How to Assess
Trustworthy AI
in practice

Roberto V. Zicari
Z-Inspection® Initiative
http://z-inspection.org

Innovation, Governance and AI4Good
The Responsible AI Forum
Munich, December 6, 2021

1. We consider the View of contemporary Western European democracy

Fundamental values

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

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

2. 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.

EU Framework for Trustworthy Artificial Intelligence

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 Requirements for Trustworthy AI

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

3. How to asses Trustworthy AI in *practice*?

photo RVZ

Challenges and Limitations of the EU Framework for Trustworthy AI

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***.

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.).

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

Z-inspection®

- We created an *orchestration 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*.

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

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

Based on our research work On Assessing Trustworthy AI: Best Practices

- Assessing Trustworthy AI. Best Practice: **AI for Predicting Cardiovascular Risks** (completed. Jan. 2019-August 2020)
- Assessing Trustworthy AI. Best Practice: **Machine learning as a supportive tool to recognize cardiac arrest in emergency calls**. (1st phase completed. September 2020-March 2021)
- Co-design of Trustworthy AI. Best Practice: **Deep Learning based Skin Lesion Classifiers**. (1st phase completed. November 2020-March 2021)
- Assessing Trustworthy AI. Best Practice: **Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients**.(completed Dec. 2021)

We use a holistic approach

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

We include the full community 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.

We use Socio-technical Scenarios to identify *issues*

By collecting relevant resources, a 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* to be assessed**.

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)

We use a consensus process based on *mappings. Open to close vocabulary*

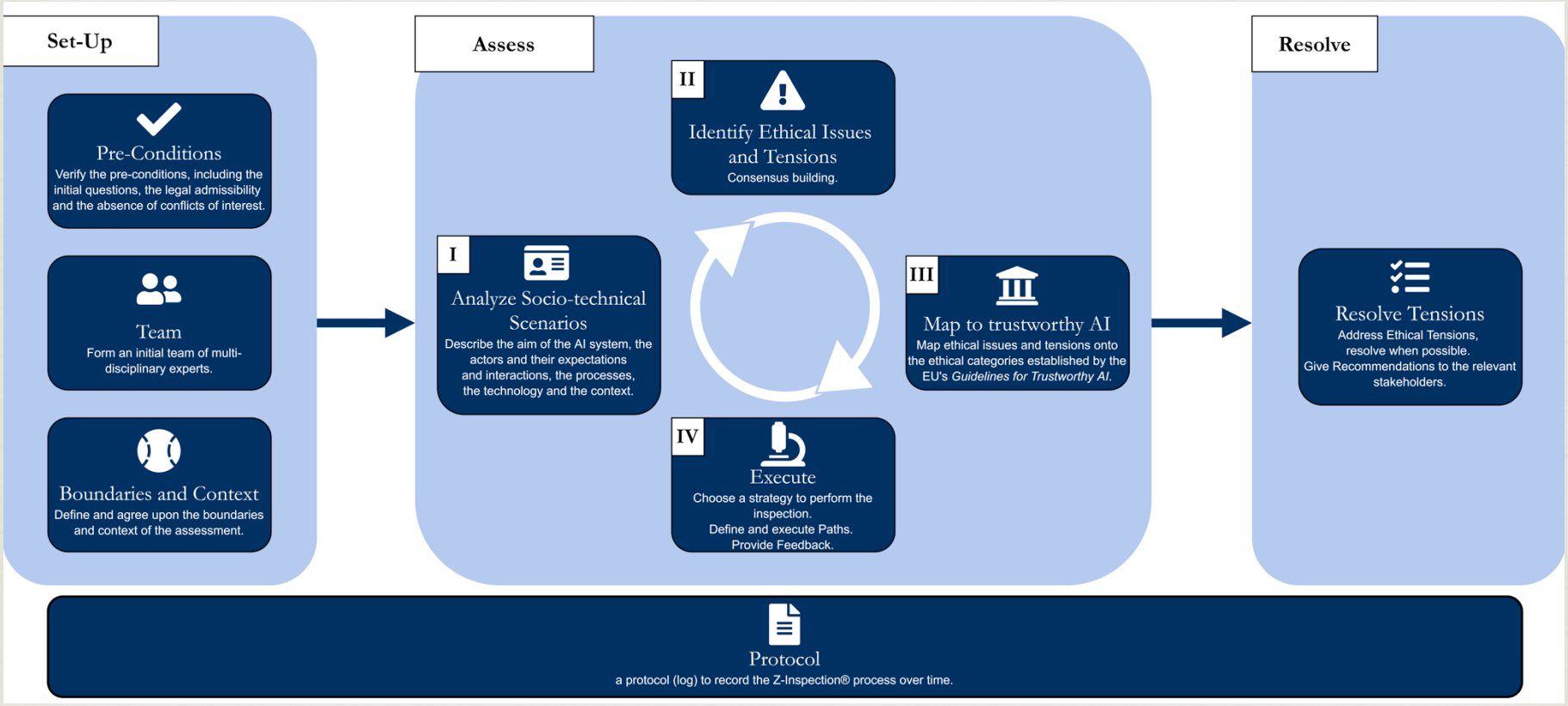
We map *issues* freely described (*open vocabulary)* by the interdisciplinary team of experts) to some of the 4 ethical principles and 7 requirements for Trustworthy AI (*closed vocabulary*)

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

We give Recommendations

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

Z-inspection® Process in a Nutshell



How to handle IP

- Clarify *what is* and *how to handle* the *IP* of the AI and of the part of the entity/company to be examined.
- Identify possible restrictions to the Inspection process, in this case assess the consequences (if any)
- Define if and when *Code Reviews* is needed/possible. For example, check the following preconditions (*):
  - There are no risks to the security of the system
  - Privacy of underlying data is ensured
  - No undermining of intellectual property

  Define the implications if any of the above conditions are not satisfied.

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

Resources

http://z-inspection.org