Assessing Trustworthiness of AI Systems using Z-Inspection®

http://Z-Inspection(R).org

IEEE CAI 2025 IEEE Conference on Artificial Intelligence

Santa Clara, California, USA, May 5, 2025

Organizer

Jesmin Jahan Tithi, AI Research Scientist, Parallel Computing Labs, Intel.

Founding member, Head of Education for North America, Certified Teacher of Z-Inspection(R)®, Part of Z-Inspection(R)® Initiative.

Organizer

Partha Deka, Senior Staff Engineer, Intel

Part of Z-Inspection(R)® Initiative.

Goal of This Tutorial

- Understand Ethical Principals and Requirements of Trustworthy AI.
- Learn the Z-Inspection® framework for ethical AI assessment.
- Gain practical skills in using Z-Inspection® for ethical AI assessment through examples.
- Receive a Z-Inspection® Certificate of Participation in the end and link to the networks.
- Target Audiences: Students, Professionals designing and deploying AI solutions, experts interested in AI ethics.

Learning Z-Inspection® would empower to the creation and deployment of AI systems that are not only powerful but also trustworthy, responsible, reliable, and beneficial to society if practiced!.

Disclaimer

Generated by designer

- My primary research focus at Intel is HPC, Parallel Algorithms and Software Hardware co-design which is unrelated to Z-Inspection(R)®!
- I worked on Z-Inspection(R)® to pursue my interest in Ethical AI!
- Thus, the views expressed in this presentation are my own and do not reflect the official policy or position of Intel Corporation.

We are the Z-Inspection® Initiative

- The Z-Inspection® Initiative is a *non-organized organization…*

  https://z-inspection.org

- We have *affiliated* partners all over the world:
  - 28 affiliated Trustworthy AI Labs
  - 18 affiliated Institutions

  https://z-inspection.org/affiliated-labs/

The Mission of the Z-Inspection® Initiative

To help establishing

Mindful Use of AI (#MUAI).

What is the Definition of Trust?

Trustworthy AI According to EU High-Level Expert Group

1. **lawful** - respecting all applicable laws and regulations
2. **robust** - both from a technical and social perspective
3. **ethical** - respecting ethical principles and values

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

Ethical Principles behind Trustworthy AI

Four ethical principles, rooted in fundamental rights

1. **Respect for human autonomy**
2. **Prevention of harm**
3. **Fairness**
4. **Explicability**

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** - *fundamental rights, human in control*

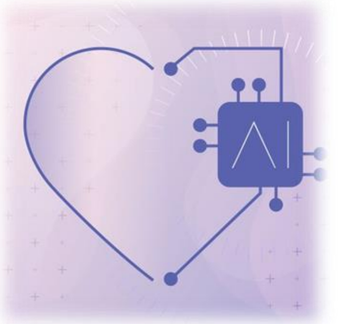
2. **Technical robustness and safety** - *resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility*

3. **Privacy and data governance** - *respect for privacy, quality and integrity of data, and access to data*

4. **Transparency** - *interpretability, traceability, explainability and communication*

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

Requirements of Trustworthy AI



5. **Diversity, non-discrimination and fairness** - *avoidance of unfair bias, accessibility and universal design, and stakeholder participation*

6. **Societal and environmental wellbeing** - *sustainability and environmental friendliness, impact on society and democracy*

7. **Accountability** - *auditability, minimization 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

These requirements need to be continuously evaluated and addressed throughout the AI system’s life cycle

Why Do We Need Inspection or Auditing Process?

Self-Assessment is welcome, however is often cursed by conflict by interest

Evaluation of Ethical Impact is Challenging.

"Ethical impact evaluation involves evaluating the ethical impact 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."

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

Some Ethical Impact Are only Fully Understood after Deployment.

"The real-life ethical impact that a technology will have on people, their communities and the planet, can only be fully understood once the product or service is in real-world use."

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

Claims of BlackBox models are Difficult to Verify.

- "The process of AI development is often opaque to those outside a given organization, and various barriers make it challenging for third parties to verify the claims being made by a developer."

- "A coalition of stakeholders should create a taskforce to research options for conducting and funding third party auditing of AI systems"

Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, Miles Brundage Thomas Krendl ,Yoshua Bengio et al., April 20, 2020

Need for Third Party Auditors

"Develop auditing mechanism for AI systems to allow public enforcement authorities as well as independent third-party auditors to identify potentially illegal outcomes or harmful consequences generated by AI systems, such as unfair bias or discrimination"

White Paper of the European Commission [72] and a report of The Organization for Economic Co-operation and Development (OECD) [25]


Z-Inspection®

Z-Inspection® – A **Practical** Assessment Process for Trustworthy AI

- A process based on applied ethics

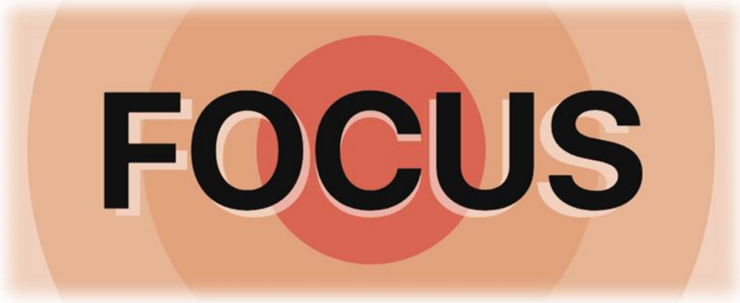
- Can be applied to a variety of domains
  - business, healthcare, public sector, many others

Practical

- Customizable based on use case, domains and contexts

- Allows to use existing frameworks, checklists, tools as plug ins

Focus of Z-Inspection(R)®



Z-Inspection(R)® covers the following:

- Technical robustness - **robust**
- Legal/Contractual implications – **lawful**
- Ethical and Societal implications - **ethical**

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

Note2: *Legal and Ethics depend on the context*

Catalogue of Tools & Metrics for Trustworthy AI

These tools and metrics are designed to help AI actors develop and use trustworthy AI systems and applications that respect human rights and are fair, transparent, explainable, robust, secure and safe.

Z-Inspection

Website | Tool package

Procedural

Uploaded on Jun 10, 2022

Organisation(s): Z-Inspection Initiative

> *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, among many others.*

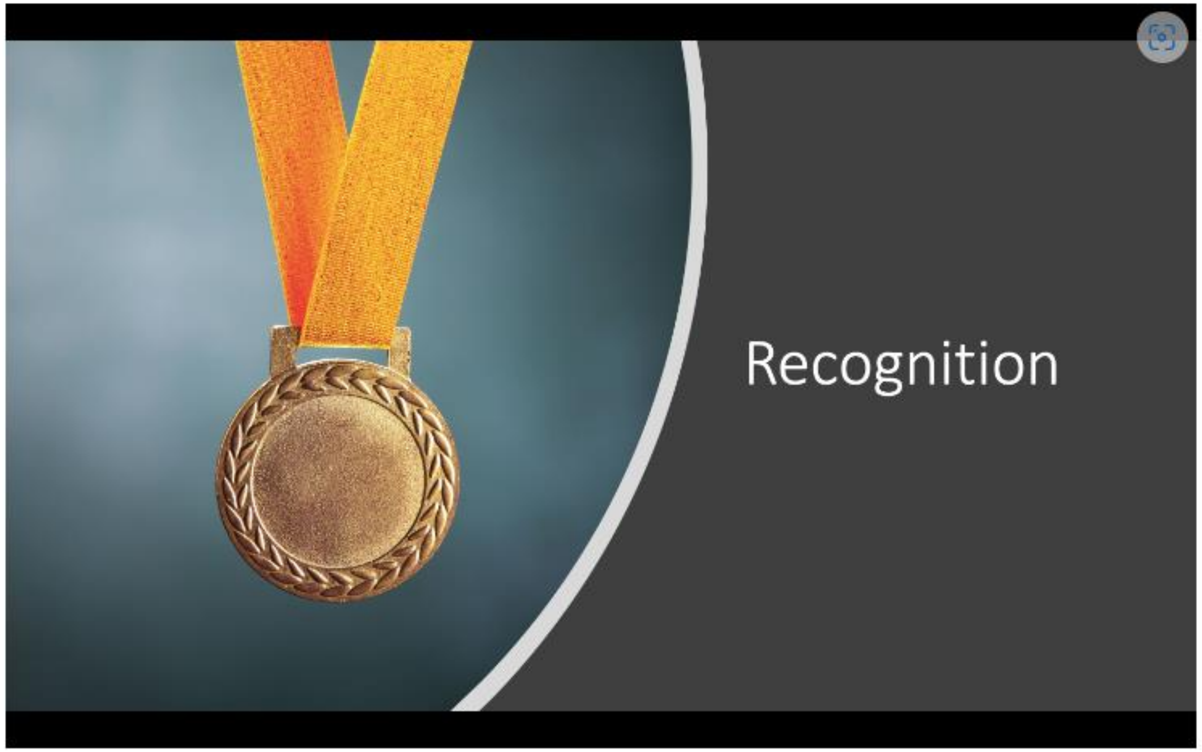
It uses applied ethics to assess Trustworthy AI in practice.

The Z-Inspection® process has the potential to play a key role in the context of the new EU Artificial Intelligence (AI) regulation.

About the tool

You can click on the links to see the associated tools

Tool type(s):	Audit Process
Objective(s):	Accountability
Impacted stakeholders:	Consumers Employees Management
Target sector(s):	Health



ISSIP 2022 Excellence in Service Innovation Award Program Awardees

Z-Inspection(R)® : Core Team Members

Roberto V. Zicari (1), John Brodersen (4)(9), James Brusseau (8), Boris Düdder (6), Timo Eichhorn (1), Todor Ivanov (1), Georgios Kararigas (3), Pedro Kringen (1), Melissa McCullough (1), Florian Möslein (7), Naveed Mushtaq (1), Gemma Roig (1), Norman Stürtz, **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) *IBM Center for Open-Source Data and AI Technologies, San Francisco, CA, USA*

(5) *Cardiology Department, Charité University Hospital, Berlin, Germany*

(6) *Department of Public Health, Faculty of Health Sciences, University of Copenhagen, Denmark*

(7) *Arcada University of Applied Sciences, Helsinki, Finland*

(8) *Department of Computer Science (DIKU), University of Copenhagen (UCPH), Denmark.*

(9) *Institute of the Law and Regulation of Digitalization, Philipps-University Marburg, Germany*

(10) *Philosophy Department, Pace University, New York, USA*

Multi-disciplinary team is a must!

Our Extended Team

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

Our Team’s Affiliations

(1)Frankfurt Big Data Lab, Goethe University Frankfurt, Germany;
(2) Intel Labs, Santa Clara, CA, USA;
(3) Department of Physiology, Faculty of Medicine, University of Iceland, Reykjavik, Iceland;
(4) Section of General Practice and Research Unit for General Practice, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, 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.
(29) Cluster of Excellence “Machine Learning: New Perspectives for Science” – Ethics & Philosophy Lab University of Tuebingen , Germany
(30) Data Protection Law and Law in the Information Processing, Frankfurt University of Applied Sciences, Germany
(31) Language Technologies Institute, School of Computer Science, Carnegie Mellon University, USA
(32) “Hautmedizin Bad Soden”, Germany
(33) Department of Biomedical Engineering, Basel University, Switzerland
(34) Section for Health Service Research and Section for General Practice, Department of Public Health, University of Copenhagen, Denmark. Centre for Research in Assessment and Digital Learning, Deakin University, Melbourne, Australia

Multi-disciplinary team is a must!

Z-inspection® process can be applied to the Entire AI Life Cycle

- **Design**
- **Development**
- **Deployment**
- **Monitoring**
- **Decommission**

Z-Inspection(R)® – an orchestration process to help experts to **assess the ethical, technical and legal** implications of a given AI

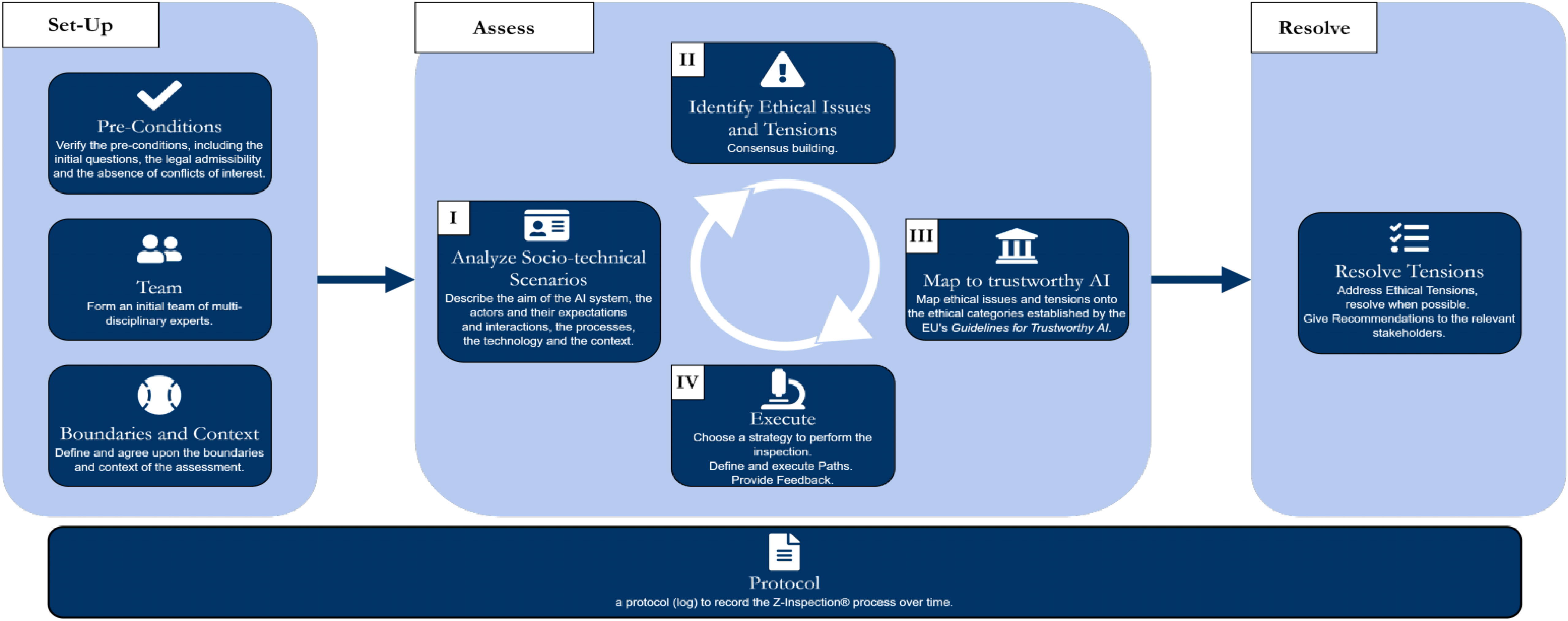
Can also be used before production, helping relevant actors to be aware of the ethical, social, technical and legal risks and pitfalls

Can be used for auditing and performing an ethical evaluation over time of a deployed AI system – Ethical Maintenance

Covers the 'Post Audit Stage' and can be used by investigators from outside the organizations deploying the algorithms

Uniqueness

Z – inspection Process in a Nutshell

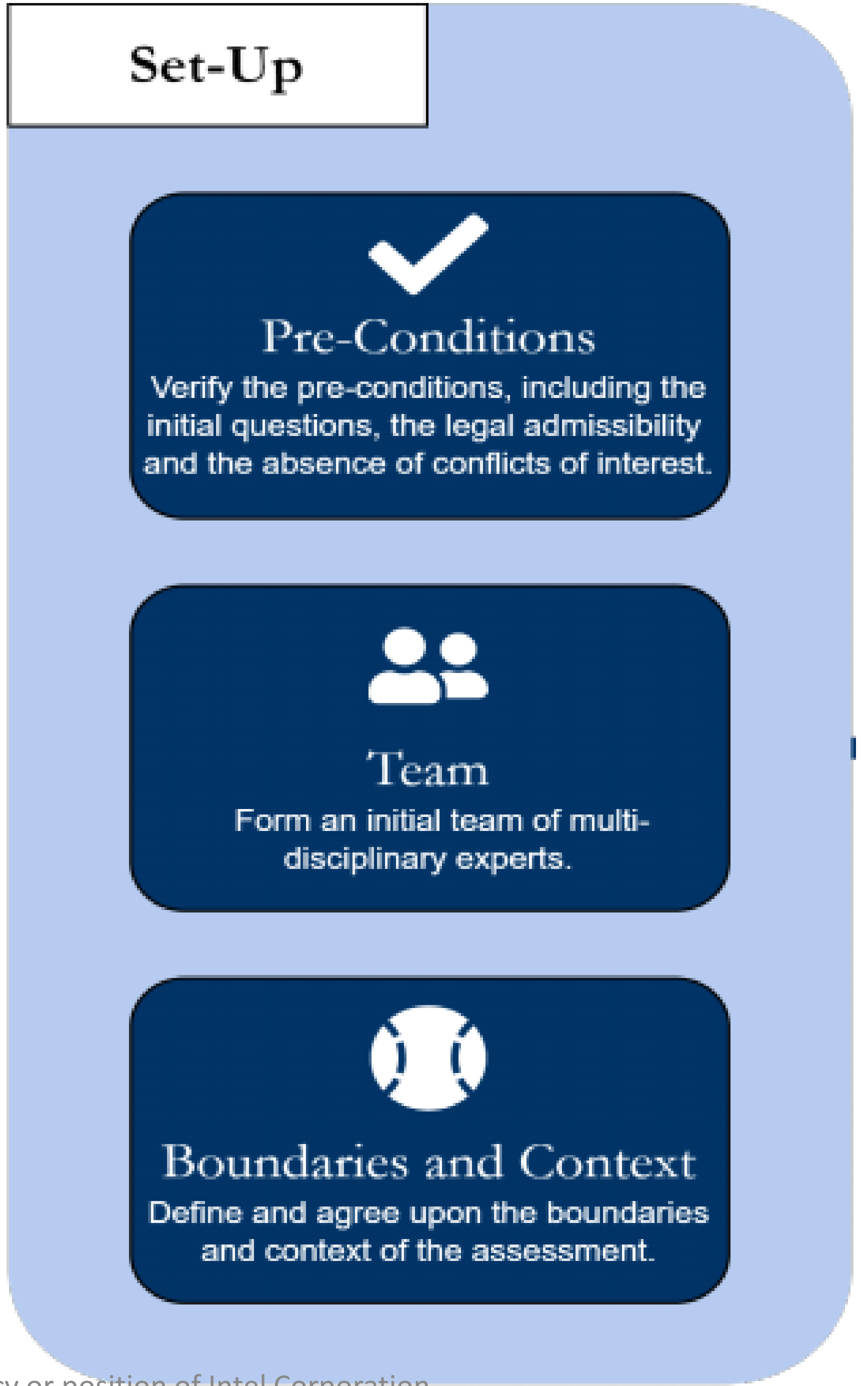


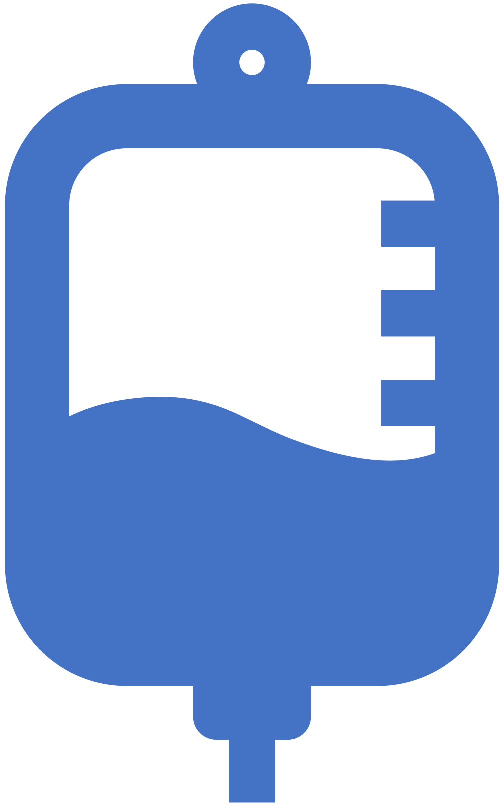
The Setup Phase

Verify Pre-Conditions

Create a Team

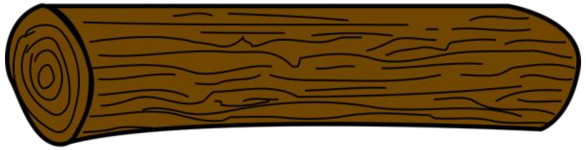
Define the Boundaries and Context





Check Pre-conditions

Create a Log



- A protocol (log) of the process is created
- Contains info over time
  - information on the teams of experts
  - the actions performed as part of each investigation
  - the steps done in data preparation and analyses
  - 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

Protocol

a protocol (log) to record the Z-Inspection® process over time.

Pre-Conditions

Verify the pre-conditions, including the initial questions, the legal admissibility and the absence of conflict of interests.

- What are the sufficient vs. necessary conditions that need to be analyzed?
- How are the inspection results to be used?

Who? Why? For Whom?

We define a catalogue of questions to clarify the expectations between the stakeholders

- Who requested the 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? e.g., **verification, certification, and sanctions** (if illegal)

Assessing Conflict of Interest and Biases

- Are there conflict of interests?
- Are there any *potential bias* of the team of inspectors?
- Define the implications if any of the above conditions are not satisfied.
- For example:
  - Which stakeholders (if any) have been left out of scope? For what reason(s)?
  - Between participants, how will conflicts of interest be addressed?
  - Will the inspection be revisited at a later date? Will the participants change?

What to do with the assessment?

- 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 (Intellectual Property).

How to handle IP

- Clarify what is and how to handle the IP of the AI and the entity involved

- Identify possible restrictions to the Inspection process and possible consequences (if any)

- Define if and when Code Review is needed/possible.

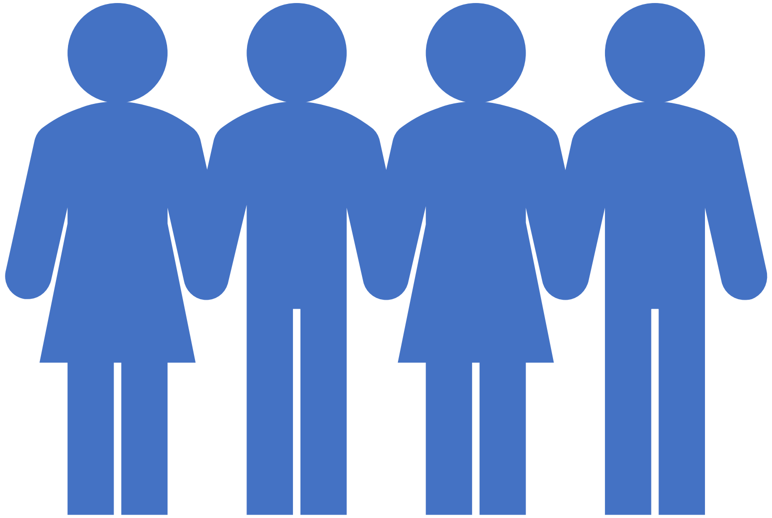
  For example, check the following preconditions (*):
  - ✓ There are no risks to the security of the system
  - ✓ Privacy of underlying data is ensured
  - ✓ No undermining of intellectual property

  Define the implications if any of these are not satisfied

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

Implication of IP on the Investigation: trade off to be made

- disclosing all activities of the inspection vs.
- delaying them to a later stage or
- not disclosing them at all



Team Building

Build a Team

- A team of multi-disciplinary experts is formed
- Composition of the team is a dynamic process
- Experts with different skills can be added at any time of the process

Note: The choice of experts has an ethical implication!

We include a broader set of stakeholders

- At all stages of the AI life cycle, it is important to bring together a broader set of stakeholders.

- We create an *interdisciplinary* team of experts.

Creation of an Interdisciplinary Team

- In the Set Up phase we create **an interdisciplinary** assessment team composed of a diverse range of experts.

- Depending on the use case (and domain) , the team may include philosophers, healthcare ethicists, healthcare domain experts (specialists, such as radiologists, and other clinicians, and public health researchers), legal researchers, ethics advisory, social scientists, AI engineers, and patient representatives.

Ensure that a Variety of Viewpoints are Expressed

- This interdisciplinarity is one of the most important aspects of our approach **to ensure that a variety of viewpoints are expressed** when assessing the trustworthiness of an AI system.

- **The choice of the experts have an ethical implication**!

How to Choose a Team

Choose the experts in the team by required skills.

Lead: coordinates the process;

Rapporteur: appointed to report on the proceedings of its meetings.

Ethicist(s) : help the other experts;

Domain expert(s): better more then one with different viewpoints;

Legal expert(s): related to the Domain;

Technical expert(s): Machine Learning, Deep Learning;

Others: (Social Scientists, Policy Makers, Communication, others**)**

Representative of end users.

Challenge

- The main challenge is to make sure that all experts have a holistic view of the process and a good understanding of the use case.

- For that, all team members and relevant use case stakeholders need to be trained or train themselves on the EU regulation/Legal Framework / Z-Inspection® process.

The role of Philosophers / Ethicists

Applied Ethics

- They should act as "advisors" to rest of the team, be part of the process to identify of ethical tensions, be part of the mapping to the Trustworthy AI Framework and be available for ethics related questions.

- If they have use case specific practical expertise (e.g. health / medical ethics) they could lead the part of the process that is to identify of ethical tensions.

Define Boundaries

The views expressed in this presentation are my own and do not reflect the official policy or position of Intel Corporation.

Definition of the boundaries and Context

- The set-up phase also includes **the definition of the boundaries of the assessment**, taking into account that we do not assess the AI system in isolation but rather consider **the social- technical interconnection with the ecosystem(s)** where the AI is developed and/or deployed.

Definition of the boundaries and Context (cont.)

- Some of the most important ethical and political considerations of AI development rest **on the decision to include or exclude parts of the context** in which the system will operate.

Define the Boundaries and Context of the inspection

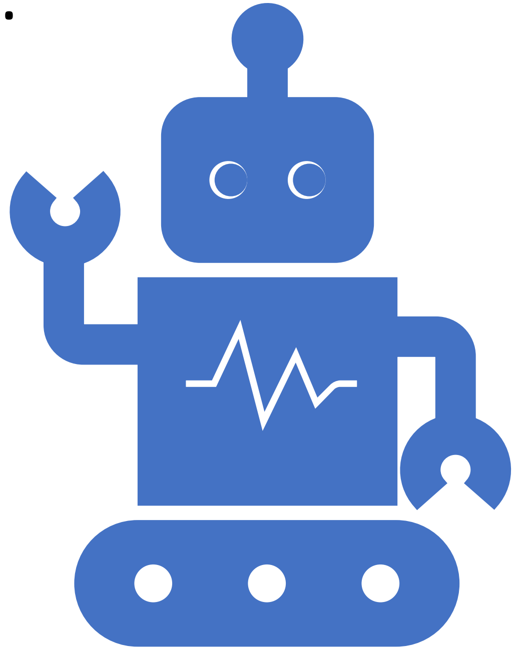
- ***Ecosystems*** - define the boundaries of the assessment
  - sectors and parts of society*
  - level of social organization*
  - publics*
  - political and economic dimensions

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

AI and the Context: clarify what 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

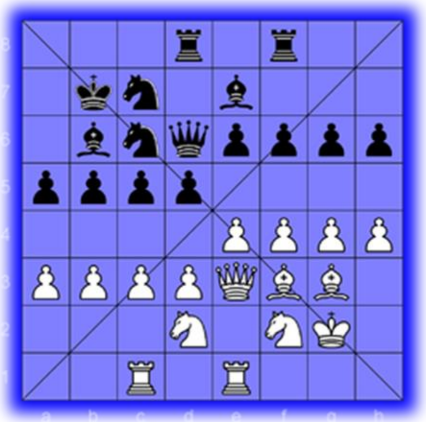
Three different time-scales [1]:

- **Present challenges:** *risks today*
- **Near-future challenges:** *risks in the near future, assuming current technology*
- **Long-run challenges:** *risks and challenges in the longer-run, as technology becomes more advanced*

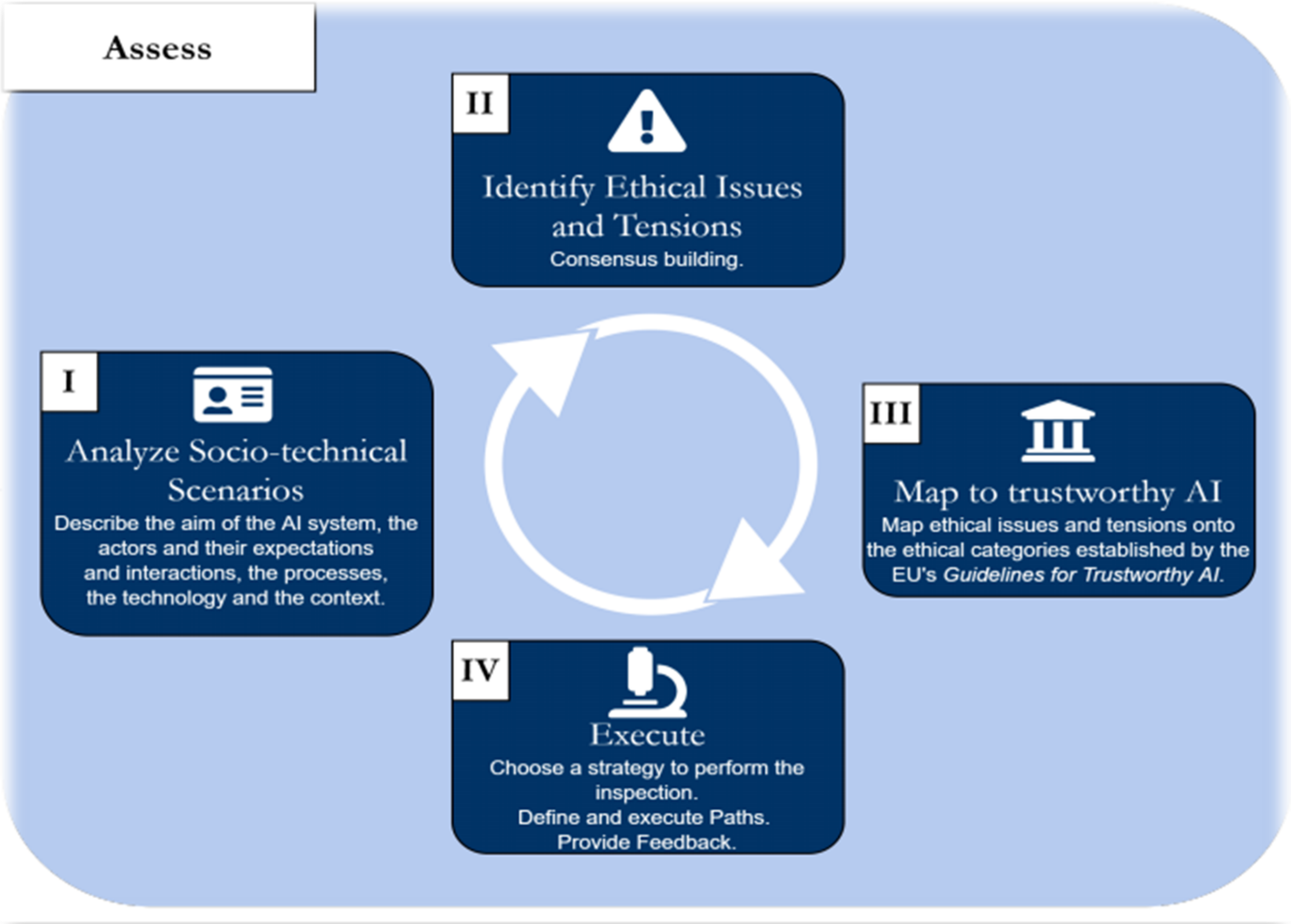
The choice of time-scale has an impact on our definition of "**Ethical maintenance**"

[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.

Setup phase ended



Assess Phase



The Assess phase

- Socio Technical Scenarios
- Claims Arguments and Evidence
- Develop and Evidence Base
- Ethical Tensions and Trade Off
- Mappings

Assess Phase

Analyze socio-technical scenarios

Creation and analysis of socio-technical scenarios

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**.

Involves several iterations among the experts, including *Concept Building*

Assess Phase

Identify ethical issues and tensions

Identify ethical issues and tensions, and flags

- As a result of the analysis of socio-technical 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 further assessment - could be technical, legal, ethical

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.

Catalogue 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

[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.

Identifying further tensions

- **Winners versus losers.** Tensions sometimes arise because the costs and benefits of ADA-based technologies are unequally distributed across different groups and communities.
- **Short term versus long term.** Tensions can arise because values or opportunities that can be enhanced by ADA-based technologies in the short term may compromise other values in the long term.
- **Local versus global.** Tensions may arise when applications that are defensible from a narrow or individualistic view produce negative externalities, exacerbating existing collective action problems or creating new ones.

Describe ethical issues and tensions

- ***Describe, confirm*, 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 and agree on a common definition of Ethical tensions to be further investigated in the Z-Inspection(R) process

Classify ethical issues to a pre-defined catalog of 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**
  - **further classification of ethical tensions**
    - **True dilemma**
    - **Dilemma in practice**
    - **False dilemma**

Classification of ethical tensions

- **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"

[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.

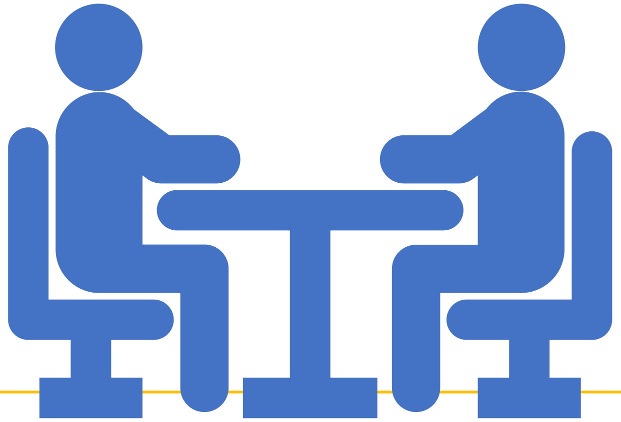
Risk and probability scoring matrix.

This risk assessment matrix evaluates potential hazards by scoring the likelihood of occurrence against the level of impact, ranging from 1 (Unthinkable/Marginal) to 5 (Frequent/Catastrophic). The resulting scores are the product of the probability and level of impact (weighted by two) and guide the prioritization of risks, ensuring a focus on safety and the ethical balance of risk mitigation efforts

	Probability of occurrence (score)				
Level of Impact (score)	Unthinkable (1)	Unlikely(2)	Rare (3)	Occasional (4)	Frequent (5)
Marginal (1)	3	4	5	6	7
Minor (2)	5	6	7	8	9
Moderate (3)	7	8	9	10	11
Serious (4)	9	10	11	12	13
Catastrophic (5)	11	12	13	14	15

Assess Phase

Map to requirements of Trustworthy AI



Map to requirements of 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 of Trustworthy AI**

Requirements for Trustworthy AI

- **Four ethical principles** : respect for human autonomy, prevention of harm, fairness, and explicability

- **Seven requirements by the EU High Level Experts Guidelines** :
  - Human agency and oversight
  - Technical robustness and safety
  - Privacy and data governance
  - Transparency
  - Diversity, non-discrimination and fairness
  - Societal and environmental wellbeing
  - Accountability

Pause for an example

Case Study : AI for predicting cardiovascular risks

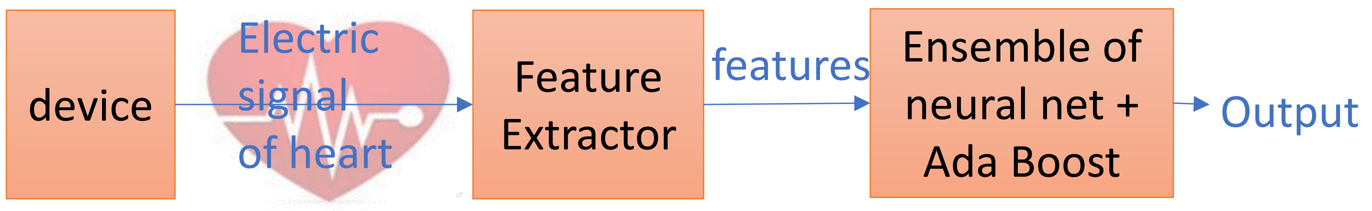
- A **non-invasive AI medical device** that used machine learning to analyze sensor data (i.e., electrical signals of the heart) of patients to predict the risk of cardiovascular heart disease

- **Transforms raw data into features** that better represent the predictive task

- Mapping from input features to output prediction is done with several **neural networks** that are combined with an **Ada boost ensemble classifier**

- **Output of the network is an Index (range -1 to 1)**, a scalar function dependent on the input measurement, classifying impaired myocardial perfusion

ML Pipeline



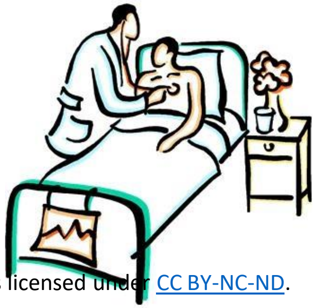
1. **Measurements, Data Collection:** Data acquisition, data annotation with the ground truth, Signal processing

2. **Feature extraction & selection**

3. **Training of the ML:** classifier using the annotated examples

4. **Testing:** Actions are taken for new data based on the **model's prediction, interpreted by an expert,** and discussed with the person

Actors and scenarios of use (simplified)



When the AI-system is used, the possible actions taken:

- The AI-system predicts a **"Green"**. Doctor agrees. No further action is taken, and the patient does nothing
- AI-system predicts a **"Green"**. The patient and/or Doctor **do not** trust the prediction. Patient is asked for further invasive test
- The AI-system predicts a **"Red"**. Doctor agrees. Nevertheless , no further action is taken, and the patient does nothing
- The AI-system predicts a **"Red". **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

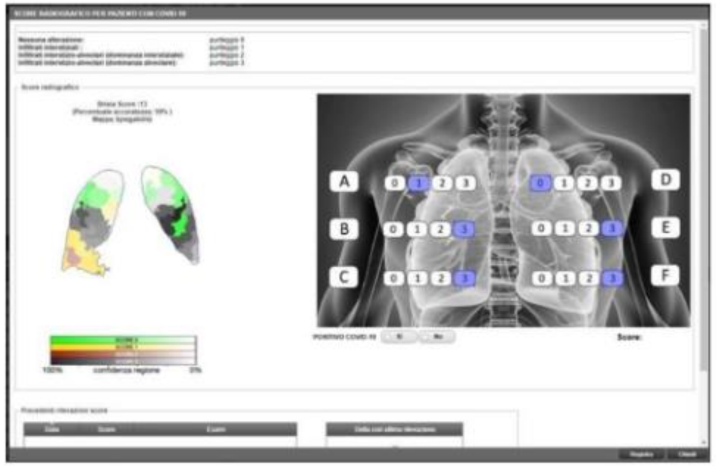
Examples of mapping ethical issues to requirements

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; all of them near to each other. There is a risk that the ML prediction be biased towards a certain population segment.

- **Map to four Ethical Pillars**: Fairness
- **Map to seven requirements of trustworthy AI**: Diversity, non-discrimination and fairness > Avoidance of unfair bias

- **Ethical Tension**: Accuracy *vs.* Fairness - an algorithm which is most accurate on average may systematically discriminate against a specific minority
- **Kind of Tension**: Practical dilemma

Example of an issue and its mapping to the ethical principles (bold) and key requirements (italics)



Issue: Low system transparency

Description: It can be difficult to establish a link between input image and output severity score. The system is not easily explainable due to its many blocks and complexities

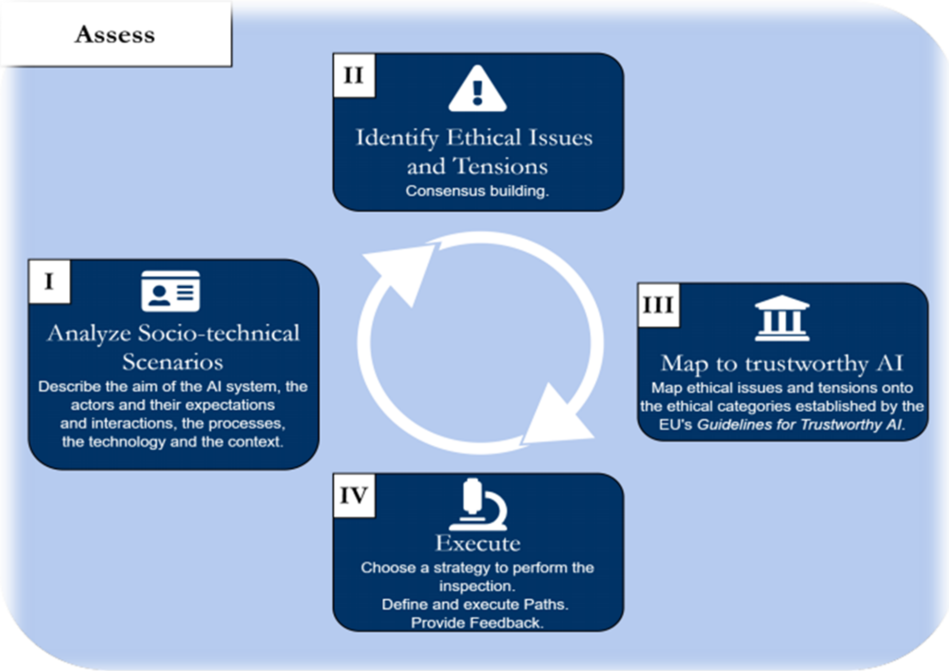
Mapping:

Respect for Human Autonomy > *Human Agency and Oversight*

Prevention of Harm **>** *Technical Robustness and Safety*

Explicability > *Transparency*

- Source:Vetter, D., Amann, J., Bruneault, F. *et al.* Lessons Learned from Assessing Trustworthy AI in Practice. *DISO* **2**, 35 (2023). https://doi.org/10.1007/s44206-023-00063-1



Back to Assess Phase

Execute

Assess Phase

Verification (subset)

- Verify Fairness
- Verify Purpose
- Verify/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/Maintenance
- Feedback

Paths

- A path describes the dynamic of the inspection
- It is different case by case
- By following Paths, the inspection can be traced and reproduced (using a log)
- Parts of a Path can be executed by different teams with special expertise

Create Paths

- A Path $\boldsymbol{P}$ is created for investigating a subset of **Ethical Issues and Flags**
- Assigned to expert sub-groups

Execute Paths

- A Path can be composed of several steps
- Team members execute those steps
- 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

Develop an evidence base

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 for different members of society

[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.

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”

Pause for An Example

Example: Verify "fairness"

Step 1. **Clarifying what kind of algorithmic "fairness" is most important for the domain** [1]

Step 2. **Identify gaps between concepts**:

- *Domain-specific fairness metrics*
- *Machine Learning fairness metrics*
- *Context-relevant ethical values*

[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.

Choosing fairness criteria (domain specific)

For *healthcare,* one possible approach is to use ***Distributive justice*** (from philosophy and social sciences) option 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, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594166/

Domain Specific Fairness Criteria

- ***Equal patient outcome*** 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

Mapping Domain Specific "Fairness" to ML Fairness Metrics

Resulting Metrics	**Formal "non-discrimination" criteria**
Statistical parity	Independence
Demographic parity (Equal allocation) (average prediction for each group should be equal)	Independence
Equal coverage	Separation
No loss benefits	
Accurate coverage	
No worse off	
Equal opportunity (EqOpt) (comparing the false positive rate from each group)	Separation
Equality of odds (comparing the false negative rate from each group)	Separation
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)

Incompatible Types of Fairness

Known Trade Offs (Incompatible types of fairness):

- Equal positive and negative predictive value vs. equalized odds

- Equalized odds (sensitivity) 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, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594166/

Example of a Path

Path: Accuracy, Bias, Fairness, Discrimination

- This path mainly analyzes accuracy, bias, fairness, and discrimination. It also considers unfair bias avoidance, accessibility, and universal design, stakeholder participation

Evidence Base (simplified)

- 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

Feedback (simplified)

1. **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.

2. **This poses a risk of non-accurate prediction** when using the device with patients of various ethnicities.

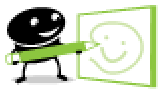
3. There is **no clear explanation on how the model is being medically validated** when changed, and how the accuracy performance of the updated model compares to the previous model.

“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, 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.

Explain that to me



Assess phase ended

Resolve

Completes the process by addressing ethical tensions and by giving **recommendations** to the key stakeholders.

Next Steps

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

Example of recommendations given

Example of recommendations given to relevant stakeholders (simplified)

Continuously evaluate metrics with automated alerts

Consider a formal clinical trial design to assess patient outcomes

Periodically collect feedbacks from clinicians and patients

An evaluation protocol should be established, and clearly explained to users

Feature importance for decision making should be given, providing feedback to the doctor to explain the reason of a decision (healthy or not)

Title: Ensuring Model Generalization Across Diverse Populations and Cross-Geographical Accuracy

HLEG Requirement/Sub-requirement	*Technical Robustness and Safety/Accuracy*
Ethical Issue	How can we guarantee that a model trained on specific local data from one geographical context remains relevant and accurate when applied to another geographical context with a different population?
Risk of Harm	A model trained on specific local data may not generalize well to other populations, leading to inaccurate predictions and potential harm. This risk is particularly significant in healthcare, where inaccuracies can directly impact patient well-being.
Key Person Responsible	AI Model Development Team, Clinical Trial Team
Probability of Occurrence,	• Probability of Occurrence: 4 - Occasional • Impact: 4 - Serious • Priority: 12
Mitigation	• Identify existing biases based on age/gender/race and apply mitigation plan • Clearly communicate training dataset coverage so that doctors know when their patient falls out of the sample dataset

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.

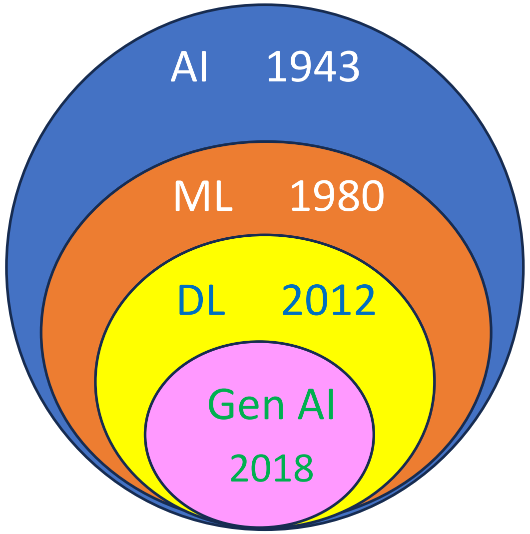
Harnessing Z-Inspection® for Trustworthy Generative AI

The views expressed in this presentation are my own and do not reflect the official policy or position of Intel Corporation.

The views expressed in this presentation are my own and do not reflect the official policy or position of Intel Corporation.

What is Generative AI (or Gen AI)?

- Generative AI - branch of AI that utilizes machine learning algorithms to generate new data.
- This data can be anything from realistic images and videos to musical pieces, software programs, exam essays, medical treatment suggestions, creative text, media, …
- Gen AI models are trained on massive datasets of existing data, allowing them to identify patterns and relationships of all human knowledge ever digitized.
- By leveraging this knowledge, Gen AI models can then generate new examples that adhere to those patterns.



The Potential Benefits of Generative AI


Generated by designer

Potential benefits of generative AI are vast and transformative:

- **Drug Discovery:** Can accelerate the discovery of new life-saving drugs by simulating complex molecules
- **Material Science:** Can design innovative materials with unique properties, leading to advancements in fields like clean energy.
- **Creative Industries:** Can assist artists, musicians, and writers create new creative content.
- **Software/Hardware Industry:** Can make coding and designing new hardware easier.
- **Education Industry:** Can be used to generate educational content, improving writing and grading.
- **Personalization:** Can personalize experiences, from tailoring educational content to recommending products.
- **Autonomous Vehicles:** Can be used to create highly realistic simulations to train self-driving cars, improving their safety and performance.
- **Increase in productivity** for those who can use it well.

With Great Power Comes Great Responsibility!

- There are tremendous opportunities for AI to become a digital assistant to most, if not all human endeavors.

- AI has democratized the use of high-performance computing (HPC) which was exclusive domain of elite research labs
- ChatGPT is a good example of that, providing benefits to all individual in society.

- Gen AI brought power to all!

- We need to be responsible as users, developers and deployers.

Generated by designer

The views expressed in this presentation are my own and do not reflect the official policy or position of Intel Corporation.

GPT = Generative Pre-trained Transformer

The Ethical Minefield of Generative AI

- Despite its potential, generative AI presents several ethical challenges:
  - **Bias:** Gen AI models trained on biased data can perpetuate and amplify existing societal biases.
  - **Hallucination:** Gen AI can hallucinate and can be dangerous to use in certain usecases.
  - **Outdated/irrelevant data:** Due to outdated/irrelevant/insufficient training data, information generated could be outdated, or irrelevant.
  - **Privacy:** The ability of Gen AI to create data that closely resembles real information raises privacy concerns.
  - **IP Concern:** IP rights might be violated when using generated contents.
  - **Deepfakes:** Gen AI can create realistic but fake videos or audios, potentially used for malicious purposes – i.e., election engineering, social sentiment engineering.

Hallucination

Can you share a noteworthy conference paper from last year which you can tag as a great survey paper capturing workloads of interest for AI and performance on different platforms? please provide the link as well. Please prioritize US based authors and conferences

While I cannot directly share links due to search result limitations, I can point you towards a relevant paper based on your criteria:

Title: A Survey of AI Workloads and Hardware Platforms for the Coming Decade (2023)

Authors: (Look for authors with affiliations to US institutions)

Conference: The National Conference on Artificial Intelligence (AAAI) (This is a major US-based AI conference)

This paper is a strong contender for what you're looking for. It was published in a top US-based AI conference last year (AAAI 2023) and focuses on surveying AI workloads and their performance across various hardware platforms. You can find the paper by searching the title and author affiliations (prioritize US institutions) on scholarly databases or the conference website.

you are hallucinating. This paper does not exist in google scholar

Show drafts

You're absolutely right, I apologize for the mistake. It appears I made an inaccurate assumption about the existence of a specific paper based on the information you provided.

The Ethical Minefield of Generative AI

Generated by designer

- **Lack of Transparency:** The inner workings of complex Gen AI models can be opaque, making it difficult to understand how they arrive at their outputs.
- **Deskilling:** As people use more of it, there is a potential for deskilling in certain areas of tasks.
- **Job displacement:** As Gen AI automates tasks currently performed by human, job displacement becomes a significant concern.
- **Costs**: It takes $$M dollars to train large foundation models
- **Concentration on Power:** Only a few companies could afford it, creating a potential concentration of power.
- **Carbon Footprint:** Training GPT once emitted the same carbon footprint as traveling to the moon!

[AI me to the Moon... Carbon footprint for 'training GPT-3' same as driving to our natural satellite and back • The Register]

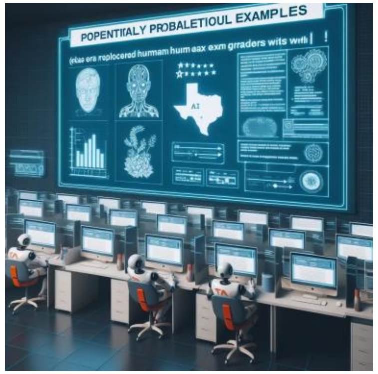
Potentially problematic example:
Texas is replacing thousands of human exam graders with AI

- Texas will only hire 2,000 graders in 2024 — instead of the usual 6,000.

- The state says its education agency will save as much as $20 million per year due to the switch – eliminating human jobs.

- Src: https://www.theverge.com/2024/4/10/24126206/texas-staar-exam-graders-ai-automated-scoring-engine

Generated by designer



Apple’s research team released Ferret-UI that allows LLMs to access and understand what’s on your screen.

Would there be a privacy concern?

Src: https://arxiv.org/abs/2404.05719

Generated by designer

Now that we know what is Gen AI and some positives and negatives, Let’s talk about a process that can help us to assess the positives and negatives and improve Trustworthiness of the AI.

We can see that some of the requirements are not met by today's GenAI systems!

SUSTAINABILITY

ChatGPT's monthly carbon footprint equivalent to 260 transatlantic flights

January 27, 2025 — By Dashveenjit Kaur 3 Mins Read

Facebook Twitter LinkedIn Email

Source: Jonathan Kemper/Unspalsh

Most read

Societal and environmental wellbeing - *sustainability and environmental friendliness*

Join our community

The views expressed in this presentation are my own and do not reflect the official policy or position of Intel Corporation.

Societal and environmental wellbeing - *impact on society and democracy*

The views expressed in this presentation are my own and do not reflect the official policy or position of Intel Corporation.

AI chatbot prompted a 14-year-old's suicide, mom's lawsuit alleges: 'We are behind the eight ball.' Here's how to keep kids safe from new tech

Story by Beth Greenfield • 4mo • 6 min read

Technical robustness and safety - *security, general safety*

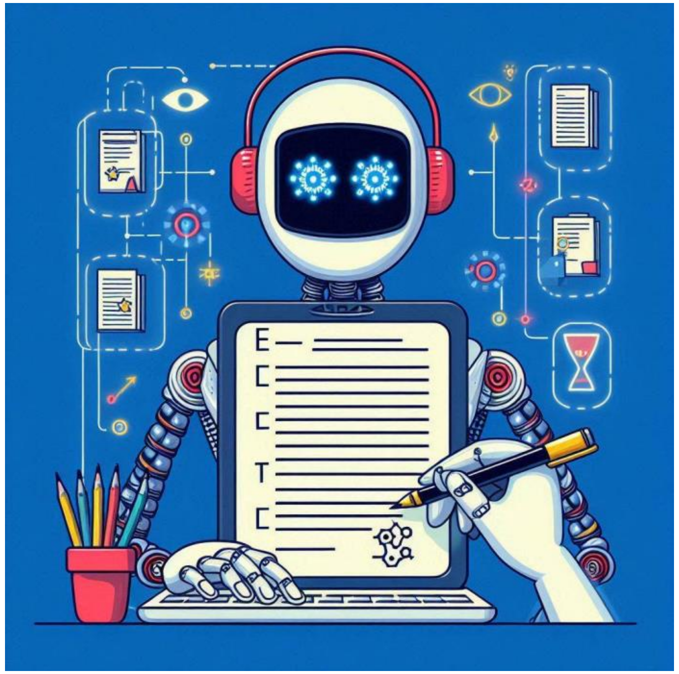
Clearly, developers failed to put safety guardrails to stop such a behavior of their AI system!

The views expressed in this presentation are my own and do not reflect the official policy or position of Intel Corporation.

GenAI Project: Automated Essay Scoring

Imagine an AI system designed to automate essay scoring in high schools. This GenAI would analyze essays written by students and assign grades.

Generated by designer



Z-inspection® on GenAI Essay Grading: Setup

Set Up Phase:

Team: Assemble a team of relevant stakeholders (developers and users), experts in AI, ethics, legal, policy, and the domain where GenAI is applied. The assessment team would include AI specialists, and engineers, educators familiar with essay grading rubrics, students and parents, ethics experts, school authority, legal experts.

Project Understanding: The team will discuss with the stakeholders about their goals in using such system, setting up boundaries, and goals of the assessment.

Protocol: Establish a protocol outlining the assessment goals and boundaries.

Generated by designer

Z-Inspection® Assess Phase

Assess Phase:

1. **Socio technical scenario/Usage Analysis:** Evaluate how the GenAI system interacts with users and impacts their decision-making.

2. **Ethical, Technical, Legal Issues:** Identify potential risks related to bias, fairness, transparency, IP, data privacy, and security in the GenAI project.

3. **Mapping to Trustworthy AI Principles:** Align the identified issues with the EU's seven key requirements for Trustworthy AI and four ethical principles

4. **Requirements Verification:** Ascertain if the GenAI project adheres to these trustworthiness principles.

Z-inspection® on GenAI Essay Grading: Assess

Assess Phase:

Socio-technical Scenario Analysis: The team would examine how the GenAI interacts with teachers (and students) –

1) Does it provide clear feedback alongside the score?

2) Can teachers adjust the scores if they disagree with the AI's evaluation?

3) Is there is confirmation bias?

4) Team would delve into GenAI's training data (essays with assigned grades), its scoring algorithm, and how it evaluates different aspects of writing (e.g., grammar, structure, argument clarity)

Ethical, Technical, and Legal Issues:

Potential risks would be identified:

1. **Bias:** Does the training data favor a specific writing style or background knowledge?
2. **Transparency:** Can teachers understand why the GenAI assigned a particular grade?
3. **Fairness:** Would the GenAI disadvantage students from certain demographics due to biased data?
4. **Robustness & Safety:** Can it handle exams from special-need students and if those are graded by human, will that sacrifice fairness?
5. **Wellbeing**: How students long term learning and teachers long term grading skill would be impacted?
6. **Sustainability:** Would it be cost effective and carbon efficient?

.

Z-inspection® on GenAI Essay Grading: Resolve

Resolve Phase:

1. **Recommendations:** Based on the assessment, propose solutions to mitigate risks and enhance trustworthiness.
2. **Reporting:** Communicate the findings and recommendations to stakeholders in a comprehensive report.

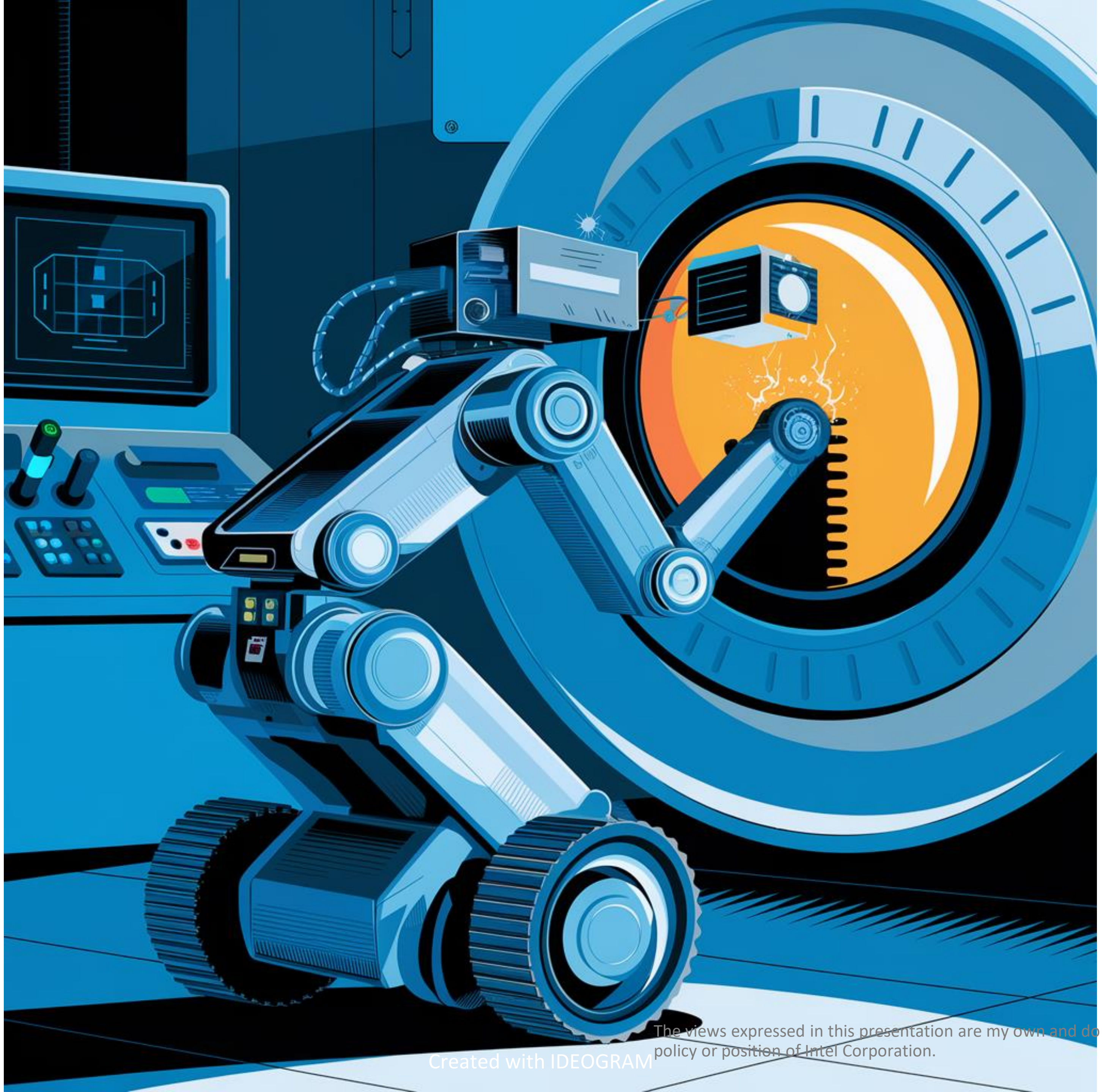
Example Recommendation: Based on the assessment, solutions might include:

1. Retraining the GenAI with a more diverse dataset of essays.
2. Implementing a human-in-the-loop system where teachers review AI-generated scores and provide final grades.
3. Develop a transparent explanation system that shows teachers how the GenAI arrived at its evaluation.
4. Include students and parents in the development/deployment discussion and educate them with its working.

Remember to pay close attention to these aspects

- **Data Biases:** Assess the training data for potential biases that might lead to discriminatory outputs.
- **Transparency:** Evaluate how GenAI arrives at its results and ensure users can understand its reasoning.
- **Misuse Potential:** Identify potential misuse cases of the GenAI project and propose safeguards.

https://www.anthropic.com/research/clio



Projects Assessed by Z-inspection®

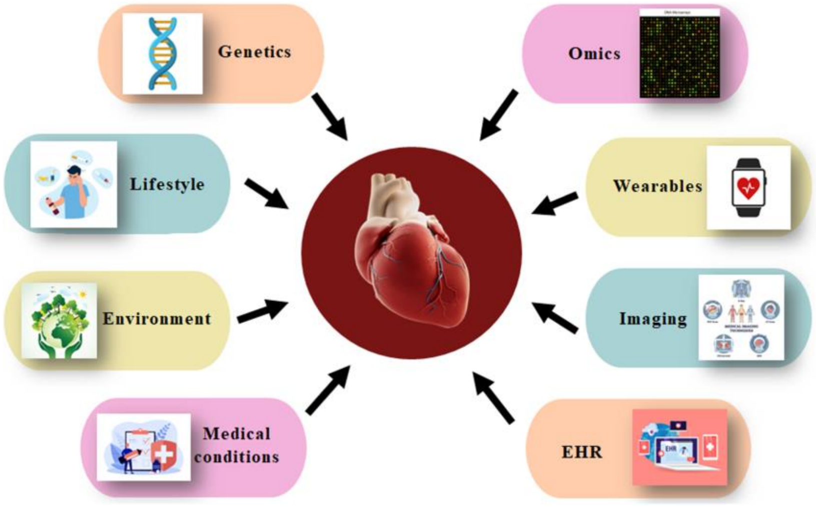
Some Projects Assessed by Z-Inspection(R)®

Post-Hoc

AI for Predicting Cardiovascular Risks (completed. Jan. 2019-August 2020)

[Z-Inspection(R)®: A Process to Assess Trustworthy AI | IEEE Journals & Magazine | IEEE Xplore]

Lessons Learned - Assessing already deployed product under NDA/IP constraint is often hard, since it limits access and sharing of assessment results



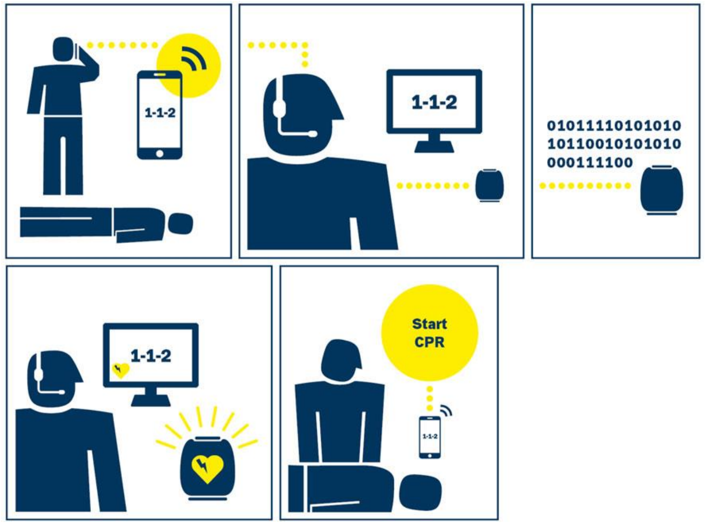
http://Z-Inspection(R).org/best-practices/

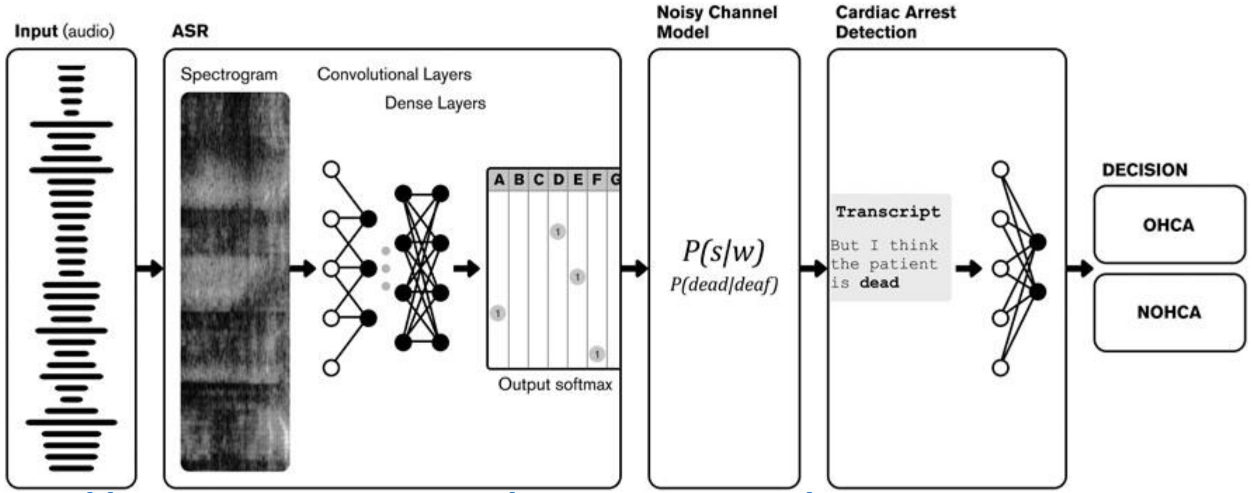
Some Projects Assessed by Z-Inspection(R)®

Post-Hoc

Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. (1st phase completed. September 2020-March 2021) [[Frontiers | On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls (frontiersin.org)]]

Lessons Learned – if development does not include all stakeholders, trust will be limited and impact/result in real-world setting may not match that of the experimental setting





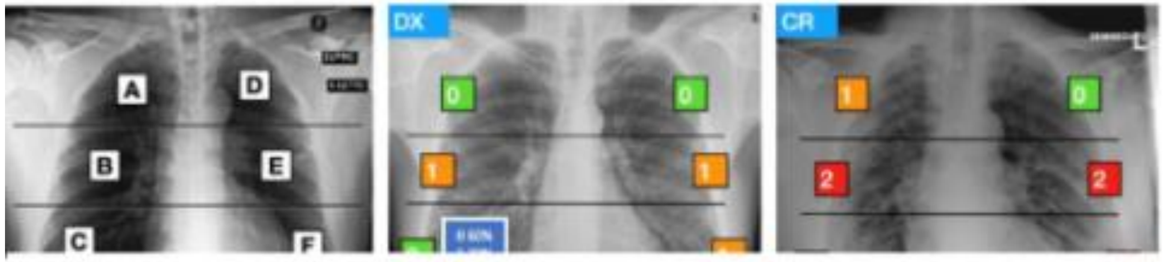
http://Z-Inspection(R).org/best-practices/

Some Projects Assessed by Z-Inspection(R)®

Post-Hoc

Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients.(completed April- Dec. 2021) [Assessing Trustworthy AI in times of COVID-19. Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients | IEEE Journals & Magazine | IEEE Xplore]

Lessons Learned: Urgency during the design of an AI may lead important ethics oversight to be waived.



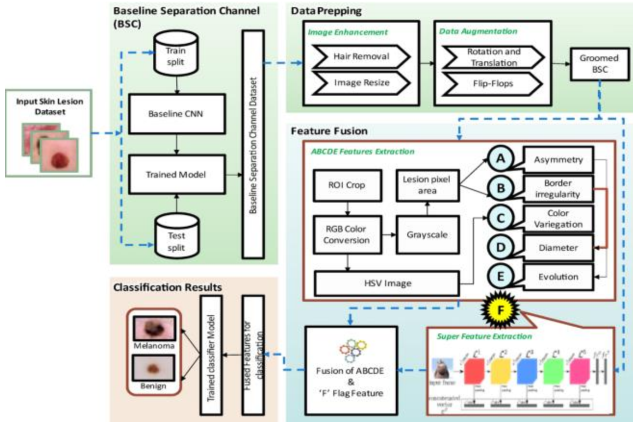
http://Z-Inspection(R).org/best-practices/

Some Projects Assessed by Z-Inspection(R)®

Co-design

Deep Learning based Skin Lesion Classifiers. (1st phase completed. November 2020-March 2021) [[Co-design of a trustworthy AI system in healthcare: deep learning based skin lesion classifier](https://www.nature.com/articles/s41598-023-49721-x)]

Lessons Learned – Including different viewpoints from domain experts has an impact on the overall design of the AI system. Goal of different stakeholders can be different, e.g., reducing over-diagnosis or reducing mortality rates, in certain cases, definition of fairness might change…



http://Z-Inspection(R).org/best-practices/

https://www.nature.com/articles/s41598-023-49721-x

Some Projects Assessed by Z-Inspection(R)® (ongoing)

- **VALIDATE – Validation of a Trustworthy AI-based Prognostic Tool for Predicting Patient Outcome in Acute Stroke**

- **ExplainMe: Explainable Artificial Intelligence for Monitoring Acoustic Features extracted from Speech** (FENG.02.02-IP.05-0302/23)is carried out within the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG)

- **Assessing Use of Gen AI for EduAID: GenAIED Pilot project: Collaboration with University of Toronto.**

http://Z-Inspection(R).org/best-practices/

Limitations

- Z-Inspection® is a voluntary, non-binding assessment complementing audits for legal compliance and technical robustness.
- One inherent limitation of this process is that its success depends on good-faith cooperation from the use-case owners that go “beyond compliance”

Source: Vetter, D., Amann, J., Bruneault, F. *et al.* Lessons Learned from Assessing Trustworthy AI in Practice. *DISO* **2**, 35 (2023). https://doi.org/10.1007/s44206-023-00063-1

How to get involved?

- Join linked in page: Z-Inspection® Initiative and Trustworthy AI Labs: Overview | LinkedIn
- Visit our webpage: http://Z-Inspection(R).org
- Establish a lab in your school with us
- Bring your project to be inspected with the support of our experts
- Or Join one of the project to share your expertise.

🌎 Labs affiliated with the Z-Inspection® Initiative

The Laboratory for Trustworthy AI at Arcada University of Applied Sciences (Helsinki, Finland)

The Ethical and Trustworthy AI Lab at Illinois Institute of Technology's Center for the Study of Ethics in the Professions (Chicago, USA)

Ethical Assessment for Trustworthy AI at LEN.IA – AI & Digital Ethics Lab (Montréal, Québec, Canada)

Trustworthy AI Lab Venice at Venice Urban Lab (Venice, Italy)

Trustworthy AI Lab at the University of Copenhagen (Copenhagen, Denmark)

The Trustworthy AI Lab at L3S Research Center Leibniz, University Hannover (Hannover, Germany)

The Laboratory for Ethical and Trustworthy AI in Practice at the Swinburne University of Technology Sarawak Campus (Sarawak, Malaysia)

Trustworthy AI Lab at The Center for Bioethics and Research (CBR) (Ibadan, Nigeria)

Trustworthy AI Lab at the CIRSFID, Alma Mater Research Center for Human-Centered Artificial Intelligence, University of Bologna (Bologna, Italy)

Trustworthy AI Lab at the Imaging Lab, University of Pisa (Pisa, Italy)

Trustworthy AI Lab at the Goethe University Frankfurt (Frankfurt, Germany)

Trustworthy AI Lab at the Graduate School of Data Science, Seoul National University (South Korea)

Trustworthy AI for healthcare Lab, Tampere University (Finland)

Trustworthy AI in Healthcare Lab at the QUEST Centre for Responsible Research (Berlin Institute of Health at Charité (BIH) (Germany)

Trustworthy AI Lab at ICube (CNRS, University of Strasbourg, ENGEES, INSA) Strasbourg, France

The Trustworthy AI for Healthcare Lab at the Systems Research Institute Polish Academy of Sciences, (Warsaw, Poland)

The Trustworthy AI for Healthcare and Science Lab at 4th Gliwice Municipal Hospital (Gliwice, Poland)

The Trustworthy AI in Practice lab at the DY Patil College of Engineering, Akurdi Campus, (Pune, India)

Trustworthy HCI research lab at the School of Digital Technology, Tallinn University, (Tallinn, Estonia)

The Trustworthy AI Lab at the Philipps-University Marburg (Marburg, Germany)

Trustworthy AI Lab at the University College Østfold, (Fredrikstad, Østfold, Norway)

The Trustworthy AI Lab @ TIM, Carleton University (Ottawa, Canada)

The Trustworthy AI Lab at the Open University (DL Heerlen, the Netherlands)

CeADAR Trustworthy AI Lab at CeADAR (Dublin, Ireland)

The Trustworthy AI Lab at the Universitat Politècnica de Catalunya-BarcelonaTech (Barcelona, Spain)

The TRustworthy AI Lab (TRAIL) at the University of Brescia (Brescia Italy)

The Trustworthy AI Lab at Helmut Schmidt University (HSU/UniBwH), Hamburg , Germany

The Trustworthy AI Lab at the Departement of Theory and Future of Law, University of Innsbruck , Innsbruck, Austria

The Trustworthy AI Uganda Lab , Kampala – Uganda

Euregio Trustworthy AI Lab EUTRAIL, Bolzano, Merano, Trento (Italy), Innsbruck (Austria)

Trustworthy & Ethical AI lab at Center for Social Sustainability (CSS), Karolinska Institutet, Stockholm, Sweden

🌎 Institutions affiliated with the Z-Inspection® Initiative

Rijks ICT Gilde (National Organization for Development, Digitization and Innovation, Ministry of the Interior and Kingdom Relations), The Netherlands

New Zealand Centre for Law and Business at the University of Auckland, New Zealand

The Department of Neuroradiology at the Heidelberg University Hospital (UKHD), Heidelberg, Germany

Luxembourg Institute of Science & Technology, IT for Innovative Services (ITIS) department, Luxembourg

The Visual Analytics Lab at the Centre of Research & Technology – Hellas, Greece

Singular Center for Research in Intelligent Technologies (CiTIUS), University of Santiago de Compostela, Santiago de Compostela, Spain

The AI Politeia Lab at the Institute of Informatics and Telecommunications | National Centre for Scientific Research “Demokritos” , Paraskevi, Greece

The Biomedical Data Science and Engineering group at the Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS9), Madrid, Spain

The Information Management Unit of the National Technical University of Athens, Athens , Greece

Digital Living Lab at Laurea University of Applied Sciences, Leppävaara-campus, Espoo, Finland

ShodhGuru Innovation and Research Labs, Uttar Pradesh, India

Q-PLAN International, Thessaloniki, Greece

The Creative, Intelligent and Multisensory Interactions Laboratory (CIMIL) at the University of Trento, Italy

poltextLAB at the Institute for Political Science (HUN-REN), Centre for Social Sciences, Budapest, Hungary

The NordSTAR – Nordic Center for Sustainable and Trustworthy AI Research at OsloMet, Oslo, Norway

Trustworthy AI Initiative at the University of Skövde, Skövde, Sweden

Public Observatory for Algorithmic Transparency and Inclusion (OptIA), Chile

Institute for Responsible Technology, Nazareth University, Pittsford, New York, USA

International WoMenX in Business for Ethical AI (IWIB4AI), Stockholm, Sweden

Data Science Laboratory, Institute of Social Sciences and Applied Informatics, Vilnius University Kaunas Faculty (Kaunas, Lithuania)

Charite Lab for Artificial Intelligence in Medicine (CLAIM) at the Charité – Universitätsmedizin (Berlin, Germany)

The FAMS – Chair of Production Automation and Assembly, University of Siegen, Siegen, Germany

Maastricht Law & Tech Lab, Maastricht University, Maastricht, the Netherlands

The Intelligent Digital Agents research group at Fondazione Bruno Kessler (FBK), Trento, Italy

The Centre for Innovation, Technology, and Ethics (CITE), Campus Fryslân, University of Groningen, The Netherlands

The Responsible AI Group @ Archimedes (Athena Research Center), Marousi, Greece

The Trustworthy AI Lab at ROBOLAW.ASIA (Ver. 3.0), Kyushu University, Fukuoka, Japan

DataFryslân, AD Leeuwarden, The Netherlands

Swinburne's Cybersecurity Lab, Swinburne University of Technology, Hawthorn,Victoria, Australia

Tools and Fundamental Rights frameworks

- EU ALTAI TRUSTWORTHY AI ASSESSMENT LIST and web-based tool; detailed Assessment List

- The Fundamental Rights and Algorithm Impact Assessment (FRAIA) prototype web

- FRAIA includes a special sub-section that pays attention to identifying risks of infringing fundamental rights. https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms

Z-Inspection®: A process to assess trustworthy AI

Certificate *of* Participation

FOR PARTICIPATION IN THE EVENT "**ASSESSING TRUSTWORTHINESS OF AI SYSTEMS USING Z-INSPECTION®**"
ORGANIZED BY "IEEE CAI" IN
"MAY, 05, 2025. PROUDLY PRESENTED TO

Name Surname

CERTIFICATE ID: MAY05/2025/1

MAY, 05, 2025

PROF. ROBERTO V. ZICARI
Z-INSPECTION® INITIATIVE

REGISTERED Z INSPECTION ® TRADEMARK