A holistic and analytical approach to assess Trustworthy AI: Z-inspection® project

Francesca Lizzi
Istituto nazionale di Fisica Nucleare, sezione di Pisa
Next-AIM general meeting, Milano, 13-15 Febbraio 2023

Outline

1. Quick intro to the EU Framework for Trustworthy Artificial Intelligence

2. Quick intro to the Z-Inspection®: A Process to Assess Trustworthy AI

3. Illustration of a Use Case: evaluation of the BrixiaNet algorithm for severity assessment of COVID-19 patients

Trustworthy AI

"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. As a result, claims about system attributes may not be easily verified." **Yoshua Bengio**

"AI may improve health care and medicine all over the world **only if ethics and human rights are a main part of its development.** Ethical guidance based on the **shared perspectives** of the different entities that develop, use or oversee such technologies is critical to build trust in these technologies, **to guard against negative or erosive effects and to avoid the proliferation of contradictory guidelines.**" **World Health Organization**

The EU framework for trustworthy AI

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.

Possible tensions between this components

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

Foundations of Trustworthy AI

Four ethical principles, rooted in fundamental rights:

(i) Respect for human autonomy;

(ii) Prevention of harm;

(iii) Fairness

(iv) Explicability

Seven requirements for Trustworthy AI

(1) Human agency and oversight

(2) Technical robustness and safety

(3) Privacy and data governance

(4) Transparency

(5) diversity, non-discrimination and fairness

(6) environmental and societal well-being and

(7) accountability

Possible 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

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.

Assessing Trustworthy AI

- The EU guidelines offers a static checklist and web tool to perform a SELF-ASSESSMENT.
- No validation of claims nor taking into account changes of AI over time.
- The AI HLEG guidelines are GUIDELINES and not a law. Some of the requirements is not anchored to the context.

We need a way to assess Trustworthy AI dynamically.

Z-inspection® project is an experiment to assess Trustworthy AI in practice

https://z-inspection.org/

Roberto V. Zicari et al. Z-Inspection ® : A Process to Assess Trustworthy AI . IEEE Transactions on Technology and Society, 2(2):83–97, 2021

H. Allahabadi *et al.*, "Assessing Trustworthy AI in Times of COVID-19: Deep Learning for Predicting a Multiregional Score Conveying the Degree of Lung Compromise in COVID-19 Patients," in *IEEE Transactions on Technology and Society*, vol. 3, no. 4, pp. 272-289, Dec. 2022, doi: 10.1109/TTS.2022.3195114.

Z-inspection

- It is an **orchestration process** to help stakeholders to assess ***ethical, technical, domain specific*** **and** ***legal implications*** of the use of an AI product.
- Since its beginning, 4 algorithms have been analysed in the health context.
- **Holistic approach**: no monolithic and static checklists, interdisciplinary.
- **Analytic approach**: any part is independently analysed.
- The team is large international and interdisciplinary, from lawyers to computer scientists.
- It can be applied **at any stage**: design, development, deployment and monitoring.

The process

1. **Set-up phase:**
   - Pre-conditions;
   - Team;
   - Boundaries and context.

2. Assess phase:
- Analyse socio-technical scenarios
- Identification of ethical issues and tensions
- Map to trustworthy AI -> categories of EU HLEG
- Strategy and feedback

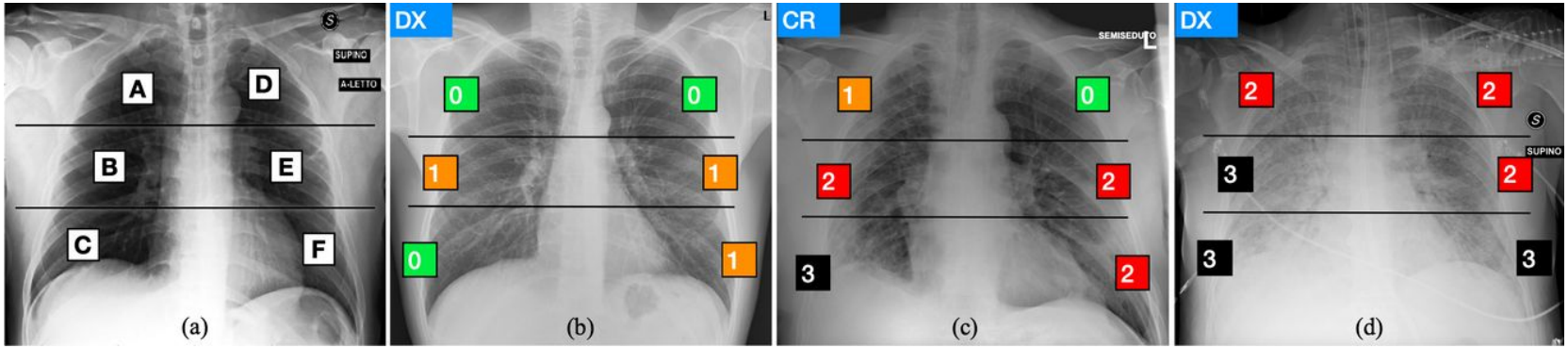
3. Resolve:
- Address ethical tensions
- When possible, give recommendations to relevant stakeholder.

Use case: BS-Net

BS-Net is an end-to-end AI system for the prediction of severity on Chest X-Ray images of COVID-19 patients based on the Brixia Score elaborated by the team of the “Spedali Civili” of Brescia.

BS-Net has been used during the first wave of COVID-19.

The system returns also explanation maps based on a sort of LIME.



A. Signoroni et al., “BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset,” Med. Image Anal., vol. 71, Jul. 2021, Art. no. 102046, doi: 10.1016/j.media.2021.102046.

Socio-technical scenarios (Assess Phase)

We considered 3 possible scenarios in which AI could be used:

1. **The *current* scenario:** single-site deployment, support radiologists by providing a second opinion to reduce errors.
2. ***Possible future applications* of the systems:** web-interface for uploading CXR to be used where radiologists availability is limited.
3. **Another *possible future application*:** use to annotate large datasets and also for retrospective studies.

For each scenario we analyzed:

- the **aim** of the system;
- identification of **actors**: primary, secondary and tertiary -> expectations;
- **context** and processes;
- **technology** used -> detailed analysis of the system;
- AI **design** and trade-offs: example continuous learning;
- **workflow**;
- intellectual property, legal framework and protocols.

Analyses of socio-technical scenarios

Team: more than 50 people! -> Working groups:

Healthcare Radiologists, Healthcare MD, Technical, Legal...

Technical Analysis

1) **Data distribution:** patients collected in one month of the first wave, 5000 CXR for classifier and 1000 CXR for segmentation.
   a) *Small size:* 5000 cases are sufficient to capture all the variance?
   b) *Representational fairness:* patient are "old", gender-biased toward male, ethnicity.
   c) *Limited set of devices*: 3 manufacturers.
2) **Data Labeling:**
   a) *No hard ground truth*: Brixia score is semi-quantitative.
   b) *Score does not describe COVID-19 specifically*
   c) *Potentially biased:* same hospitals and interaction between software developers and Brixia score.
3) **Model definition and maintenance:**
   a) No detailed evaluation of the existing techniques
   b) Subtask may not need AI

Mapping to Trustworthy AI and consolidations

Following the EU guideline, we mapped 3 levels: 1) 4 ethical pillars 2) seven key requirements 3) multiple sub-requirements.

Examples of issues:

Concerns about protection of patients' data
WG: ethics, HC & ethics, technical, social, legal
In brief: informed consent difficult, missing data management plan, anonymized or pseudo?
EP: Prevention of harm, explicability
Req: Privacy and Data Governance, Transparency

System lacks transparency
WG: radiologists, HC, technical
In brief: is patient informed?, no patient history, no COVID-19 specificity.
EP: Prevention of harm, explicability
Req: Technical Robustness and Safety, Transparency

AI system may biased radiologists
WG: radiologists, ethics, social, technical
In brief: MDs see the score before CXR, priming or anchoring effect.
EP: Respect for human autonomy, Fairness
Req: Human Agency and oversight, accountability

Dataset small and not representative
WG: HC, ethics, technical
In brief: origins, age, gender, past medical history, too little diverse
EP: Prevention of harm, Fairness
Req: Diversity, nondiscrimination and fairness, Technical robustness and safety

Recommendations...

1) **Need of a large dataset** with diverse, high-quality images from multiple institutions and different geographic areas to claim generalizability of the AI system.

2) A feedback mechanism to allow the radiologists to review the system output **after reporting.**

3) A study on how AI can be incorporated into **clinical decision making.**

4) A detailed **risk management** plan and governance structure to apply if the AI system is scaled up or expanded.

It is not fair to claim for clinical advancement without a **CLINICAL TRIAL**

Conclusions

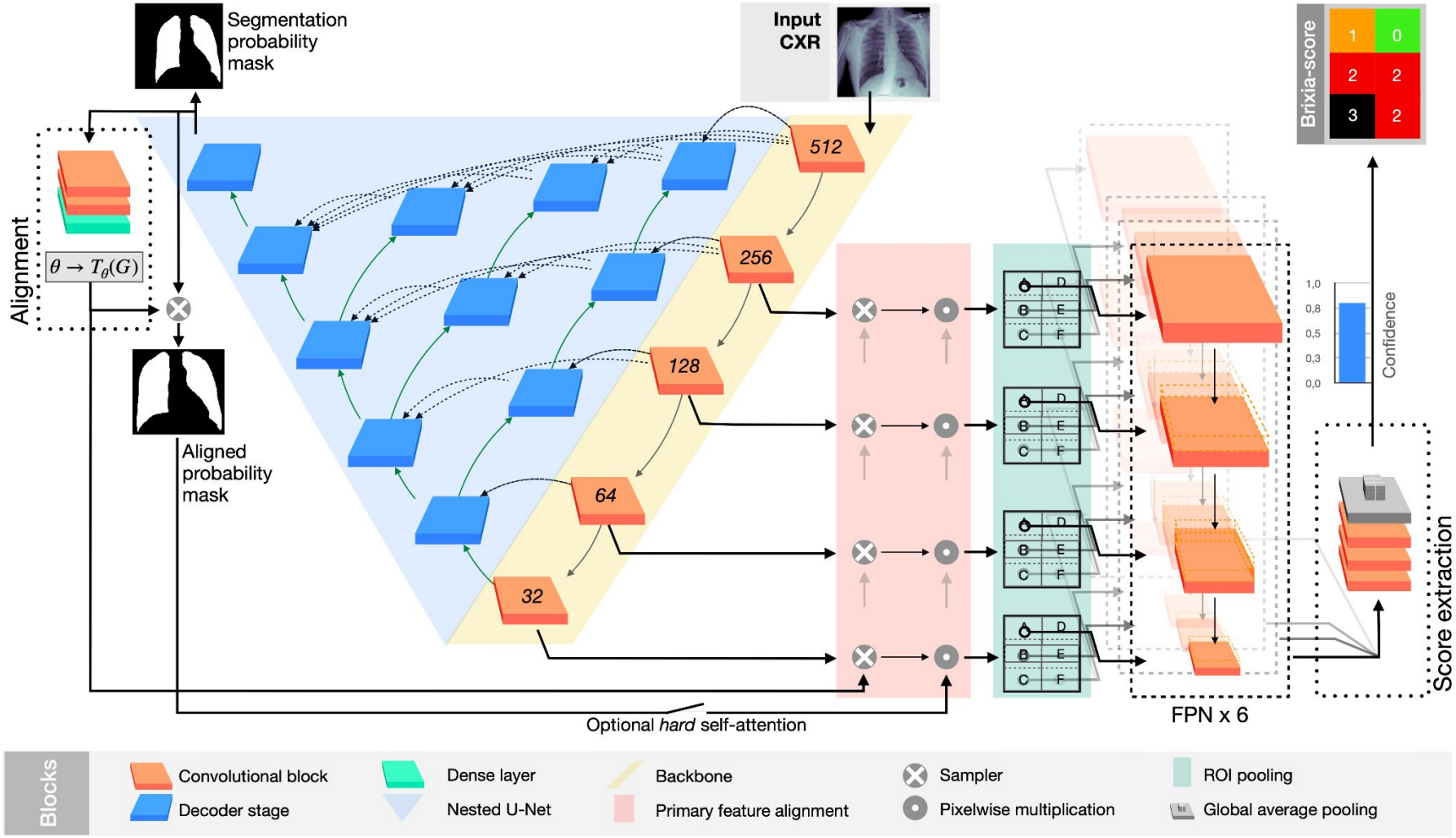
Evaluating Trustworthy AI in practice is hard and it requires a dynamic approach.

Z-inspection project offers the possibility to assess Trustworthy AI at any point of the research.

However, it is a long and complicated process that depends on the team, the problem, the context and so on...

What can we do to build algorithms that are compliant with the EU Guidelines for Trustworthy AI?

Thank you for your kind attention!
Questions?


Segmentation probability mask
Input CXR
Brixia-score
1 0
2 2
3 2
Alignment
$\theta \rightarrow T_\theta(G)$
Aligned probability mask
512
256
128
64
32
A D
B E
C F
1,0
0,8
0,5
0,3
0,0
Confidence
Score extraction
FPN x 6
Optional *hard* self-attention
Blocks
Convolutional block
Decoder stage
Dense layer
Nested U-Net
Backbone
Primary feature alignment
Sampler
Pixelwise multiplication
ROI pooling
Global average pooling