them);
- ❧ Give recommendations to relevant stakeholders.

<http://z-inspection.org/the-process-in-a-nutshell/>

# Mid-Term Report Requirements



- ❧ The report contributes to your grade and every team member will receive the same grade.
- ❧ The report must be delivered as a google doc (we will create one document per team).
- ❧ It must be min. 3-max. 5 pages long, including references and written with the google docs presets (i.e. Normal Text: Arial, 11pt).

# Mid-Term Report Requirements



The Mid Term report should cover the following:

- ❧ Define and agree upon the **boundaries and context** of the assessment.
- ❧ Analyze **Socio-technical scenarios**
- ❧ Identify **Ethical Issues and Tensions**
  
- ❧ In particular, the mid term report should relate to the topics covered by the lecture recordings and paper recommendations for up until the due date, as well as the ALTAI assessment list published by the EU and the ALTAI web tool.



# Mid-Term Report Requirements



Questions that need to be answered in the Mid term report are:

## **Analyze Socio-technical scenarios**

By collecting relevant resources, socio-technical scenarios should be created and analyzed by the team of students:

- ❧ Describe the aim of the AI system;
- ❧ Who are the actors;
- ❧ What are the actors expectations;
- ❧ How Actors interacts with each other and with the AI system;
- ❧ What are the process where the AI system is used;
- ❧ What AI technology is used;
- ❧ What is the context where the AI is used;
- ❧ What are any legal and contractual obligations related to the use of AI in this context
- ❧ And anything else you wish to be relevant.

# Mid-Term Report Requirements



## Identify Ethical Issues and Tensions

We use the term 'tension' as defined in [Whittlestone et al. 2019] :

„tensions between the pursuit of different values in technological applications rather than an abstract tension between the values themselves.“

Use the Catalog of Examples of Ethical 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

# Mid-Term Report Requirements



## Classify ethical tensions

according to the three dilemma defined in [Whittlestone et al. 2019] :

- ❧ – **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 dilemma** , i.e. “situations where there exists a third set of options beyond having to choose between two important values”.

# Remarks



**NEVER EVER COPY AND PASTE** text from the internet and other sources.

Two options:

1. You can describe what the source is saying using your words and quoting the source.

e.g As indicated in **[Roig and Vetter 2020]** the moon is flat. **Reference [Roig and Vetter 2020]** Why the Moon is Flat. Roig Gemma, Vetter Dennis Journal of Dreams, Issue No. 1, November 2020-

2. You can quote what the source is saying using their words in "... "

e.g **[Roig and Vetter 2020]** has made a case that "the moon is completely flat and not round". **Reference [Roig and Vetter 2020]** Why the Moon is Flat. Roig Gemma, Vetter Dennis Journal of Dreams, Issue No. 1, November 2020

❧ -Make sure to **READ** the source when you use them to make sure you understand the context! In the example above, if you quote that "the moon is flat" without understanding that this was a dream and not a scientific evidence, you are using a quote in a **WRONG** way!



# High-Level Expert Group on Artificial Intelligence (AI HLEG)



- ❧ In 2019 the High-Level Expert Group on Artificial Intelligence (AI HLEG), set up by the European Commission, published the Ethics Guidelines for Trustworthy Artificial Intelligence.
- ❧ The third chapter of those Guidelines contained an Assessment List to help assess whether the AI system that is being developed, deployed, procured or used, adheres to the **seven requirements of Trustworthy Artificial Intelligence (AI)**, as specified in our Ethics Guidelines for Trustworthy AI

# ALTAI WEB SITE



- ❧ This website contains the Assessment List for Trustworthy AI (ALTAI).
- ❧ ALTAI was developed by the High-Level Expert Group on Artificial Intelligence set up by the European Commission to help assess whether the AI system that is being developed, deployed, procured or used, complies with the seven requirements of Trustworthy AI, as specified in our Ethics Guidelines for Trustworthy AI.
- ❧ <https://altai.insight-centre.org>

# How to complete ALTAI



- ✧ For each question ALTAI provides **guidance in the glossary** and by **referencing to the relevant parts of the Ethics Guidelines for Trustworthy AI** and examples in text boxes alongside the questions.

Upon completing ALTAI, the following will be generated:

- ✧ **A visualisation** of the self-assessed level of adherence of the AI system and its use with the 7 requirements for Trustworthy AI. These results are based on your organisation's own assessment and are solely meant to help you identify the areas of improvement.
- ✧ **Recommendations** based on the answers to particular questions.

# Register to the ALTAI Online Tool.



- ❧ You need to create an account to use the Assessment List for Trustworthy AI Online Tool. This allows you to save and edit ALTAs.
- ❧ <https://altai.insight-centre.org/Identity/Account/Register>



# ALTAI web tool



- ❧ The self-assessment results and the list of recommendations are confidential and available to you only.
- ❧ ALTAI has seven sections, corresponding to the 7 requirements for Trustworthy AI. You can move between the sections by clicking on the side menu.

## Colour code

Questions with **white background** are aimed at (describing) the features of the AI system.

Answers to **blue questions** will contribute to recommendations.

**Text in red** allows you to self-assess your organisation's compliance with the respective requirement.

# Fundamental Rights



✧ Fundamental rights encompass rights such as human dignity and non-discrimination, as well as rights in relation to data protection and privacy, to name just some examples.

# Fundamental rights impact assessment



- ❧ Prior to self-assessing an AI system with this Assessment List, a fundamental rights impact assessment (FRIA) should be performed.
- ❧ A FRIA could include questions such as the following – drawing on specific articles in the Charter and the European Convention on Human Rights (ECHR) its protocols and the European Social Charter.
- ❧ <https://altai.insight-centre.org/Home/FundamentalRights>

# Fundamental rights impact assessment



❧ Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively):

sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?



# Fundamental rights impact assessment



- ❧ Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds
- ❧ Have you put in place processes to test and monitor for potential discrimination (bias) during the development, deployment and use phase of the AI system?
- ❧ Have you put in place processes to address and rectify for potential discrimination (bias) in the AI system?

# Fundamental rights impact assessment



- ❧ Does the AI system respect the rights of the child, for example with respect to child protection and taking the child's best interests into account?
- ❧ Have you put in place processes to test and monitor for potential harm to children during the development, deployment and use phase of the AI system?
- ❧ Have you put in place processes to address and rectify for potential harm to children by the AI system?

# Fundamental rights impact assessment



- ❧ **Does the AI system protect the right to privacy, including personal data relating to individuals in line with GDPR?**
- ❧ Have you put in place processes to assess in detail the need for a data protection impact assessment, including an assessment of the necessity and proportionality of the processing operations in relation to their purpose, with respect to the development, deployment and use phases of the AI system?
- ❧ Have you put in place measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data with respect to the development, deployment and use phases of the AI system?
- ❧ See the section on Privacy and Data Governance in this Assessment List, and available guidance from the European Data Protection Supervisor.

# Fundamental rights impact assessment



- ❧ Does the AI system respect the freedom of expression or assembly?
- ❧ Could the AI-system potentially limit a person's
- ❧ freedom to openly express an opinion, partake in a peaceful demonstration or join a union?



# Foundations. The European View Framework for Trustworthy AI

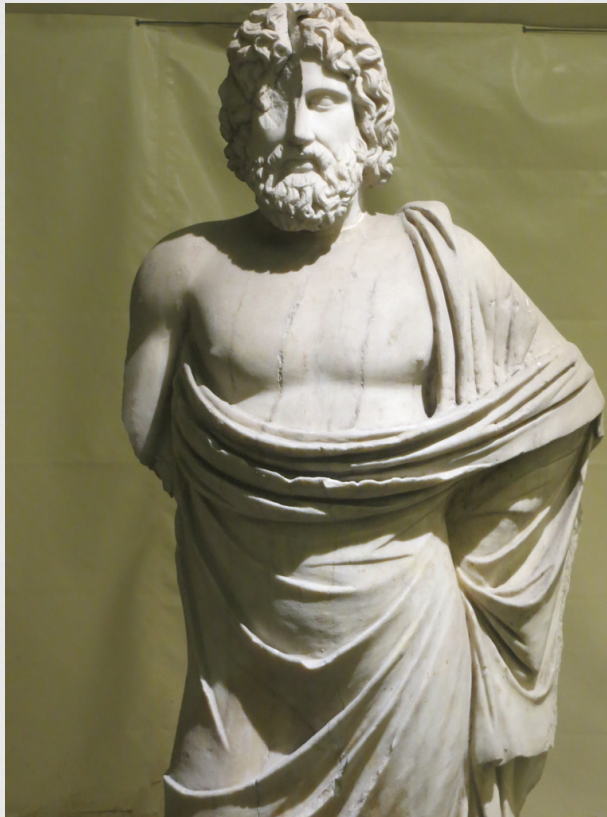


Photo RVZ

# 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.

# What is Ethics?



- ❧ A branch of philosophy
  - ❧ ➡ An appraisal of what is right (good) and wrong (bad)
  - ❧ ➡ An assessment of human actions

❧ Source: **The Ethics of Artificial Intelligence (AI)** (Dr. Emmanuel Goffi)

[http://www.bigdata.uni-frankfurt.de/wp-content/uploads/2020/01/Goffi-2020-The-Ethics-of-Artificial-Intelligence-AI\\_Ethical-Implications-of-AI-SS2020.pdf](http://www.bigdata.uni-frankfurt.de/wp-content/uploads/2020/01/Goffi-2020-The-Ethics-of-Artificial-Intelligence-AI_Ethical-Implications-of-AI-SS2020.pdf)

# Framework for Trustworthy AI


## Four Ethical Principles.



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.



# *The principle of respect for human autonomy*



- ❧ “The fundamental rights upon which the EU is founded are directed towards ensuring **respect for the freedom and autonomy of human beings.**
- ❧ Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in **the democratic process.** “

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

## *The principle of respect for human autonomy (cont.)*



- ❧ “ AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans.
- ❧ Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and **leave meaningful opportunity for human choice.**
- ❧ This means **securing human oversight** over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work. “

# *The principle of prevention of harm*

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

---

- ❧ “AI systems **should neither cause nor exacerbate harm or otherwise adversely affect human beings.**
- ❧ This entails the **protection of human dignity as well as mental and physical integrity.** AI systems and the environments in which they operate must be safe and secure. They must be technically robust and it should be ensured that they are not open to malicious use.”

## *The principle of prevention of harm (cont.)*



- ❧ “ **Vulnerable persons** should receive greater attention and be included in the development, deployment and use of AI systems.
- ❧ Particular attention must also be paid to situations where AI systems can **cause or exacerbate adverse impacts due to asymmetries of power or information**, such as between employers and employees, businesses and consumers or governments and citizens.
- ❧ Preventing harm also entails consideration of the **natural environment and all living beings.**”



# *The principle of fairness*

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

---



- ❧ **“The development, deployment and use of AI systems must be fair.** While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension.
- ❧ The substantive dimension **implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.**
- ❧ **If unfair biases can be avoided, AI systems could even increase societal fairness.** Equal opportunity in terms of access to education, goods, services and technology should also be fostered.
- ❧ **Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice.** Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives. “

## *The principle of fairness (cont.)*



- ❧ **“The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.**
- ❧ **Fairness is closely linked to the rights to Non-discrimination, Solidarity and Justice”**

# *The principle of explicability* (

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

April, 2019



✧ **Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested.**

## *The principle of explicability (cont.)*



- ❧ An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as **'black box' algorithms** and **require special attention**.
- ❧ In those circumstances, **other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities)** may be required, provided that the system as a whole respects fundamental rights.
- ❧ The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.



# Tensions and Trade-offs



- ✧ „**Tensions** may arise between the above ethical principles, **for which there is no fixed solution.**
- ✧ In line with the EU fundamental commitment to democratic engagement, due process and open political participation, methods of accountable deliberation to deal with such tensions should be established. “

## Tensions and Trade-offs (cont.)



- ❧ “For instance, in various application domains, *the principle of prevention of harm* and *the principle of human autonomy* may be in conflict. Consider as an example the use of AI systems for ‘predictive policing’, which may help to reduce crime, but in ways that entail surveillance activities that impinge on individual liberty and privacy.
- ❧ Furthermore, AI systems’ overall benefits should substantially exceed the foreseeable individual risks. While the above principles certainly offer guidance towards solutions, they remain abstract ethical prescriptions.
- ❧ **AI practitioners can hence not be expected to find the right solution based on the principles above, yet they should approach ethical dilemmas and trade-offs via reasoned, evidence-based reflection rather than intuition or random discretion.**
- ❧ There may be situations, however, where no ethically acceptable trade-offs can be identified. Certain fundamental rights and correlated “

# Requirements of Trustworthy AI Including *Sub-requirements*



## **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 Including *Sub-requirements*



## **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.



# Human agency and oversight



☞ All potential impacts that AI systems may have on fundamental rights should be accounted for and that the human role in the decision- making process is protected.

☞ Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.

# Technical robustness and safety



✧ AI systems should be secure and resilient in their operation in a way that minimizes potential harm, optimizes accuracy, and fosters confidence in their reliability;

✧ Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to

Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.

# Privacy and data governance



Given the vast quantities of data processed by AI systems, this principle impresses the importance of protecting the privacy, integrity, and quality of the data and protects human rights of access to it;

Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.

# Transparency



✧ AI systems need to be understandable at a human level so that decisions made through AI can be traced back to their underlying data. If a decision cannot be explained it cannot easily be justified;

✧ Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.



# Diversity, non-discrimination, and fairness



✧ AI systems need to be inclusive and non- biased in their application. This is challenging when the data is not reflective of all the potential stakeholders of an AI system;

✧ Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications

# Societal and environmental wellbeing



✧ In acknowledging the potential power of AI systems, this principle emphasizes the need for wider social concerns, including the environment, democracy, and individuals to be taken into account;

✧ Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications

# Accountability



- ✧ This principle, rooted in fairness, seeks to ensure clear lines of responsibility and accountability for the outcomes of AI systems, mechanisms for addressing trade-offs, and an environment in which concerns can be raised. However, the interpretation, relevance, and implementation of trustworthy AI depends on the domain and the context where the AI system is used.

✧ Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications

# Challenges and Limitations



✎ Although these requirements are a welcome first step towards enabling an assessment of the societal implication of the use of AI systems, there are some challenges in the practical application of requirements, namely:

The AI HLEG trustworthy AI guidelines are not contextualized by the domain they are involved in. The meaning of some of the seven requirements is not anchored to the context (e.g., fairness, wellbeing, etc.).

They mainly offer a static checklist (AI HLEG 2020) and do not take into account changes of the AI over time.

They do not distinguish different applicability of the AI HLEG trustworthy AI guidelines (e.g., during design vs. after production) as well as different stages of algorithmic development, starting from business and use-case development, design phase, training data procurement, building, testing, deployment, and monitoring (Morley et al., 2019).

There are not available best practices to show how to implement such requirements and apply them in practice.

The AI HLEG trustworthy AI guidelines do not explicitly address the lawful part of the assessment.

Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications



# A Process for Trustworthy AI Assessment



photo CZ

# 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.

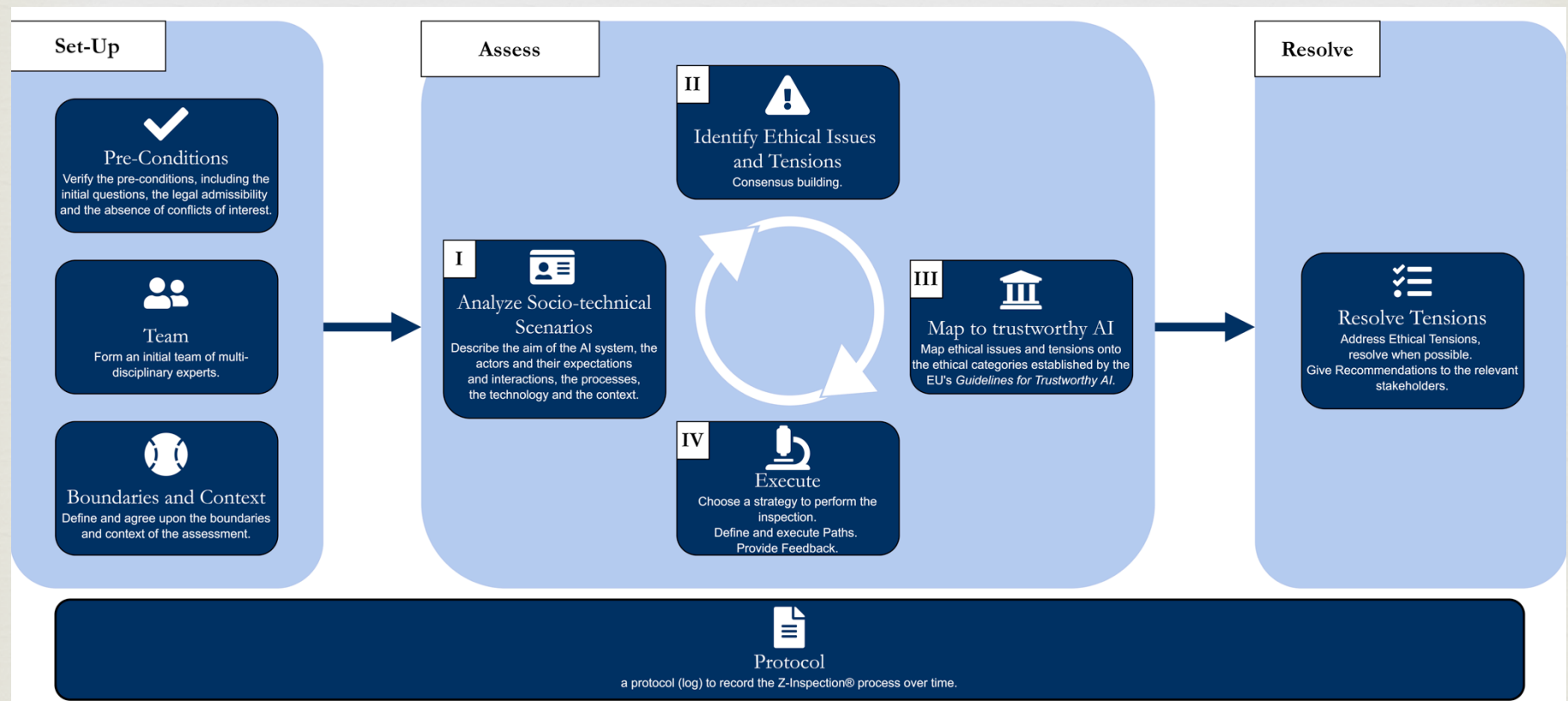


## Orchestration Process



- ❧ 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 propertyDefine 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. Nyrupe, 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.



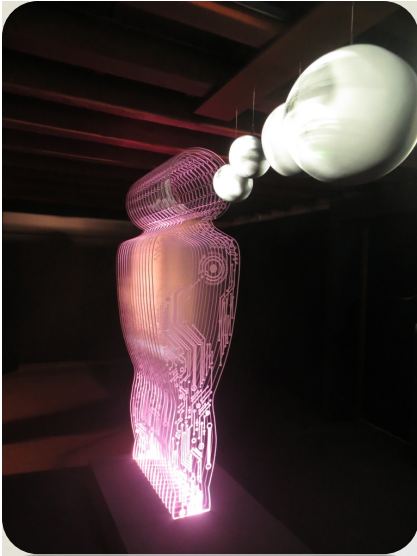
- ✧ When designing, training and testing an AI-system (e.g. Machine-Learning algorithm) we do “embed” into the system notions such as “good”, “bad”, “healthy”, “disease”, etc. mostly not in an explicit way.

# “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



## 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*

# 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

## *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



Making *trade-offs* between values:  
Choosing to prioritize one value at the expense of another.

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>