How to Assess Trustworthy AI *with* Z-inspection®

Roberto V. Zicari

Z-Inspection® Initiative

http://z-inspection.org

GSDS SNU, Seoul, November 15, 2022

Work to do for the Final report

1. Study the comments received for the Mid Term report;
2. Do not edit the Mid Term report
3. Take into account some of the comments for the Final report

Final Report: To Do List

1. Revise (if any)
the part on Claim/Evidence/Arguments,
the list of Issues and Ethical Tensions,
the kind of tensions and the trade offs.

Final Report: To Do List (cont.)

2. Use the ALTAI web tool to double check the issues identified and perhaps adding new issues and tensions (see step 1.)

Check if the general recommendations received from the ALTAI tool are appropriate for the use case and comment on it in the report.

Final Report: To Do List (cont.)

3. When appropriate go deeper with the Fundamental Rights Assessment part

Comment in the report the relationship you have found between fundamental human rights and the ethical issues.

Final Report: To Do List (cont.)

4. Proceed with the mapping from Ethical issues to the EU Framework for Trustworthy AI

Use the template we provided (See examples later)

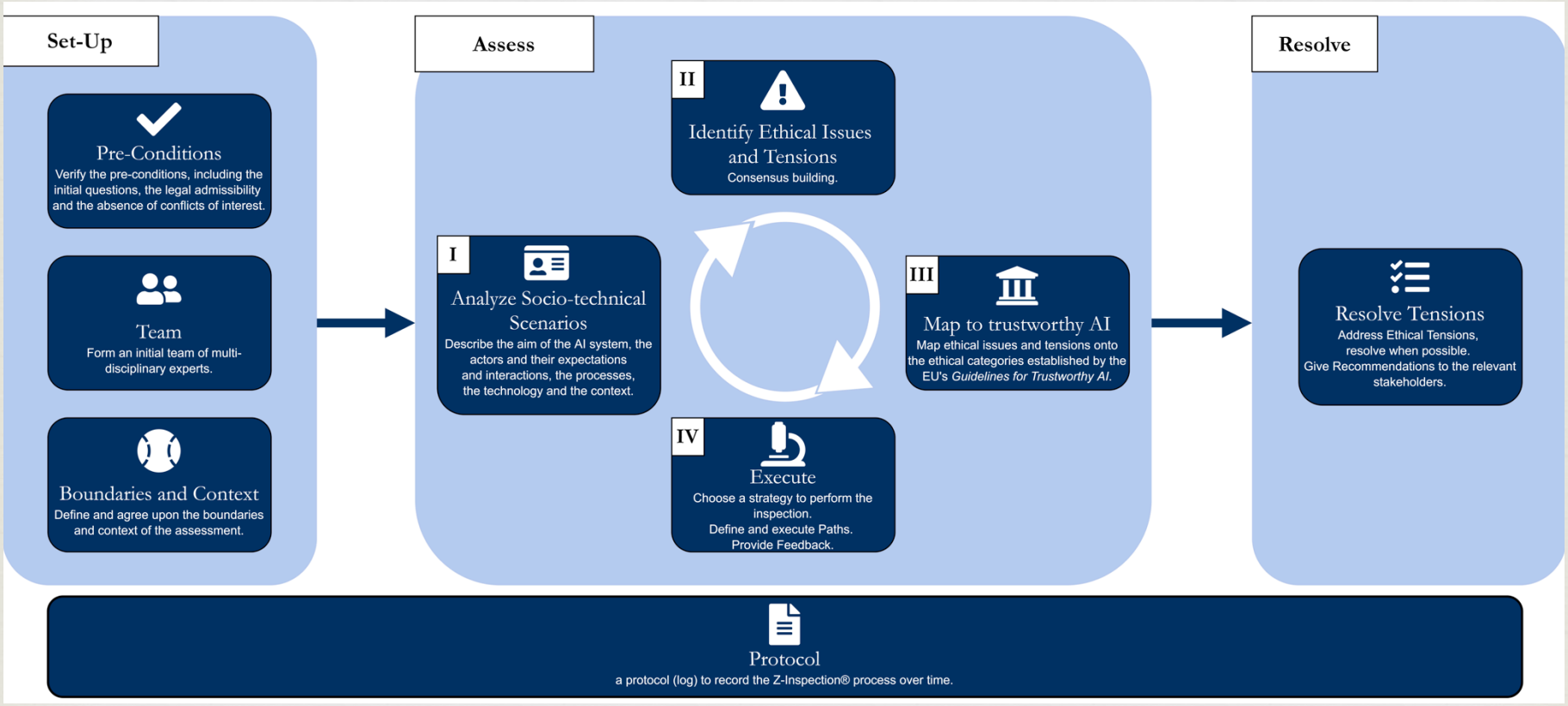
and explain briefly how did you come up with the mapping.

Final Report: To Do List (cont.)

5. Offer recommendations to stakeholders.

Decide which stakeholders you wish to give recommendations.

Z-inspection® Process in a Nutshell



References on How to do the Mapping

- **Z-Inspection®: A Process to Assess Trustworthy AI**

Roberto V. Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas , Pedro Kringen, Melissa McCullough, Florian Möslein, Karsten Tolle, Jesmin Jahan Tithi, Naveed Mushtaq, Gemma Roig , Norman Stürtz, Irmhild van Halem, Magnus Westerlund.

IEEE Transactions on Technology and Society,
VOL. 2, NO. 2, JUNE 2021

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9380498

References on How to do the Mapping

- **How to Assess Trustworthy AI in Practice.**

Roberto V. Zicari (1)(2)(3), Julia Amann (4), Frédérick Bruneault (5)(6), Megan Coffee (7), Boris Düdder (8), Alessio Gallucci (3), Thomas Krendl Gilbert (9), Thilo Hagendorff (10), Irmhild van Halem (3), Eleanore Hickman (11), Elisabeth Hildt (12)(1), Sune Holm (8), Georgios Kararigas (13), Pedro Kringen (3), Vince I. Madai (14)(15), Emilie Wiinblad Mathez (3), Jesmin Jahan Tithi (16)(17), Dennis Vetter (3)(18), Magnus Westerlund (1)(19), Renee Wurth (3). On behalf of the Z-Inspection® initiative (2022).

https://arxiv.org/pdf/2206.09887.pdf

References on How to do the Mapping and Recommendations

On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls

Roberto V. Zicari • James Brusseau • Stig Nikolaj Blomberg • Helle Collatz Christensen • Megan Coffee • Marianna B. Ganapini • Sara Gerke • Thomas Krendl Gilbert • Eleanore Hickman • Elisabeth Hildt • Sune Holm • Ulrich Kühne • Vince I. Madai • Walter Osika • Andy Spezzatti • Eberhard Schnebel • Jesmin Jahan Tithi • Dennis Vetter • Magnus Westerlund • Renee Wurth • Julia Amann • Vegard Antun • Valentina Beretta • Frédérick Bruneault • Erik Campano • Boris Düdder • Alessio Gallucci • Emmanuel Goffi • Christoffer Bjerre Haase • Thilo Hagendorff • Pedro Kringen • Florian Möslein • Davi Ottenheimer • Matiss Ozols • Laura Palazzani • Martin Petrin • Karin Tafur • Jim Tørresen • Holger Volland • Georgios Kararigas

Front. Hum. Dyn., 08 July 2021 |
https://www.frontiersin.org/articles/10.3389/fhumd.2021.673104/full

Example of a mapping

If you have found an issue, e.g.:

- **Potential Harm Resulting From Tool Performance**

now use the template to do the mapping

ID Ethical Issue: E5, Potential Harm Resulting From Tool Performance

- **Description**

The tool's characteristic performance, such as a higher rate of false positives compared to human dispatchers, could adversely affect health outcomes for patients.

Map to Ethical Pillars/Requirements/ Sub-Requirements (Closed Vocabulary)

Mapped to:

- *Ethical Pillar:* **Prevention of Harm**

Requirements > Technical Robustness and Safety

Sub requirements > Accuracy.

Narrative Response

- The algorithm did not appear to reduce the effectiveness of emergency dispatchers but also did not significantly improve it. The algorithm, in general, has a higher sensitivity but also leads to more false positives. There should be a firm decision on thresholds for false positive vs. false negatives. The risk of not doing CPR if someone needs CPR exceeds the risk of doing CPR if not needed. On the other hand, excessive false positives put a strain on healthcare resources by sending out ambulances and staff to false alarms. This potentially harms other patients in need of this resource. The gold standard to assess whether the tool is helpful for the given use case is to analyze its impact on outcome. Given, however, the low likelihood of survival from out of hospital cardiac arrest, there wasn't an analysis attempting to assess the impact on survival, as it would take years in a unicentric study.

Example of Classification of tension

- Kind of tension: True dilemma.
- Trade-off: *Fairness vs. Accuracy*.
- Description: The algorithm is accurate on average but may systematically discriminate against specific minorities of callers and/or dispatchers due to ethnic and gender bias in the training data.

Recommendations to the key stakeholders

- The output of the assessment is a report containing recommendations **to the key stakeholders**. Such recommendations should be considered as a source of qualified information that help decision makers make good decisions, and that help the decision-making process for defining appropriate trade-offs. They would also help continue the discussion by engaging additional stakeholders in the decision-process.

Who are the key stakeholders?

- You decide who are the key stakeholders for your use case

Examples

- The developers of the vendor company
- The management and communication people at the vendor company
- The Primary Actors identified in the Socio-Technical Scenarios
- The Secondary Actors Actors identified in the Socio-Technical Scenarios
- The intended end users of the AI system

Example of Recommendations

- *Recommendation 1*: **It is important to ensure that dispatchers understand the model predictions so that they can identify errors and detect biases that could discriminate against certain populations.**

Stakeholders: **dispatchers**

Recommendations

- Here, the model is a statistical black-box, and the clinical trial conducted with the model showed an important lack of trust that had an impact on the outcome of the trial.
- An improvement to the model would include interpretable local approximations [such as SHAP (Lundberg and Lee, 2017)], which are easy for stakeholders to understand and provide different levels of interpretation for judging the relevance of an individual prediction. In our example, explanation may involve words that were more predictive, tone of voice, or breath sounds.

  Stakeholders: **developers**

Recommendations

- *Recommendation 2*: We believe that the team should either intentionally sample the entire training set in order to prevent discrimination, or define a heuristic that could inform dispatchers when to use and when not to use the model. ..

- *Stakeholders:* **management of the 112 emergency service, and developers**

Recommendations

- *Recommendation 3*: Involve stakeholders. The group of (potential future) patients and (potential future) callers could be interested in how the system functions and is developed. User involvement/ stakeholder involvement could be very helpful in the process of re-designing the AI system. ..

- *Stakeholders:* **management of the 112 emergency service**

Recommendations

- *Recommendation 4*: It is important to learn how the protocol (what questions, how many, etc.) does or does not influence the accuracy of the ML output. Further research work should be performed to answer this question….

- *Stakeholders:* **medical doctors, management responsible for the 112 emergency service and developers**

Recommendations

- *Recommendation 5:* Although we did not assess the legal aspects of the AI system, we suggest to the prime stakeholder to verify with legal local competent authorities if the AI system needed a CE-certification as a medical device…

- *Stakeholders:* **management of the 112 emergency service, ethical board, legal officers**

Team presentations

- **November 17: there will be a Q&A in presence (NOT teams presentations) Bldg. 942, Room #302**

- Teams presentations will be on

Tuesday Nov. 22 and Thursday Nov. 24:. in presence. Bldg. 942, Room #302

.

Team presentations

For the team presentation here are the requirements.

Each team will present a **Power Point presentation**.

- Time is **7 minutes**.

Content of the presentation

- Brief introduction of the team members
- Brief introduction to the AI system chosen
- Main results obtained in the Mid Term report
- Next steps: what will be done for the Final Report

Team presentations

- You will receive feedback but **no gra**des for the presentations.
- The aim is to **help you to write up the Final Report** and to learn **how to present your results in public.**