Lessons Learned in Performing a Trustworthy AI and Fundamental Rights Assessment.

Roberto V. Zicari

Seoul National University
Z-Inspection® Initiative Lead
https://z-inspection.org

Seoul, September 16, 2025

Technology is not created in a vacuum.

- **“It carries the values, perspectives, and power structures of those who design it."**

Source UN Women Conference, Seoul :https://sites.google.com/view/ai-and-gender-conference-websi/home

We are the Z-Inspection® Initiative (started Jan. 2019)

The Z-Inspection® Initiative is a *non-commercial* initiative

100+ individual experts

78 *affiliated* Institutions and Labs

in 40 countries all over the world.

Australia, Austria, Belgium, Brasil, Canada, Chile, Cyprus, Croatia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, India, Ireland, Italy, Japan, Latvia, Lithuania, Luxemburg, Malaysia, the Netherlands, Nigeria, Norway, Poland, Portugal, Rumania, Sierra Leone, South Korean, Spain, Sweden, Switzerland, Uganda, Turkey, United Kingdom, USA, New Zealand.

https://z-inspection.org

Our Mission

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

Motivation of our work

How to asses Trustworthy AI in *practice*?

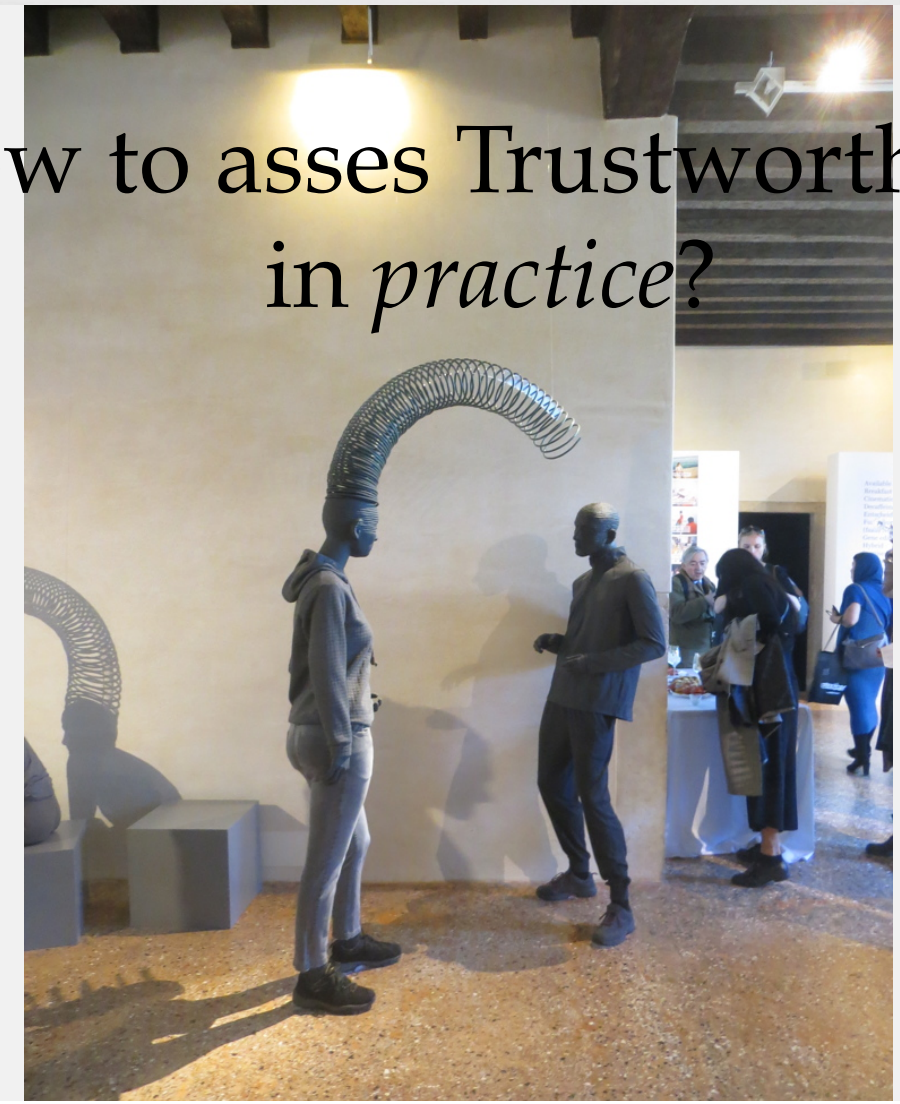


photo RVZ

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**

"Responsible use of AI" Pilot Project

- ***Pilot Project with the Province of Fryslân, Rijks ICT Gilde & the Z-Inspection® Initiative.***
- Marjolein Boonstra, Frédérick Bruneault, Subrata Chakraborty, Tjitske Faber, Alessio Gallucci, Eleanore Hickman, Gerard Kema, Heejin Kim, Jaap Kooiker, Elisabeth Hildt, Annegret Lamade, Emilie Wiinblad Mathez, Florian Möslein, Genien Pathuis, Giovanni Sartor, Marijke Steege, Alice Stocco, Willy Tadema, Jarno Tuimala, Isabel van Vledder, Dennis Vetter, Jana Vetter, Magnus Westerlund, Roberto V. Zicari.

- The pilot project took place from May 2022 through January 2023.

- During the pilot, the practical application of a deep learning algorithm from the province of Fryslân was assessed.

Environmental Monitoring

- The AI maps heathland grassland by means of satellite images for monitoring nature reserves. **Environmental monitoring** is one of the crucial activities carried on by society for several purposes ranging from maintaining standards on drinkable water to quantifying the CO2 emissions of a particular state or region.
- Using satellite imagery and machine learning to support decisions is becoming an important part of environmental monitoring.

Share the experiences, results and lessons learned from performing

- **A Trustworthy AI assessment** using the Z-Inspection® process and the EU framework for Trustworthy AI, combining it with

- **A Fundamental Rights assessment** using the Fundamental Rights and Algorithms Impact Assessment (FRAIA) as recommended by the Dutch government for the use of AI algorithms by the Dutch public authorities.

The "Assessment for Trustworthy AI" pilot sought to answer to the following questions

- As a government, how do you govern the development and use of *responsible* AI?
- What frameworks, laws and regulations are important, and how do we assess them in the development and use of AI?
- How do you analyze, assess and improve AI applications?
- And are the applications consistent with public values and human rights?
- What ethical issues does the AI system raise?
- What fundamental rights could be affected by the AI system?
- What measures could be met for the AI system to be *trustworthy*?

- The pilot gave some answers to these questions and in addition helped to stimulate awareness and dialogue about AI within the Dutch government, and provided guidelines to be able to confidently deploy AI technology for the questions of tomorrow.

- **“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)

The background

- The Province of Fryslân is investing, in the coming years, in the smart and effective use of data. The province sees that almost all provincial developments and social tasks contain a data component. This creates urgency in the subject. To be able to responsibly respond to technological developments as a province, a sharp vision on data and AI is needed. Participation in the pilot helped design the future digital infrastructure and outline ethical frameworks.

- As directly related to this pilot, the Province of Fryslân is required by law to monitor biodiversity in natural areas. This is done by conducting a manual, visual inspection once every 10 years. There is a need to monitor and map the natural areas more often and monitor heather fields for grassification of heathlands and faster. To facilitate the process, reduce its costs and streamline the procedure, the Province commissioned a third party to develop an AI system for this purpose.

Scope of the Pilot

- The scope of the pilot was to assess whether the use of this AI system is trustworthy, which fundamental human rights are affected by the AI system, and how it can be used responsibly in practice.

Aim of the AI System

- The aim of the AI system is to **help ecologists** to quickly and frequently image the natural area so that it can be checked whether the intended nature quality objectives are being met, the right management measures can be taken and whether the approach to increasing biodiversity is working.
- Specifically the AI system **aims to provide information about the diffusion of the invasive and unwanted grass species** Molinia caerulea, known as moor grass or pipestraw, and Avenella flexuosa (common name wavy-hair grass) in heather fields using satellite images. The satellite images are made available by The Netherlands Space Office (NSO) on the free Satellite Data Portal where generic high resolution optical satellite images are available to be used in GIS.

Approach

In the pilot, the AI system was examined from three different perspectives:

- 1. Technical
- 2. Ecological
- 3. Ethical and Fundamental Human Rights

Ethics and Fundamental Rights Assessment

In light of the introduction of a fundamental rights impact assessment tool for algorithms in the Netherlands in March 2022, a hybrid approach was adopted in the pilot.

- First, the AI system was assessed against the **human rights** requirements using **the FRAIA**. This assessment did not only consider human rights violations but also rights which could be protected or strengthened by applying the AI system, such as the right to a healthy environment.
- Then **ethical issues** were identified and assessed based on the **European guidelines for Trustworthy AI** and the system was assessed from this broader perspective.

Claim, Arguments and Evidence

Our approach is unique, as it combines a fundamental rights assessment with a Trustworthy AI Assessment using a **evidence based approach**, using a framework called *Claim, Arguments and Evidence*.

Focus: fundamental rights

- In identifying the fundamental rights being affected by the AI system the assessment looked at the list of fundamental rights provided in the FRAIA, which are clusters around four groups with specific rights listed under each of the areas. Rights related to:
- The person
- freedom-related fundamental rights
- equality rights
- procedural fundamental rights

Focus: Ethics

- Following this step, the assessment considered more broadly ethical issues arising from the AI system. Specifically, the ethics guidelines for trustworthy artificial intelligence were considered as defined by the EU High-Level Expert Group on AI (AI HLEG, 2019). The four ethical principles of the guidance were used, acknowledging that tensions may arise between them:
- (1) Respect for human autonomy (2) Prevention of harm (3) Fairness (4) Explicability
- Furthermore, the seven requirements of Trustworthy AI defined in (AI HLEG, 2019) were considered. Each requirement has a number of sub-requirements .

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

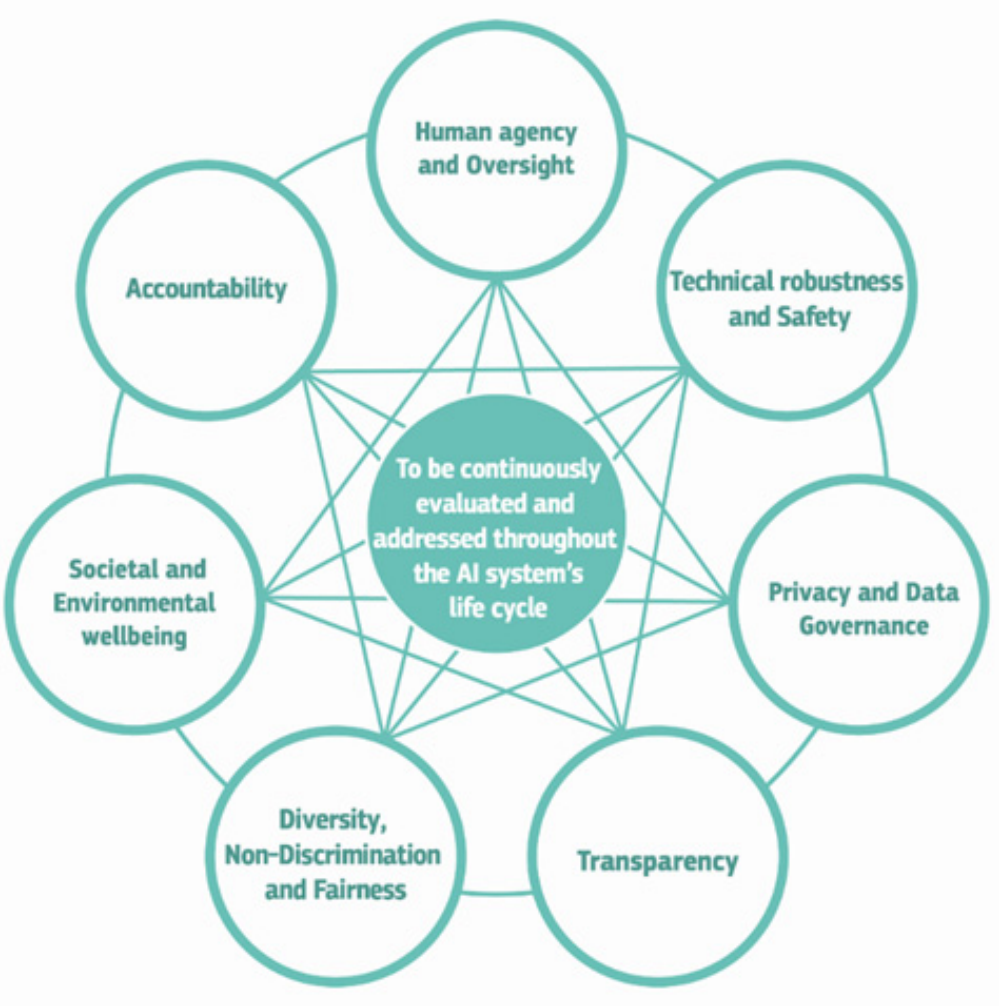
(iii) Fairness

(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.

EU 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.

- **Requirements Sub-Requirements**
- 1 Human agency and oversight *Including fundamental rights, human agency and human oversight*
- 2 Technical robustness and safety *Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility*
- 3 Privacy and data governance *Including respect for privacy, quality and integrity of data, and access to data*
- 4 Transparency *Including traceability, explainability and communication*
- 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, minimization and reporting of negative impact, trade-offs and redress.*

Human Rights Assessment with a Evidence base approach

In this pilot, for each of the rights identified as potentially affected, the assessment concludes with a ***claim*** whether the right is

- a) affected (regardless of whether this is positively or negatively affected),
- b) not affected, or
- c) might be affected depending on certain clarifications.

A brief argument is made in respect of each claim and ***evidence*** is provided in support of whether the right is affected. The assessment identified five fundamental rights clusters which were potentially affected by the AI system.

I. Rights related to the Person

I. Rights related to the Person:

- Rights related to Healthy living Environment
- Rights related to Personal identity/personality rights/personal autonomy
- Rights related to Protection of data and informational privacy rights
- Rights related to Territorial privacy

II Procedural Rights

II Procedural Rights

- 5. Rights related to Right to good administration

Additional relevant aspects that arise from the Trustworthy AI Assessment

- In addition to the ethically relevant aspects discussed in the context of the fundamental rights-based assessment, additional ethical issues were identified for reflection.

Ethical issues

- Transparency and lack of transparency
- Receiving relevant information
- Human agency and oversight
- Technical robustness and safety
- Justice and fairness
- Cost reduction
- Diversity and Inclusion
- Responsibility and Accountability
- Due diligence in decision-making

Comparing the Trustworthy AI assessment process with the fundamental rights-based FRAIA assessment tool

- While the FRAIA tool was useful and clear, the question about how to frame the human rights assessment nevertheless arose and more specifically, how to consider the fundamental rights as part of an assessment of trustworthiness and ethical reflections on an AI system.
- **Should we consider the rights as they are defined in law and interpreted through the courts only**?
- **Or should the rights be considered more broadly, as part of the assessment, linking the rights to ethical principles beyond their narrower legal definition?**
- If only the legal definitions are used, an assessment of whether specific legislation applies would be required.

A two-tiered, integrated approach.

- **If the human rights assessment is defined too narrowly it risks being an assessment separate from the ethical assessment, or the assessment of trustworthiness.**
- **If it is too broad, the human rights standards risk being watered down.**

A two-tiered, integrated approach, looking both at legal requirements and the broader ethical questions, could be envisaged, depending on the organizational set up and use case.

Lessons Learned from the two assessment approaches

- The fundamental rights assessment and the ethics assessment based on the Trustworthy AI guidelines go hand in hand; both approaches provide critical insights with regard to the AI use case.

Similarities and Differences

Reflecting on AI from an ethics perspective clearly overlaps with a fundamental rights assessment.

- Both ethics and fundamental rights are about norms and fundamental values held in society.
- As ethics reflection and ethics guidelines influence law, scholars from both fields must work together when thinking about the shaping of technology and its societal implications.
- Even though there are great similarities, there are several considerable differences between the two approaches.

Differences

- **A fundamental rights-based approach is more closely linked to existing law and focuses on aspects that are legally relevant and thus enforceable.**
- **Compared to this, an ethics-based approach is much broader and also more open to reflection on potential implications that may not be worth considering from a legal perspective.**
- For example, from an ethics perspective, personal autonomy, freedom of decision-making, and fairness were found to be concepts of clear relevance in the context of the pilot project's AI tool, whereas, from a rights-based perspective, rights related to personal autonomy in a strictly legal sense were considered not infringed by the AI tool.

Different perspectives

- **While a fundamental rights-based assessment focuses on whether fundamental rights are negatively affected or infringed, from an ethics perspective, both positive and negative implications of AI technology are considered.**
- In this pilot, for example, the potential positive implications of the AI tool on the environment proved to be central.

Final Note

- Approaching the use case from a fundamental rights perspective implies that ethical and societal aspects and implications of AI are discussed only in-sofar as they are related to fundamental rights and existing law.

For more information

Report with the results

Lessons Learned in Performing a Trustworthy AI and Fundamental Rights Assessment.

Cite: arXiv:2404.14366 [cs.CY]

https://arxiv.org/abs/2404.14366

YouTube Video:

https://www.youtube.com/watch?v=z_RCysclXdk

Winner of the 2025 ISSIP Awards Excellence in Service Innovation with Distinguished Recognition – Impact to Society.

https://z-inspection.org/wp-content/uploads/2025/03/ISSIP2025-certificate.pdf