

Mindful Use of AI
Z-inspection®
**A holistic and analytic process to
assess Ethical AI**



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

October 15, 2020

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/>)

Z-inspection® : Core Team Members



Roberto V. Zicari (1), John Brodersen (4)(9), James Brusseau (8), Boris Döder (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, Denmark; (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

Ethical and societal implications of Artificial Intelligence systems



“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

Status quo

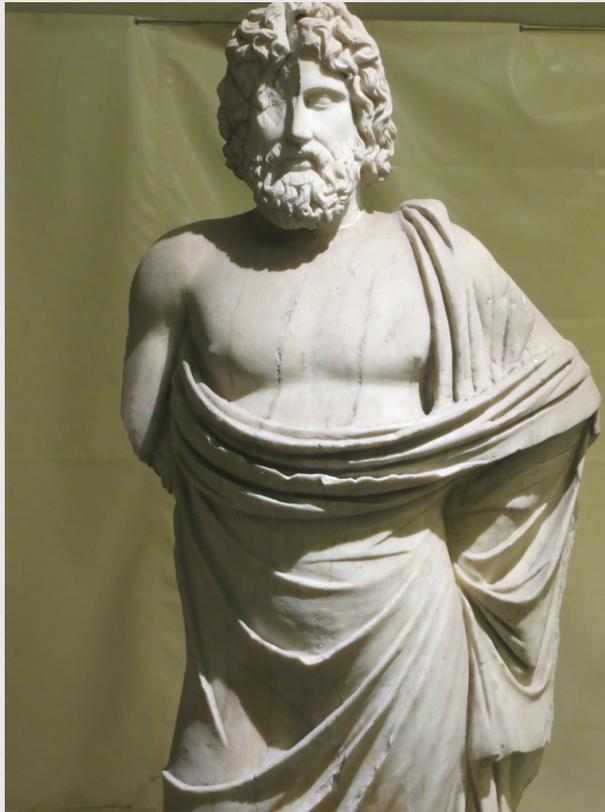


Photo RVZ

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

German Data Ethics Commission



- ❧ Established in mid 2018 with the mission to develop, within one year, an ethical and regulatory framework for data, ADM and AI
- ❧ Co-chaired by Christiane Wendehorst and Christiane Woopen
- ❧ Opinion presented in Berlin on 23 October 2019
- ❧ Includes ethical guidelines and 75 concrete recommendations for action regarding data and algorithmic systems

❧ Source: **Opinion of the German Data Ethics Commission** Christiane Wendehorst
Co-Chair of the Data Ethics Commission

❧ http://www.bigdata.uni-frankfurt.de/wp-content/uploads/2020/01/CWe-Presentation_Zicari.pdf

Data-driven technologies (including AI)



- ❧ Ethics of handling personal data
- ❧ Ethics of handling data in general (including non-personal data)
- ❧ **Ethics of handling data and data-driven technologies (including algorithmic systems, such as AI)**
- ❧ **Ethics of the digital transformation in general (including issues such as the platform economy or the future of work)**

❧ Source: **Opinion of the German Data Ethics Commission** Christiane Wendehorst
Co-Chair of the Data Ethics Commission

❧ http://www.bigdata.uni-frankfurt.de/wp-content/uploads/2020/01/CWe-Presentation_Zicari.pdf

General ethical principles



- ❧ Human dignity
- ❧ Autonomy
- ❧ Privacy
- ❧ Security
- ❧ Democracy
- ❧ Justice and Solidarity
- ❧ Sustainability

❧ Source: **Opinion of the German Data Ethics Commission** Christiane Wendehorst
Co-Chair of the Data Ethics Commission

❧ http://www.bigdata.uni-frankfurt.de/wp-content/uploads/2020/01/CWe-Presentation_Zicari.pdf

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

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.

Conceptual clusters



- Bias/Fairness/discrimination
- Transparencies/Explainability/ intelligibility/interpretability
- Privacy/ responsibility/ Accountability
- Safety
- Human-AI
- Uphold human rights and values;
- Promote collaboration;
- Acknowledge legal and policy implications;
- Avoid concentrations of power,
- Contemplate implications for employment.

The Gap



✧ “Putting ethical 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”.

Source: *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.

Z-inspection®: *Holistic and Analytic*



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 Trustworthy 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 to assess Trustworthy AI **in practice**.

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



Our approach is learning by doing.

We developed Z-inspection® by assessing an *AI-based medical device for enhancing decision-making (cardiology)*.

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), Elisabeth Hildt (17), Ludwig Christian Hinske (24), Sune Holm (25), Todor Ivanov (1), Georgios Kararigas (3), Pedro Kringen (1), Vince Madai (27), Melissa McCullough (1), Carl-Maria Mörch (12), Florian Möslein (7), Naveed Mushtaq (1), David Ottenheimer (20), Matiss Ozols (14), Laura Palazzani (10), Eberhard Schnebel (1), Andy Spezzati (11), Gemma Roig (1), 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) 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, Denmark;
- (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.



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.

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).

☞ 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.

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*



photo RVZ

Who? Why? Whom?



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®: 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

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.

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.

“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

On AI Ethics Scores (Labeling)



Scoring denotes the assignment of a numerical value (a score) to an AI-based software for the purpose of evaluating certain areas of investigation.

Scoring may have different meaning, depending *when* and *why* and by *whom* 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.

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.

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.

AI and the Contexts



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.

Z-inspection® *Layered Reference Model*

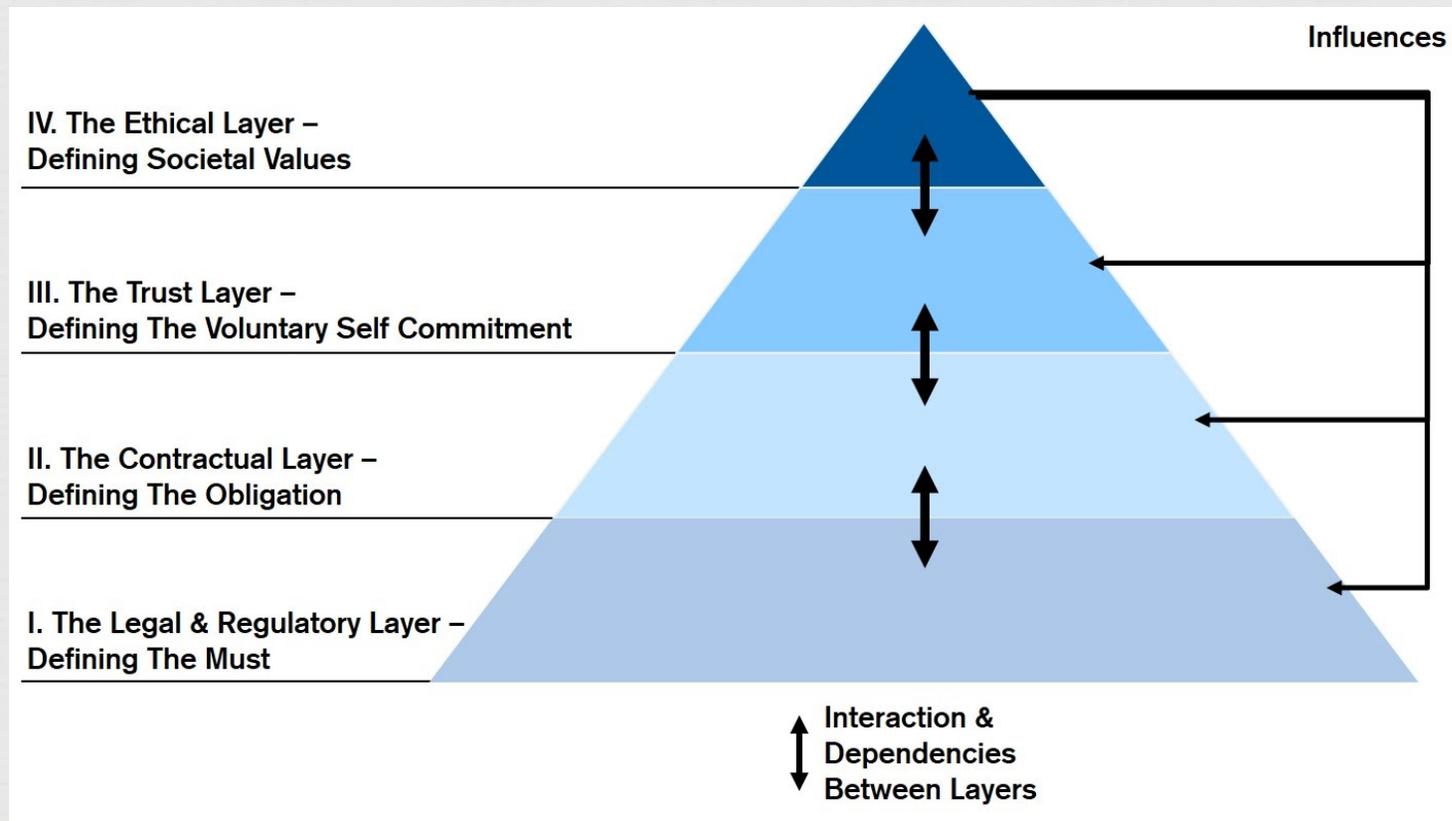


Z-inspection® *Layered Reference Model*



- ✧ The Z-Inspection reference model fulfills the need to establish a framework that helps us identify where ethical related actions are located and whether certain ethical issues and actions are part of a more general, broad value system, legally or regulatory prescribed or indeed part of a specific contractual obligation.

Z-inspection® Layered Reference Model



Z-inspection® *Layered Model:*

Legal, Contractual



I. AI Legal/Regulatory Must Layer

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



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.

Z-inspection® *Layered Model: Ethical*



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.

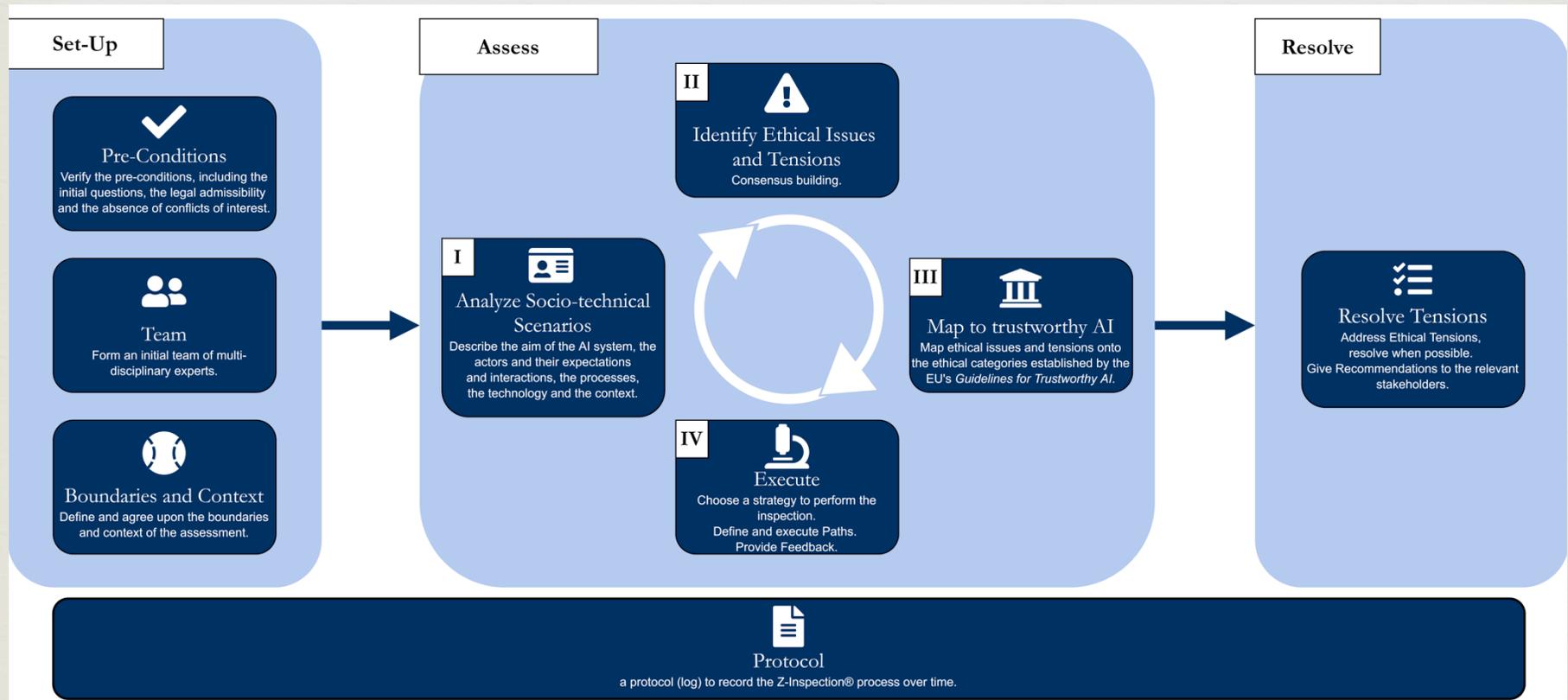
It is a layer which is not mandatory but where ethical principles are discussed and as such implicitly made a goal post to orient at.

Z-inspection® Methodology



Photo RVZ

Z-inspection® Process in a Nutshell





Build a Team



A. The Pre-conditions are verified;

B. 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 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.



Define the time-frame



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 challenges we are already aware of and already facing today?”*
- ❧ Near-future challenges: *“What challenges might we face in the near future, assuming current technology?”*
- ❧ Long-run challenges: *“What 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.



Use Socio-technical scenarios



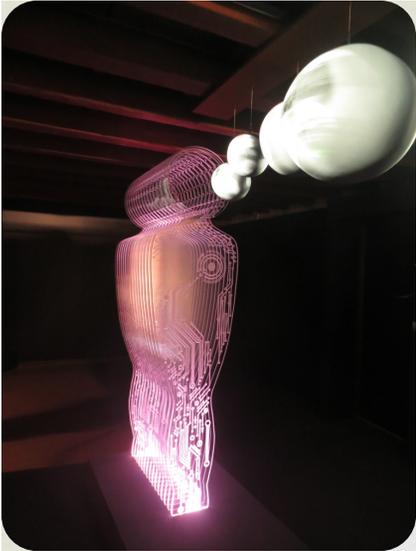
- ❧ Socio-technical scenarios are created (or given to) by the team of experts to represent possible scenarios of use of the AI. This is a process per se, that involves several iterations among the experts, including using *Concept Building*.

Socio-technical 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)



Identify Ethical Issues



- ❧ As a result of the analysis of the scenarios, **Ethical issues**, $E_1 \dots E_i$, and **Flags**, $F_1 \dots F_j$ 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.

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.



Describe Ethical issues.



- ☞ 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.

Identify, Classify and Describe Tensions



This is part of the iterative process among experts with different skills and background.

Catalog of 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)

Ambiguity



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.

Concept Building



- ❧ *“The goal is building a shared understanding of key concepts that acknowledges and resolves ambiguities, and bridges disciplines, sectors, publics and cultures. Scenarios are dependent on the domain”. [1]*
- ❧ In our experience, concept building in practice is a complex iterative process!
- ❧ Need of a clear moderation, and setting of common “goals” to reach, revised at each iteration.

❧ 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.

Why using Concept Building?



Defined by Whittlestone, J et al [1], *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. The choice of experts has ethical implications!

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.



Map Ethical issues and Flags onto Trustworthy AI indicators.

- ❧ 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*.

Identify gaps and map conceptual concepts at different levels



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.

Case Study

Illustration of Ethical Issues and Mappings to the Areas of Investigation.



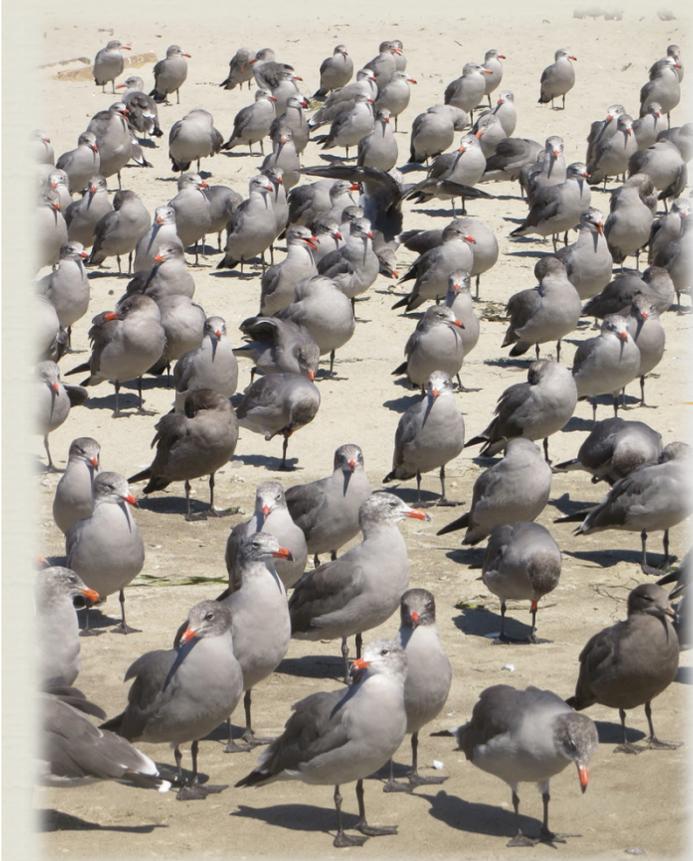
- ❧ Ethical "Issue: E1
- ❧ Description: When the AI is being used in screening asymptomatic people who are "*notified*" with a "minor" CAD problem that might not impact their lives, they might get worried- change their lifestyles after the *notification* even though this would not be necessary.
- ❧ MAP TO 4 ETHICAL Pillars: Respect for human autonomy
- ❧ MAP TO 7 trustworthy AI REQUIREMENTS: Human agency and oversight > Human agency and Autonomy
- ❧ MAP TO 4 ETHICAL Pillars: Prevention of Harm
- ❧ MAP TO 7 trustworthy AI REQUIREMENTS: Technical Robust and Safety > Accuracy
- ❧ Ethical Tensions: N/A

Case Study

Illustration of Ethical Issues and Mappings to the Areas of Investigation.



- ❧ Ethical "Issue": E2
- ❧ Description: If due to the AI test more patients with minor CAD problems are being “notified” and sent to cardiologists, this might result in significant increase of unnecessary health care costs for society, due to further diagnostics tests.
- ❧ MAP TO 4 ETHICAL Pillars: Prevention of Harm
- ❧ MAP TO 7 trustworthy AI REQUIREMENTS: Societal and environmental wellbeing > Impact on Society at large.
Acknowledge Legal and Policy implications
- ❧ Ethical Tensions: Human agency (individual rights) vs. social wellbeing (social welfare)
- ❧ Kinds of tensions: True Dilemma

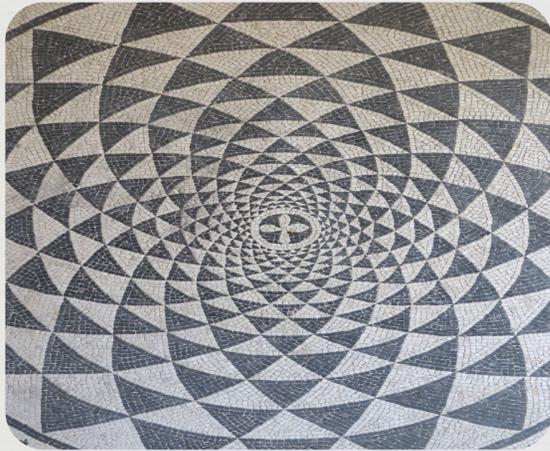
A large group of seagulls, likely Mediterranean Gulls, are gathered on a sandy beach. They are mostly grey with white heads and necks, and some have distinctive red beaks. The birds are scattered across the frame, some standing in small groups and others alone. The background is a bright, sandy beach under a clear sky.

Choose 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

Set the initial level of abstraction
(Macro vs Micro).



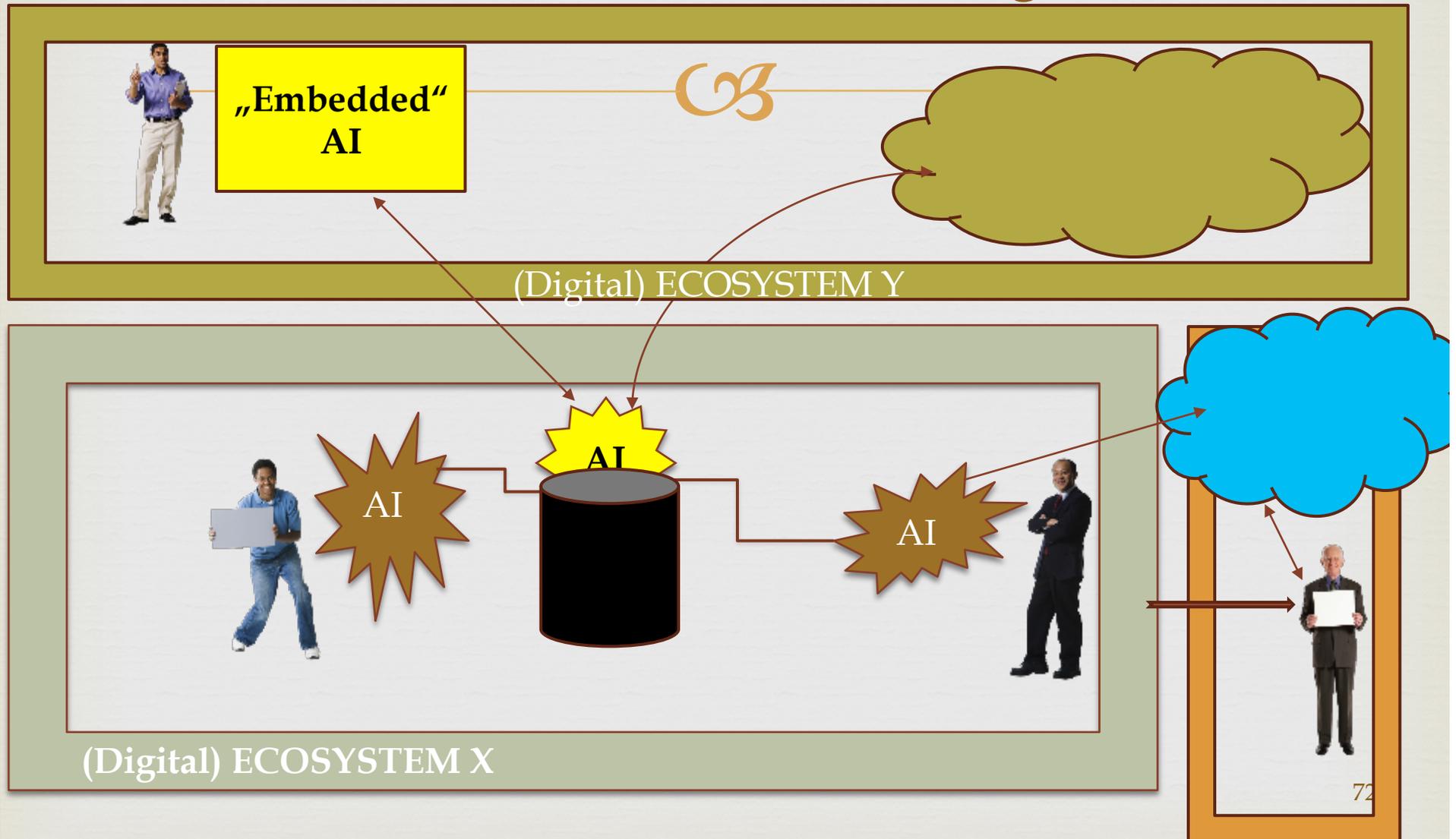


Create Layers



- ❧ 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.

Ethical AI "Macro"-Investigation



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

Ethical AI "Micro"-Investigation



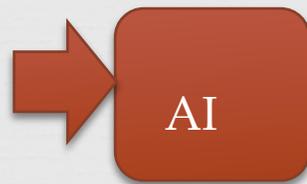
Context
Culture
People/Company Values



Feedback



People
+
Algorithms
+
Data



"Good"

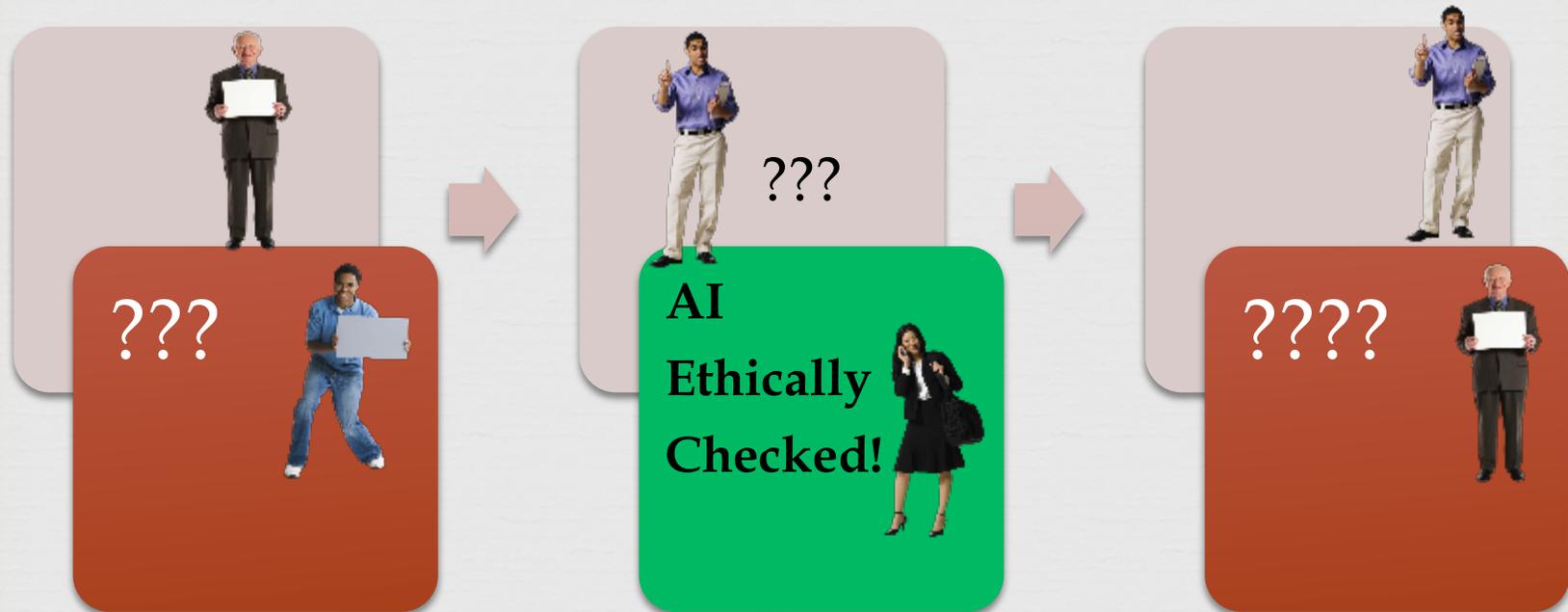


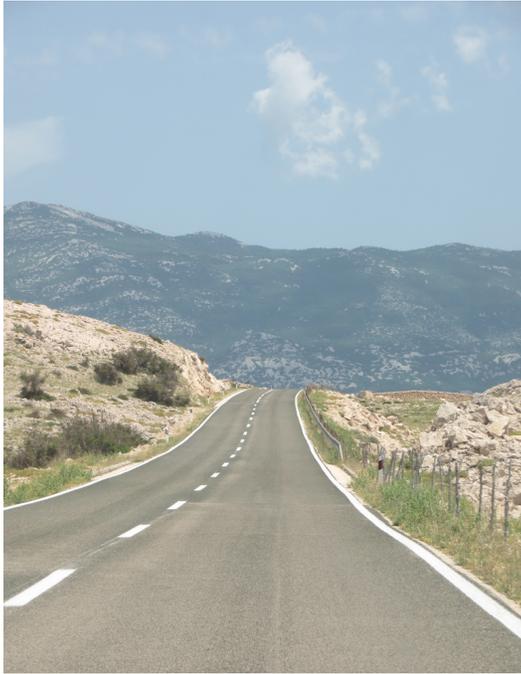
"Bad"

???



Micro-validation does not imply Macro-validation





Create Paths



- ✧ A *Path* P is created for investigating a subset of *Ethical Issues* E_i and *Flags* F_j
- ✧ E_i and F_j each, are associated to a cluster area C of investigations
- ✧ 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.

Z-inspection 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: Assessing 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.

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)

Incompatible types of fairness



Known Trade Offs (Incompatible types of fairness):

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

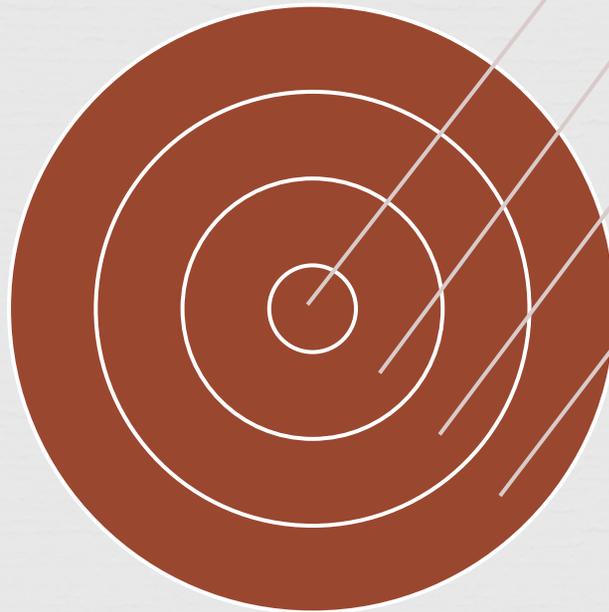
Which type of fairness is appropriate for the given application and what level of it is satisfactory?

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

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

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

Example: Iterative Inside Out Approach



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

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

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

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

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 beginning (The boundaries and context of the assessment are agreed upon and defined) till end (Ethical issues and Flags are re-assessed) , until a *stop* is reached.



Classify Trade-offs



Iterative process.

A useful classification [1]:



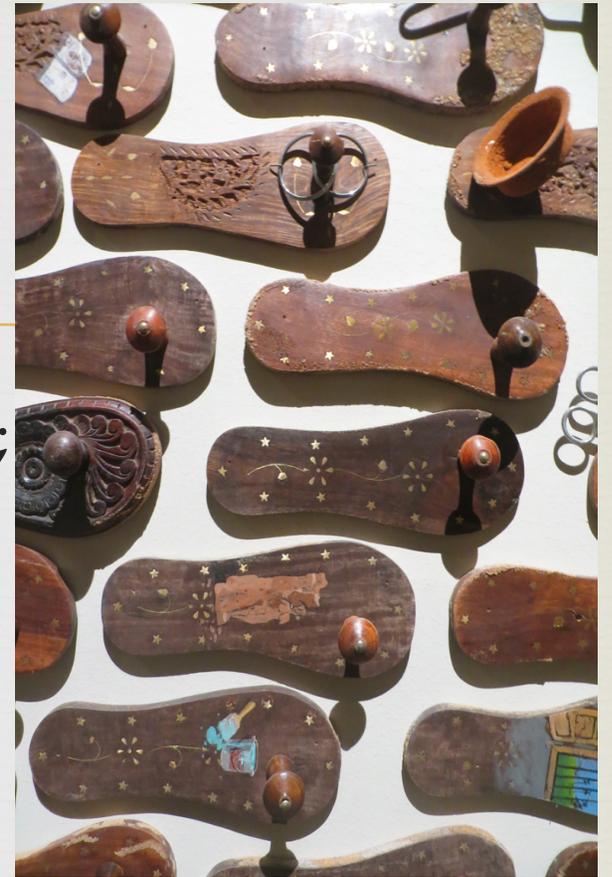
- ❧ *True ethical dilemma* - the conflict is inherent in the very nature of the values in question and hence cannot be avoided by clever practical solutions.
- ❧ *Dilemma in practice*- the tension exists not inherently, but due to our current technological capabilities and constraints, including the 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)

Next Steps



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





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>

Acknowledgements



Many thanks to

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

for providing valuable comments and feedback.

All photos: Roberto V. Zicari, Carla Zicari