How to Assess Trustworthy AI *with* Z-inspection® : An Overview and Lessons Learned

Roberto V. Zicari

Z-Inspection® Initiative Lead

https://z-inspection.org

December 10, 2024

Motivation of our work: How to asses Trustworthy AI in *practice*?

Research started January 2019

photo RVZ

We consider the View of Modern Democracy

Fundamental values

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

  -- **Christopher Hodges**, *Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford*

We use the EU Framework for Trustworthy Artificial Intelligence

The EU High-Level Expert Group on AI defined 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.

We use the Four Ethical Principles of the EU Trustworthy AI Framework

Four ethical principles, rooted in fundamental rights

(i) Respect for human autonomy

(ii) Prevention of harm

(iii) Fairness

(iv) Explicability

- There may be **tensions** between these principles.

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

We use the Seven Requirements and Sub-requirements for Trustworthy AI

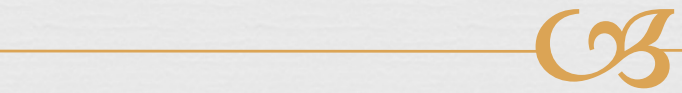
source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

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

They offer a **static checklist** and web tool (ALTAI) for self-assessment, but ***do not validate claims, nor take into account changes of AI over time.***

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

Holistic Approach



- We use a *holistic* approach, rather than monolithic and static ethical checklists.

Z-inspection® Process

We created a *participatory process* to help teams of skilled experts to assess the ***ethical, technical, domain specific*** **and** ***legal*** implications of the use of an AI-product/services within given *contexts*.

- **Published in IEEE Transactions on Technology and Society VOL. 2, NO. 2, JUNE 2021**

OECD Catalogue of AI Tools & Metrics

- Z-Inspection® is listed in the OECD Catalogue of AI Tools & Metrics (including links to Best Practices):

  https://oecd.ai/en/catalogue/tools/z-inspection

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

With Z-Inspection® we want to help to establish what we call a Mindful Use of AI (#MUAI).

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

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

Trustworthy AI *Assessments*

- We have used the process to assess AI Systems already designed, implemented and deployed. So called **post hoc** assessment.

  Post hoc in Latin means '*after this*'.

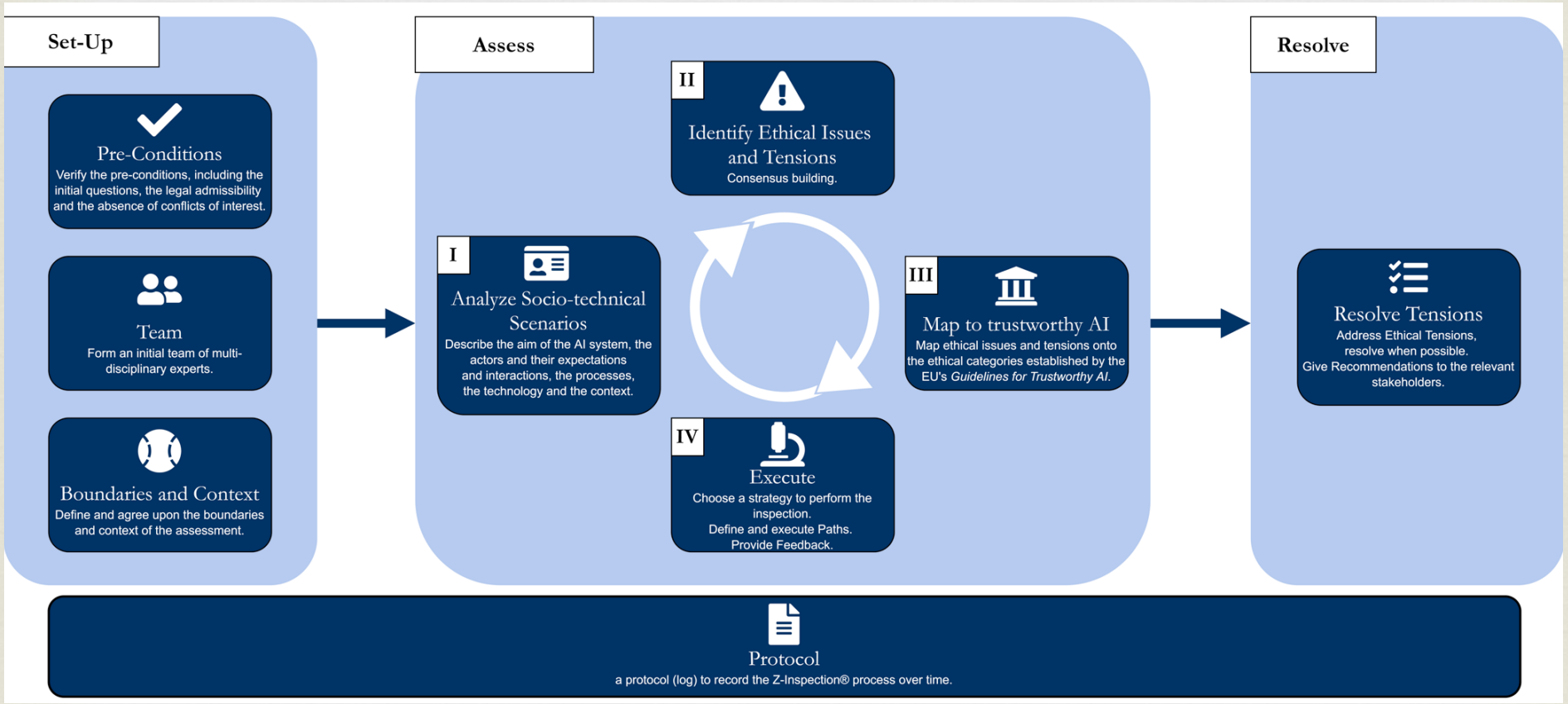
- We also did **ante hoc** (in Latin means "*before this*") assessments. i.e. ***co-design/ co-creation*** of AI systems

International AI Ethical Frameworks

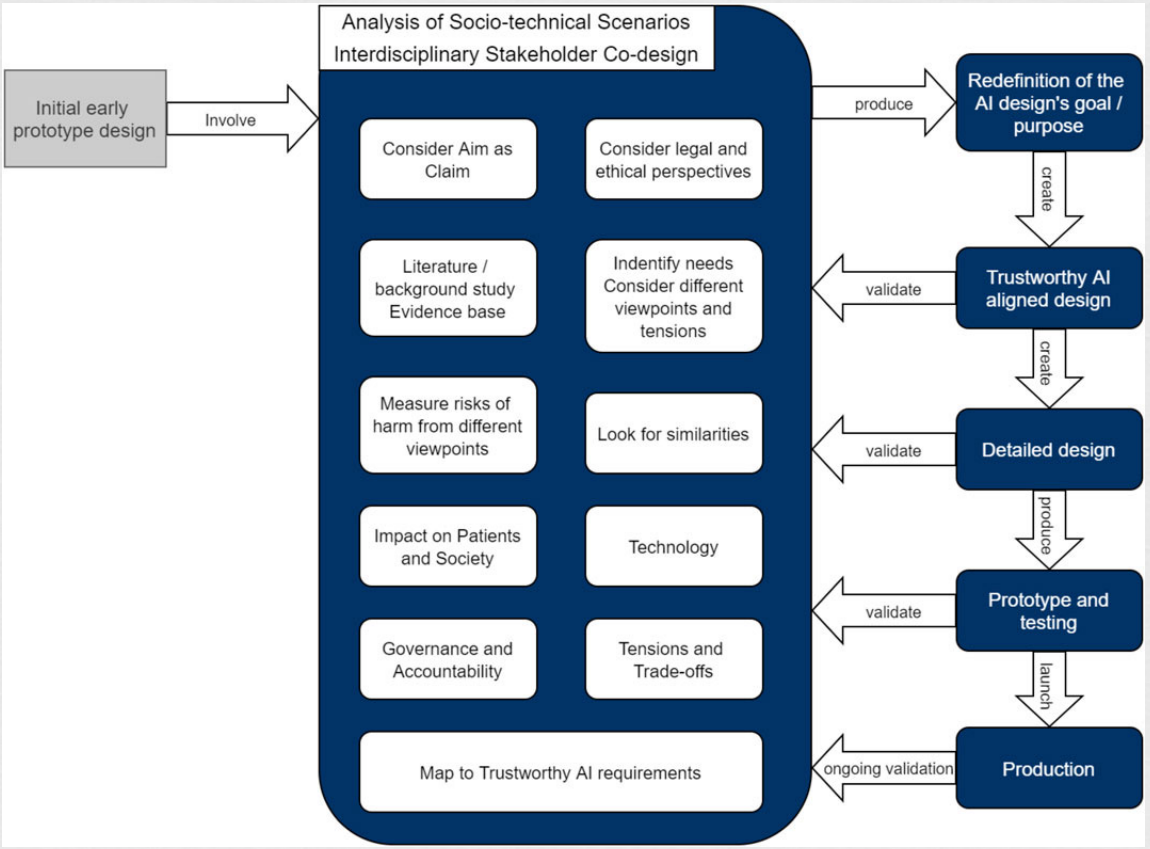
- The Z-inspection® process can also be used for a variety of other AI Ethical Frameworks, e.g. UNESCO, OECD Guidelines, IEEE, etc.

- … besides the EU Trustworthy AI Framework.

Z-inspection® Process in a Nutshell



Ethically aligned co-design methodology



A. Set Up

- Verify Pre-Conditions
- Create a Team
- Define the Boundaries and Context

Pre-Conditions

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

- Who requested the inspection?
- Why carry out an inspection?
- For whom is the inspection relevant?
- Is it recommended or required (mandatory inspection)?

Pre-Conditions (cont.)

- What are the sufficient vs. necessary conditions that need to be analyzed?
- How are the inspection results to be used?
- Will the results be shared (pubic) or kept private?
- **Are there conflict of interests?**

Pre-Conditions (cont.)

- 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?

*We include a broader set of stakeho*lders

- 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 a 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 view points;
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 / 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.

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.

B. The Assess phase

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

1. We use Socio-technical Scenarios to identify "*issues*"

By collecting relevant resources, the team of interdisciplinary experts create socio-technical scenarios and analyze them to describe:

the aim of the AI systems,

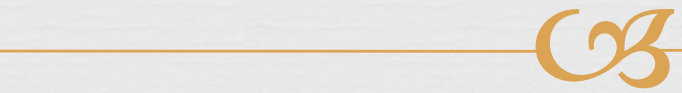
the actors and their expectations and interactions,

the process where the AI systems are used,

the technology and the context (*ecosystem*).

Resulting in a number **of *issues*** *(possible risks)* **to be assessed**

2. We use the Claims, Arguments and Evidence (CAE) framework



Look for Claims:

Claims – "assertions put forward for general acceptance. They are typically statements about a property of the system or some subsystem.

Claims that are asserted as true without justification become **assumptions** and claims supporting an argument are called sub claims. "

Claims, Arguments and Evidence (CAE)

Provide Evidence:

Evidence "that is used as the basis of the justification of the claim.

Sources of evidence may include the design, the development process, prior field experience, testing, source code analysis or formal analysis", peer-reviewed journals articles, peer-reviewed clinical trials, etc.

Claims, Arguments and Evidence (CAE)

Arguing:

Arguments link the evidence to the claim.

They are defined as *Toulmin's warrants* and are the "statements indicating the general ways of arguing being applied **in a particular case** and implicitly relied on and whose trustworthiness is well established", together with the validation for the scientific and engineering laws used.

3. We develop an evidence base

This is an iterative process among experts with different skills and background with goal to:

- Understand technological capabilities and limitations
- Build a stronger evidence base to support claims and identify tensions (*domain specific*)
- Understand the perspective of different members of society

Source: Whittlestone, J et al (2019)

Develop an evidence base (cont.)

- Technology is generally designed for a highly specific purpose, however, it is not always clear what the technologies unintended harm might be.

- Therefore, an important part of our assessment process is **to build an evidence** base through the **socio-technical scenarios to identify tensions as potential ethical issues.**

4. Identifying “issues”

- When a Claim has no evidence it becomes an assumptions, and this could be a potential risk.

  We call them “issues”.

- How to describe “issues”?

- Use free text and an open vocabulary

Concept building

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

- Terminological overlaps
- Differences between disciplines
- Differences across cultures and publics
- Conceptual complexity

- Source: *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.

5. Identify Tensions and Trade-offs

- **Tensions** may arise between ethical principles**, for which there is no fixed solution**.

- "In line with the EU fundamental commitment to democratic engagement, due process and open political participation, *methods of accountable deliberation to deal with such tensions should be established.*"

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

Ethical Tensions

- "We use the umbrella term 'tension' to refer to different ways in which values can be in conflict, some more fundamentally than others."

- "When we talk about tensions between values, we mean tensions between the pursuit **of different values in technological applications** rather than an abstract tension between the values themselves."

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.

We use a Catalog of predefined ethical tensions

- To help the process, especially as a help to experts who might have not sufficient knowledge in ethics, we used a sample of catalog of predefined ethical tensions.

- We have chosen the catalog defined by the Nuffield Foundations (Whittlestone et al., 2019)

- Source: *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***

Source: Whittlestone, J et al (2019) – *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

„Thinking about tensions could also be enhanced by systematically considering different *ways* that tensions are likely to arise.

We outline some conceptual lenses that serve this purpose“

Source: Whittlestone, J et al (2019) – *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

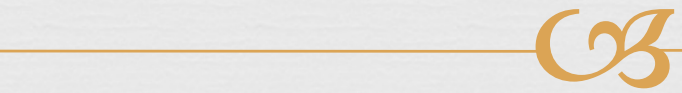
- **Winners versus losers**. Tensions sometimes arise because the costs and benefits of ADA-based technologies are unequally distributed across different groups and communities.
- **Short term versus long term**. Tensions can arise because values or opportunities that can be enhanced by ADA-based technologies 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.

Use a Classification of Ethical Tensions

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

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

6. Consensus Based *Mappings:* from *Open to Close vocabulary*



We map *issues* freely described (*open vocabulary)* by the interdisciplinary team of experts) to some of the EU 4 ethical principles and 7 requirements (sub-requirements) for Trustworthy AI (*closed vocabulary*)

(or to another AI Ethical Framework of choice)

We rank *mapped issues* by relevance depending on the context. (e.g. Transparency, Fairness, Accountability)

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

The resolve phase

- The resolve phase completes the process by addressing ethical tensions and by giving **recommendations** to the key stakeholders.

Recommendations

- **Appropriate use;**
- **Remedies**: If risks are identified, we recommend ways to mitigate them (when possible);
- **Ability to redress.**

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

Tools and Fundamental Rights frameworks

- **EU ALTAI TRUSTWORTHY AI ASSESSMENT LIST and web-based tool;**

- **The Fundamental Rights and Algorithm Impact Assessment (FRAIA)**

ALTAI Check List and web tool

1. **TRUSTWORTHY AI ASSESSMENT LIST : Check List of questions.** The AI HLEG translated these requirements into a detailed Assessment List, taking into account feedback from a six month long piloting process within the European AI community.

2. **ALTAI web tool**: the Vice-Chair of the AI HLEG and his team at the Insight Centre for Data Analytics at University College Cork, developed a prototype web based tool, to practically guide developers and deployers of AI through an accessible and dynamic checklist.

   https://altai.insight-centre.org/

The Fundamental Rights and Algorithm Impact Assessment (FRAIA)

- **The Fundamental Rights and Algorithm Impact Assessment (FRAIA)** helps to map the risks to human rights in the use of algorithms and to take measures to address this In all stages, respect for fundamental rights must be ensured.
- The FRAIA includes a special sub-section that pays attention to **identifying risks of infringing fundamental rights and to the need to provide a justification for doing so.**

https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms

Best Practices

Post hoc assessments

- **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 in times of COVID-19**.
  Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients.*(April- Dec. 2021)*

Pilot Project: Assessment for Responsible Artificial Intelligence

- together with Rijks ICT Gilde -Ministry of the Interior and Kingdom Relations (BZK) and the province of Fryslân (The Netherlands) (*April –September 2022*)

- **During this six-month pilot, the practical application of a deep learning algorithm from the province of Fryslân was investigated and assessed. The algorithm maps heathland grassland by means of satellite images for monitoring nature reserves. The assessment of this algorithm was done in collaboration with an international interdisciplinary team, using the Z-Inspection® process.**

Pilot Project: Assessment for Responsible Artificial Intelligence

- *" The results of this pilot are of great importance for the entire Dutch government, because we have developed a best practice with which administrators can really get started, and actually incorporate ethical values into the algorithms used."*

  – *Rijks ICT Gilde* -Ministry of the Interior and Kingdom Relations (BZK)

Link to the project web site and results:
https://www.rijksorganisatieodi.nl/rijks-ict-gilde/mycelia/pilot-kunstmatige-intelligentie

Best Practices

Ante hoc Assessment

- **Co-design of Trustworthy AI. Best Practice**:
  Deep Learning based Skin Lesion Classifiers.
  (*November 2020-March 2021*)
  In collaboration with the team of Dr. Andreas Dengel at the German Research Center for Artificial Intelligence (DFKI)
  Published in *Front. Hum. Dyn. Human and Artificial Collaboration for Medical Best Practices, July 13, 2021*

Co-design of Trustworthy AI. Best Practice: Deep Learning based Skin Lesion Classifiers.

- The main contribution of our work is to show the use of an **ethically aligned co-design methodology** to ensure a trustworthiness early design of an artificial intelligence (AI) system component for healthcare.
- The system is aimed to explain the decisions made by deep learning networks when used to analyze images of skin lesions.
- Our research work is addressing the need for **co-design of trustworthy AI using a holistic approach,** rather then using static ethical checklists.

New Pilot Project (started August 2023): Assessing Trustworthiness of the use of Generative AI for higher Education.

- For this pilot project, we will assess the ethical, technical, domain-specific (i.e. education) and legal implications of the use of Generative AI-product/service within the university context.
- We follow the UNESCO guidance for policymakers on AI and education. In particular the policy recommendation 6. : *Pilot testing, monitoring and evaluation, and building an evidence base.*

https://z-inspection.org/pilot-project-assessing-trustworthiness-of-the-use-of-generative-ai-for-higher-education/