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

Roberto V. Zicari

Z-Inspection® Initiative

http://z-inspection.org

Graduate School of Data Science, SNU, Seoul

Structure of the Lessons

Part I: **Quick intro to our approach.**

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

Part III: **Use Case. On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls.**

Part I: A Quick intro to our approach

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*

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 Four Ethical Principles

Four ethical principles, rooted in fundamental rights

(i) Respect for human autonomy

(ii) Prevention of harm

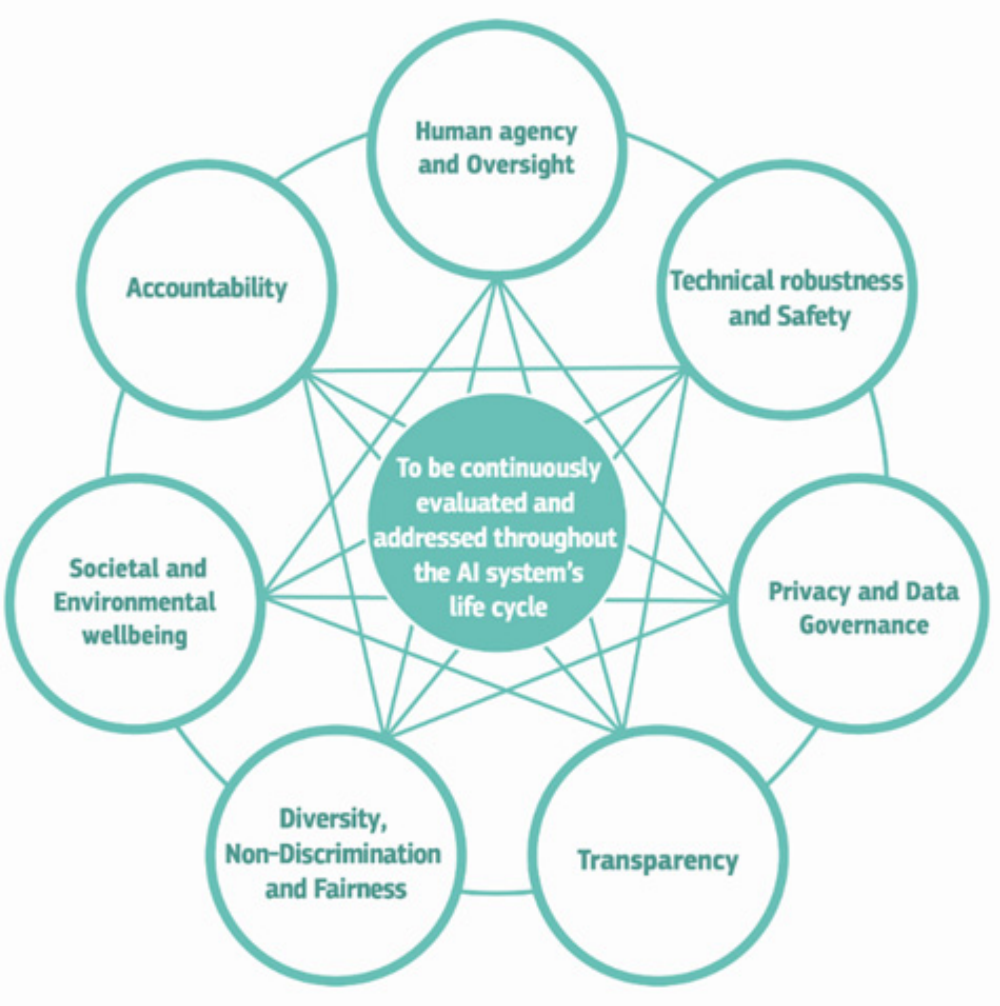
(iii) Fairness

(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 Seven Requirements for Trustworthy AI


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

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

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® is a registered trademark.

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

We use a holistic approach

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

We include a broader set 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)

Tools and other frameworks

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

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

- **Claims, Arguments and Evidence (CAE).**

They can be used in the Z-Inspection® process.

TRUSTWORTHY AI ASSESSMENT LIST : Check List of questions.

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.

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

FRAIA (cont).

- Fundamental right: **does the algorithm affect (or threaten to affect) a fundamental right?**
- Specific legislation: **does specific legislation apply with respect to the fundamental right that needs to be considered?**
- Defining seriousness: **how seriously is this fundamental right infringed?**
- Objectives: **what social, political, or administrative objectives are aimed at by using the algorithm?**

FRAIA (cont).

- Suitability: **is using this specific algorithm a suitable tool to achieve these objectives?**

- Necessity and subsidiarity**: is using this specific algorithm necessary to achieve this objective, and are there no other or mitigating measures available to do so?**

- Balancing and proportionality: **at the end of the day, are the objectives sufficiently weighty to justify affecting fundamental rights?**

Four main fundamental rights clusters

- Fundamental rights relating to **the person**
- **Freedom-related** fundamental rights
- **Equality** rights
- **Procedura**l fundamental rights

Claims, Arguments and Evidence (CAE)

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 subclaims. "

Claims, Arguments and Evidence (CAE)

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)

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.

We give Recommendations

- **Appropriate use;**

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

- **Ability to redress.**

End of Part I

Part II

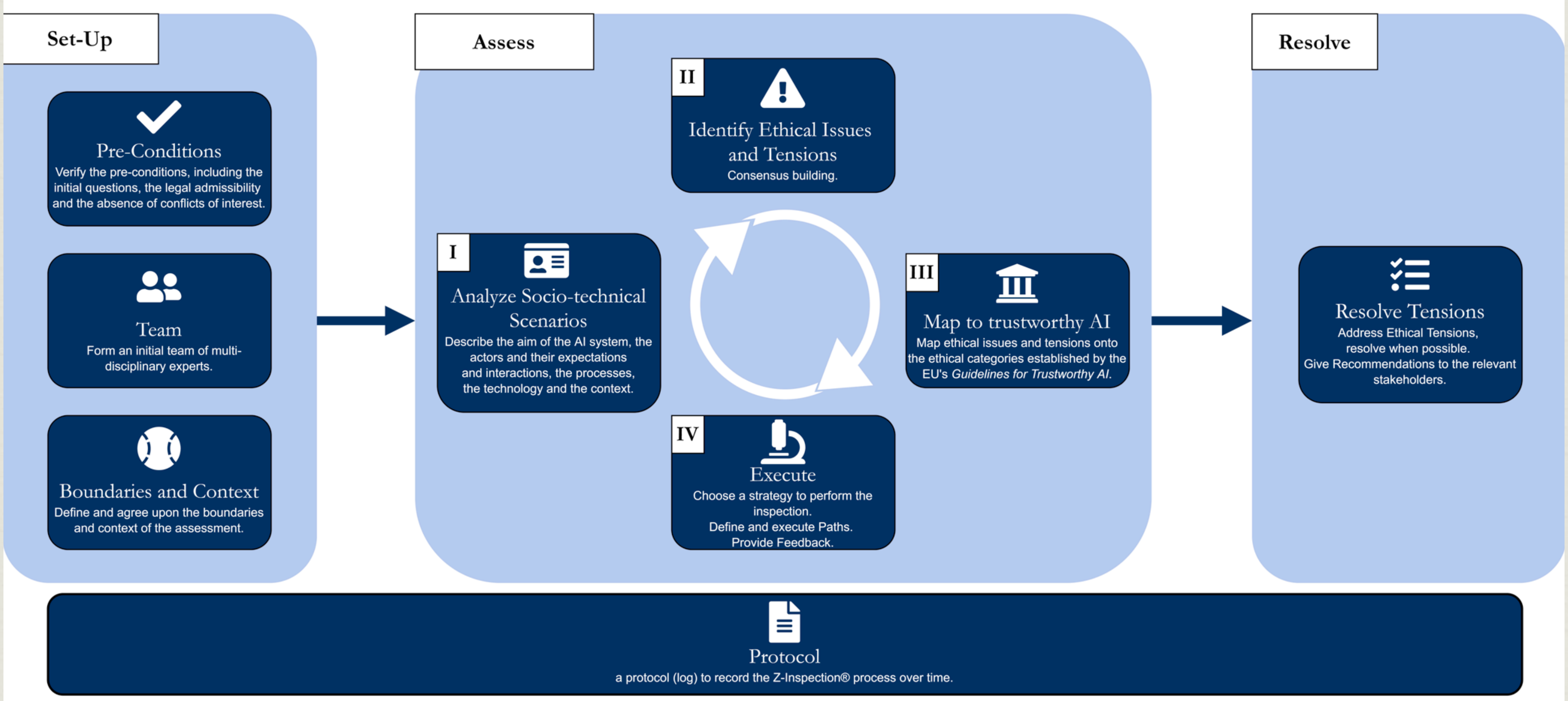
Z-Inspection®: A process to assess trustworthy AI

The Z-Inspection® process

- Z-Inspection® is a holistic process for evaluating the **design, deployment, and use** of AI- based systems towards, aimed at ensuring that the final system iteration is both trustworthy and trusted.

- It can be used at various stages of the AI development and maintenance process.

Z-inspection® Process in a Nutshell



Set-Up

Pre-Conditions
Verify the pre-conditions, including the initial questions, the legal admissibility and the absence of conflicts of interest.

Team
Form an initial team of multi-disciplinary experts.

Boundaries and Context
Define and agree upon the boundaries and context of the assessment.

Assess

I
Analyze Socio-technical Scenarios
Describe the aim of the AI system, the actors and their expectations and interactions, the processes, the technology and the context.

II
Identify Ethical Issues and Tensions
Consensus building.

III
Map to trustworthy AI
Map ethical issues and tensions onto the ethical categories established by the EU's *Guidelines for Trustworthy AI.*

IV
Execute
Choose a strategy to perform the inspection.
Define and execute Paths.
Provide Feedback.

Resolve

Resolve Tensions
Address Ethical Tensions, resolve when possible.
Give Recommendations to the relevant stakeholders.

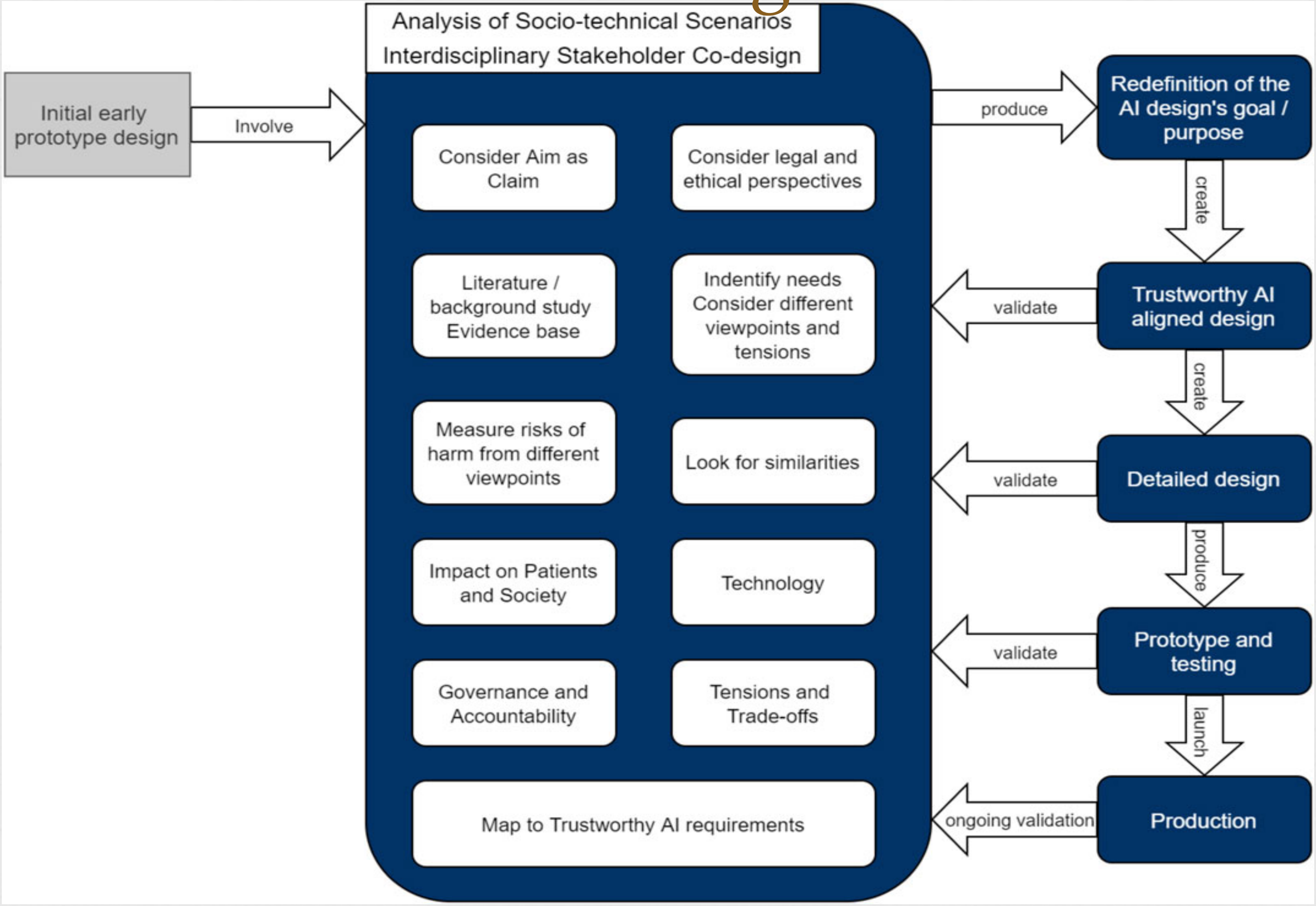
Protocol
a protocol (log) to record the Z-Inspection® process over time.

The Z-Inspection® process: Co-Design

In **design and development phases:**

Z- Inspection® can be used as a **co-creation process** to help AI engineers, domain experts to ensure that the design of their AI system meets the trustworthy AI criteria.

Co-Design



Initial early prototype design

Involve

Analysis of Socio-technical Scenarios
Interdisciplinary Stakeholder Co-design

Consider Aim as Claim

Consider legal and ethical perspectives

Literature / background study Evidence base

Indentify needs Consider different viewpoints and tensions

Measure risks of harm from different viewpoints

Look for similarities

Impact on Patients and Society

Technology

Governance and Accountability

Tensions and Trade-offs

Map to Trustworthy AI requirements

produce

Redefinition of the AI design's goal / purpose

create

validate

Trustworthy AI aligned design

create

validate

Detailed design

produce

validate

Prototype and testing

launch

ongoing validation

Production

When in Co-design.

- **Consider the AI *initial design as a Claim that needs to be verified with evidence.***

- Example: When designing, training and testing an AI-system (e.g. Machine-Learning algorithm) we do "embed" into the system notions such as "**good**", "**bad**", "**healthy**", "**disease**", etc. mostly not in an explicit/transparent way.

The Z-Inspection® process

In deployment and after deployment:

Z-Inspection® can be used as a validation process to assess the trustworthiness of the AI system being developed.

Additionally, it can form part of an AI certification, audit or **monitoring process**.

The latter can be considered a part of "*ethical maintenance*" for trustworthy AI.

When the AI is Deployed

- ***Verify Claims of the producer of the AI with Evidence***

- *Example:* “Embedded” Ethics in AI for healthcare: *Medical Diagnosis*

“Embedded” Ethics in AI for healthcare: *Medical Diagnosis*

"In case medical diagnosis or treatment recommendations are being deferred to machine learning algorithms, **it is the algorithm who sets the bar about how a disease is being defined.”**

“The deployment of machine learning in medicine might resurge the debate between *naturalists* and *normativists*. “

-- Thomas Grote , Philipp Berens

Source: Grote T, Berens P.
J Med Ethics Epub ahead of print: [please include Day Month Year]. doi:10.1136/ medethics-2019-105586

Set Up: 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!**

Practical Suggestion

- 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;

(Social Scientists, Policy Makers, Communication, others**)**

Representative of end users.

Practical Suggestion

- Team members should be selected based primarily on **skills required / expertise – availability and interest in the case**
- Motivation is essential but should not be #1 criteria for involvement.
- Later additions of experts to the team should be limited.

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.
- This could also be done by reading documents backed by online training sessions and/or by producing “video-presentations” available to all the stakeholders related to the assessment.

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.

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)?
- 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?

Set-up: Definition of the boundaries

- 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.

- 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.

How to start?

- Initially the experts team meet together with the stakeholders owning the use case in a number of workshops (via video conference) to define socio-technical scenarios of use of the AI systems.

- We use the term *stakeholders* to denote the actors who have direct ownership on the development and deployment of the AI system.

Practical Suggestions (Optional)

- First presentation of the use case by the use case “owners” to a small but diverse team of Z-Inspection® team members.

- First round of questions. This should be followed by a decision if the use case will be accepted by Z-Inspection® .

- Based on this presentation together with the case owners a decision needs to be taken for the interdisciplinary skills necessary to do a Z-Inspection® .

Small vs. Large Team of Experts

- In one use case, we had a large team of interdisciplinary experts (over 40) and we had to split the work in parallel working groups.

- In other use cases, instead, we had a mid size team (ca. 20) and we did not have to split the work in parallel working groups.

Practical Suggestion

- Consider a small team of inter disciplinary experts (8 to 10) for your first use case.
e.g.:
1 Lead
1 Ethicist
1 Legal
2 Technical
2 Domain
1 Representative of end users.

Practical Suggestion

- Agree on a time frame for inspection with set dates

- Meetings should be agreed with the final inspection team.

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* are 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)*

Implication of IP on Trustworthy AI

- There is an inevitable trade off to be made between disclosing all activities of a trustworthy AI assessment vs. delaying them to a later stage.

Benjamin Haibe-Kains, et al. The importance of transparency and reproducibility in artificial intelligence research. (Submitted on 28 Feb 2020 (v1), last revised 7 Mar 2020 (this version, v2))
https://arxiv.org/pdf/2003.00898.pdf

Create socio-technical scenarios

This is Self-Assessment: By collecting relevant resources, a team of interdisciplinary experts **together with the main stakeholder** 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**.

Socio-technical scenarios

- **Aim of the system**

Goal of the system, context WHY it is used, ...

- **Actors (primary, secondary, tertiary)**

Who is using the system, who is influenced by decisions of the system, who has interest in the system being deployed, ...

Socio-technical scenarios

- **Actors' Expectation and Motivation**

Why would the different groups of actors want the system? What are their expectations towards the system behavior? What benefits are they expecting from using the system? Are there any conflicts?

- **Actors' Concerns and Worries**

What problems / challenges can the actors foresee? Do they have concerns regarding the use of the system?

Socio-technical scenarios

- **Context where the AI system is used**
Additional context for the situations where the AI system is used, how it could be used in the future, ...

- **Interaction with the AI system**
What is the intended interaction between the system and its users? Why is it like this?

Socio-technical scenarios

- **AI Technology used**

Technical description of the AI system, so technically inclined people can get a better intuition how it is working

- **Clinical studies /Field tests**

Was the system's performance validated in (clinical/ test fied) studies? What were the results of these studies? Are the results openly available?

Socio-technical scenarios

- **Intellectual Property**

What IP regulations need to be considered when assessing / disseminating the system? Is it open access? Does it contain confidential information that must not be published?

- **Legal framework**

What is the legal framework for use of the system? What special regulations apply?

Socio-technical scenarios

- Ethics oversight and/or approval
Has the AI system already undergone some kind of ethical assessment or other approval?

If not - why not? If so, was this internal/external, volunteer/regulated, what was covered?

One could argue that if they already had some, it is unlikely they will engage with Z-inspection®, but from asking the question - we also get an understanding of the gap(s).

Did they get a waiver? Was there a clearing, but it was very light or internal and not considered sufficient?

Socio-technical scenarios

- With the information organized this way, people are encouraged to go through the materials again and ask questions in the document where they need additional clarifications from the stakeholders or developers. **This helps bring everyone to a shared understanding of the system which helps create socio-technical scenarios based on possible uses.**
- It also helps stakeholders to identify where they need to provide additional information on the system.

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.

Concept building

1. Mapping and clarifying ambiguities
2. Bridging disciplines, sectors, publics and cultures
3. Building consensus and managing disagreements

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.

Infrastructure

- We have been using Zoom for meetings and Google docs for shared content, Google groups for communication to the entire team and e-mails.
- Good experience in using Zoom, with recorded most of the meeting.
- Mix experience with Google docs. When a document becomes too big with many comments, it is not the optimal tool for co-working
- We did not find a good solution for creating a joint "library" with all relevant articles, working documents etc.
- Mix experience with Google groups and e-mails.

Develop an evidence base

- 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.**

Claims

- “AI developers regularly make **claims regarding the properties of AI systems they develop as well as their associated societal consequences. “**

- Source: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims https://arxiv.org/pdf/2004.07213.pdf

Identify Claims

- **Claims** for technological capability (for example aim, performance, architecture, or functionality, etc.) serve as an important input in developing the **evidence base.**
- This is an iterative process among experts of the assessment team with different skills and backgrounds with a goal to understand technological capabilities and limitations

Verifiable Claims

- „**Verifiable claims** are statements for which **evidence** and **arguments** can be brought to bear on the likelihood of those claims being true.

- The degree of attainable certainty in such claims will vary across contexts. „

- Source: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims

https://arxiv.org/pdf/2004.07213.pdf

Examples of Claims

- We will adhere to the data usage protocols we have specified;

- The cloud services on which our AI systems run are secure;

- We will evaluate risks and benefits of publishing AI systems in partnership with appropriately qualified third parties;

Examples of Claims

- We will not create or sell AI systems that are intended to cause harm;
- We will assess and report any harmful societal impacts of AI systems that we build; and
- Broadly, we will act in a way that aligns with society's interests.

It is highly desirable for claims made about AI development to be verifiable.

“First, those potentially affected by AI development–as well as those seeking to represent those parties’ interests via government or civil society–deserve to be able to scrutinize the claims made by AI developers in order **to reduce risk of harm** or foregone benefit. “

- Source: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims https://arxiv.org/pdf/2004.07213.pdf

It is highly desirable for claims made about AI development to be verifiable

- „Second, to the extent that claims become verifiable, various actors such as civil society, policymakers, and users can raise their standards for what constitutes responsible AI development.
- This, in turn, can **improve societal outcomes** associated with the field as a whole. „

- Source: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims https://arxiv.org/pdf/2004.07213.pdf

It is highly desirable for claims made about AI development to be verifiable

- “Third, a lack of verifiable claims in AI development could foster or worsen a "race to the bottom" in AI development, whereby developers seek to gain a competitive edge even when this trades off against important societal values such as safety, security, privacy, or fairness “

- Source: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims
https://arxiv.org/pdf/2004.07213.pdf

Building a solid knowledge / evidence base

- We suggest **building a solid knowledge / evidence base** among all team members of the use case before the inspection starts and also a solid Q&A log during the inspection process.
- Experts may approach the use case quite differently:
- Interpretations of and expectations for the AI tool being inspected may differ
- Focus of interest may be very different

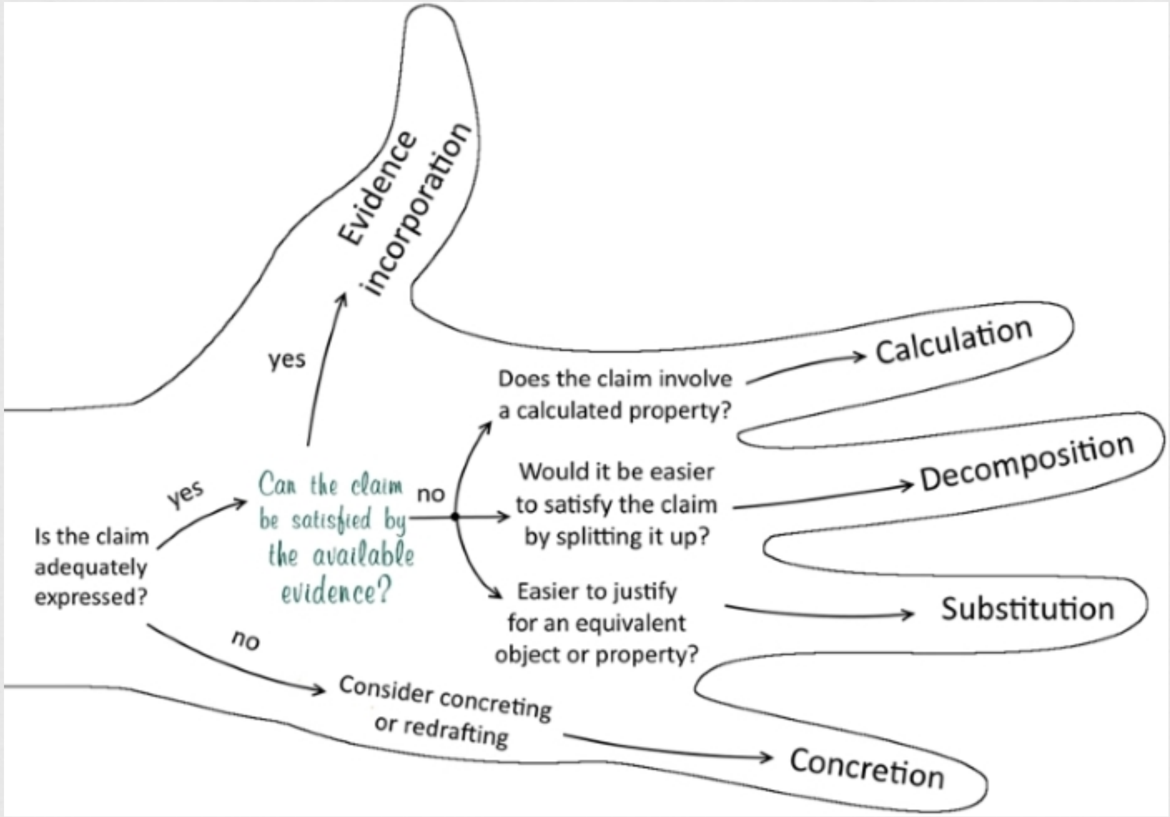
Claims, Arguments and Evidence

- **The claims, arguments and evidence (CAE) framework (*)** can help with the structuring of the use case in a clear and precise form that is supported by evidence.
- For example, each of the claims should be about only one specific property of the system and at the same time, it should be phrased in a way that is clearly verifiable or falsifiable. The CAE framework also provides guidance on how to disseminate complex claims into easier ones .

(*) https://claimsargumentsevidence.org

Claims, Arguments and Evidence

Source: https://claimsargumentsevidence.org



Evidence incorporation
yes
Is the claim adequately expressed?
yes
Can the claim be satisfied by the available evidence?
no
Does the claim involve a calculated property?
Calculation
Would it be easier to satisfy the claim by splitting it up?
Decomposition
Easier to justify for an equivalent object or property?
Substitution
no
Consider concreting or redrafting
Concretion

What is Evidence?

- Who is “qualified” to give strong evidence? We could introduce different levels of what constitutes “evidence”.
- Strong evidence is when testing is possible. However, testing is not always possible. We look at peer-reviewed journal articles supporting a claim. This is also evidence.
- When domain experts have different viewpoints, then we list such different viewpoints and related supporting evidence as tensions.

Different ViewPoints

- Experts in different fields will see the tool quite differently. What may be considered a lack of knowledge can just be a different lens. It's crucial the team understands that there will be very different perspectives based on the specific role or subdomain different experts represent.

Managing Different View Points

- Managing different viewpoints between experts composing the assessment team is an essential part of the process.

- One of the key lessons learned is that there may be tensions when considering what the relevant existing evidence to support a claim is.

Tensions in Evidence Base

- For example in the case of a skin cancer detection AI tool, there were *tensions* between the various arguments linking evidence to support the choice of a design decision derived from the different viewpoints expressed by domain experts.

Tensions in Evidence Base

Claim:

This AI System helps dermatologists to early detection of malignant melanoma.

Argument **:**

Malignant melanoma is a very heterogeneous tumor with a clinical course that is very difficult to predict. To date, there are no reliable biomarkers that predict prognosis with certainty. Therefore, there exist subgroups of melanoma patients with different risks for metastasization, some might never metastasize and diagnosing them would be overdiagnosed.

Tensions in Evidence Base

- *View Point*: **Dermatologis**t.
  Early detection of malignant melanoma is critical, as the risk of metastasis with worse prognosis increases the longer melanoma remains untreated.

- *View Point:* **Evidence Based Medicine Professio**nal.
  There are no reliable biomarkers that can predict the prognosis of melanoma before excision. There are patients who survive their localized melanoma without therapy. **Therefore, the early diagnosis does not necessarily mean a better prognosis; on the contrary, there is a risk of poor patient care due to overdiagnosis.**

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

Describing Issues

- **Use free text and an open vocabulary to describe the possible risks and issues found when analyzing the AI system.**

- The report may list the identified **ethical, technical, domain-specific (i.e. medical) and legal issues described using an open vocabulary**.

Example Ethical Issue

ID Ethical Issue: E4, Fairness in the Training Data

Description

The training data is likely not sufficient to account for relevant differences in languages, accents, and voice patterns, potentially generating unfair outcomes.

Narrative Response (*Open Vocabulary*)

- There is likely empirical bias since the tool was developed in a predominantly white Danish patient group. **It is unclear how the tool would perform in patients with accents, different ages, sex, and other specific subgroups.**

Narrative Response (*Open Vocabulary*)

- **There is also a concern that this tool is not evaluated for fairness with respect to outcomes in a variety of populations.** Given the reliance on transcripts, non-native speakers of Danish may not have the same outcome. It was reported that Swedish and English speakers were well represented but would need to ensure a broad training setmay not have a diverse enough representation.

Narrative Response (*Open Vocabulary*)

- It would also be important to see if analyses show any **bias** in results regarding age, gender, race, nationality, and other sub-groups. The concern is that the training data

Example

Possible Risks and Harm: False Positives and False Negatives

We could not find a justification for choosing a certain balance between sensitivity and specificity.

- If *specificity* is too low, CPR is started on people who do not need it and administered CPR over a longer period of time can lead to rib cage fractures, for example. However, it is unlikely that CPR would be performed on a conscious patient for a longer time, as the patient probably would fight back against it.
- If *sensitivity* is too low, cardiac arrests may not be detected. This results in no CPR being administered and the patient remains dead. In this context "too low" is when the AI system performs poorer than the dispatchers, hence will not be of any help. The AI system is evaluated against human performance, as this system is only useful if it can assist humans; otherwise, it is just a distraction.

Example

Lack of Explainability

- Our team of experts did not sign a Non Disclosure Agreement (NDA) with the vendor company, and that means that the AI system is considered a "black box," with no details of the implementation of the AI algorithms and the AI model. To avoid possible conflict of interests, no direct communication between our team of experts and the vendor company was (and is) taking place.
- The prime stakeholder cooperates with the vendor company, and they have declared no conflict of interest with them.
- The main issue here is that it is not apparent to the dispatchers how the AI system comes to its conclusions. It is not transparent to the dispatcher whether it is advisable to follow the system or not. Moreover, it is not transparent to the caller that an AI system is used in the process.

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.* "

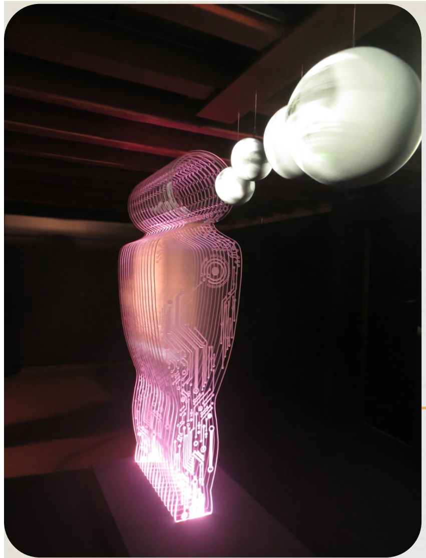
source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

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.



Identify Ethical Issues and Tensions, and Flags

- An **Ethical** issue or tension refers to different ways in which values can be in conflict.

- A **Flag** is an issue that needs to be assessed further.
  (it could be a technical, legal, ethical issue)

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

From [1]:

- ***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***

[1] 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.

Resources

http://z-inspection.org