A few Notes on XAI Validation

Ioanna Chouvarda
Aristotle University of Thessaloniki, Greece

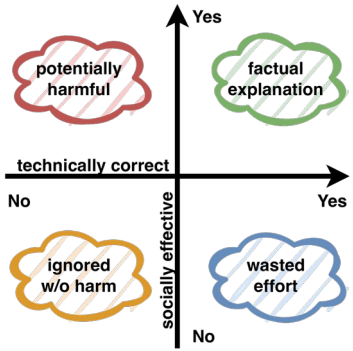
contents

XAI validation & evaluation, users and evaluation goals

Main ideas: Why validate

Validation of the explainability components is an important aspect in the field of XAI, to ensure that the insights provided are

a) accurate and
b) practical to the end-user.



Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand. (A. Barredo Arrieta, N. Díaz-Rodríguez and J. Del Ser et al, 2020)

the necessity to explain

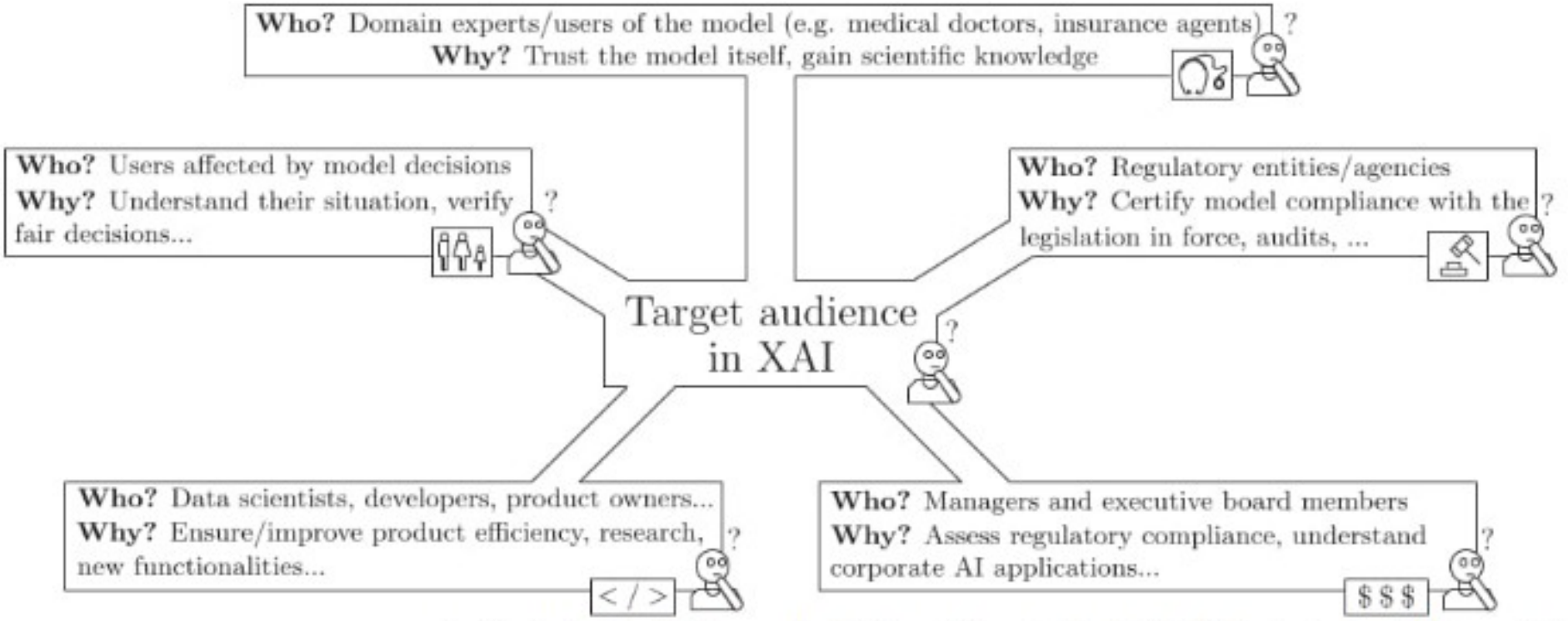
explain to justify - the decisions made by utilising an underlying model should be explained in order to increase their justifiability;

explain to control - explanations should enhance the transparency of a model and its functioning, allowing its debugging and the identification of potential flaws;

explain to improve - explanations should help scholars improve the accuracy and efficiency of their models;

explain to discover - explanations should support the extraction of novel knowledge and the learning of relationships and patterns.

A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," in *IEEE Access*, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.



Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Volume 58, 2020, Pages 82-115, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.12.012.

Main ideas: Who is the user of XAI

Medical context explanations

- technical correctness
- meaningful and actionable information to medical experts, and all involved users.

Users

- **AI novices** refer to end-users who use AI products in daily life but have no (or very little) expertise on machine learning systems.
- **Data experts** include data scientists and domain experts who use machine learning for analysis, decision-making, or research.
- **AI experts** are machine learning scientists and engineers who design machine learning algorithms and interpretability techniques for XAI systems, for model interpretation or instance explanations.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact. Intell. Syst. 11, 3–4, Article 24 (December 2021), 45 pages. https://doi.org/10.1145/3387166

Main ideas: What is the focus per user

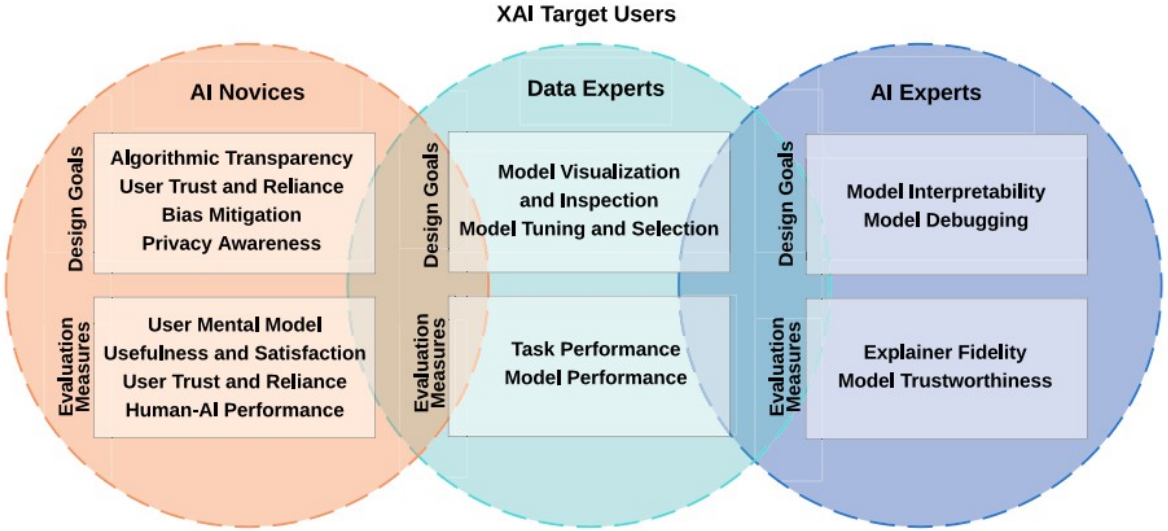


Fig. 3. A summary of our categorization of XAI design goals and evaluation measures between user groups. **Top:** Different system design goals for each user group. **Bottom:** Common evaluation measures used in each user group. Notice that similar XAI goals for different user groups require different research objectives, design methods, and implementation paths.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact. Intell. Syst. 11, 3–4, Article 24 (December 2021), 45 pages. https://doi.org/10.1145/3387166

General Evaluation goals

1. **User mental model** : **Measures what an explanation contributes to how well the user understands the system.** User Understanding of Model: interview, Likert-scale Questionnaire . Subjective and objective measures for understandability, usefulness

2. **Usefulness and satisfaction** : **Measures to what extent a user perceives a benefit from the provided explanation.** Sufficiency of details to assess explanatory value for users, questionnaires and interviews, Task Duration and Cognitive Load

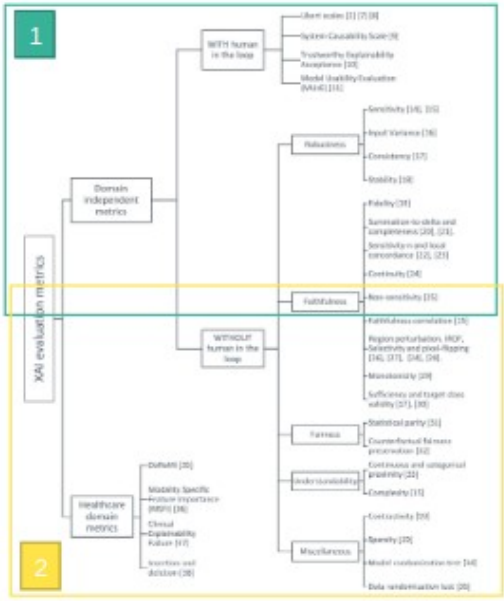
3. **User trust and reliance** : **Measures to what extent the explanation improves the user's trust in the system.** Affective and cognitive factor that influences positive or negative perceptions of a system - development of trust over time Subjective Measures (Interview, Likert scale) / Objective Measures (eg Compliance, Understandability)

4. **Human-AI task performance** : **Measures to what extent the explanation helps the combination of human + AI to perform a task.**
Relevant to all three groups of user types, users' performance in terms of success rate and task completion time to evaluate the impact of different types of explanations.

5. **Computational Measures** : **Measures to what extent the explanation is correct and complete with computational approaches.**
correctness and completeness in terms of explaining what the model has learned, correctness of generated explanations, consistency of explanation results, and fidelity of ad-hoc interpretability techniques to the original black-box model

Some Basic Concepts and terms

XAI metrics e



" When an explanation is not enough: An overview of evaluation metrics of explainable AI systems in the healthcare domain", E. Pietilä and P.A Moreno-Sanchez, Medicon 2023, paper accepted.

Domain independent metrics with human in the loop

Likert scales for conducting user studies

System Causability Scale (SCS) proposed by Holzinger et al. which is used to **evaluate the perceived explainability and ensure that the XAI system fits its purpose.** The SCS consists of 10 questions:

1. “I found that the data included all relevant known causal factors with sufficient precision and granularity.”
2. “I understood the explanations within the context of my work.”
3. “I could change the level of detail on demand.”
4. “I did not need support to understand the explanations.”
5. “I found the explanations helped me to understand causality.”
6. “I was able to use the explanations with my knowledge base.”
7. “I did not find inconsistencies between explanations.”
8. “I think that most people would learn to understand the explanations very quickly.”
9. “I did not need more references in the explanations: e.g., medical guidelines, regulations.”
10. “I received the explanations in a timely and efficient manner.

Holzinger A., Carrington A., Müller H., Measuring the Quality of Explanations: The System Causability Scale (SCS). KI - Künstliche Intelligenz. 2020; 34: 193-198 https://doi.org/10.1007/s13218-020-00636-z

challenges around XAI validation from a clinical study

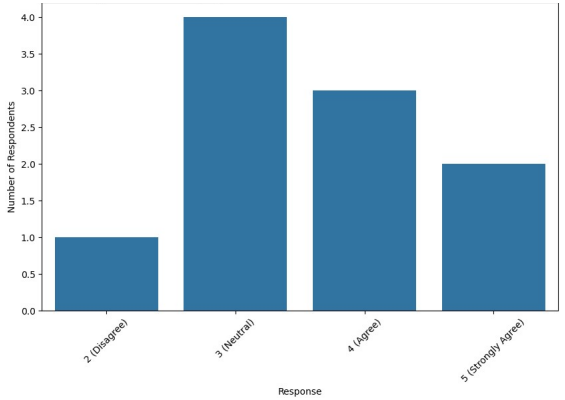
a positive reception of the XAI component and a general agreement that the explanations help in understanding, trusting and using the services of the toolbox.

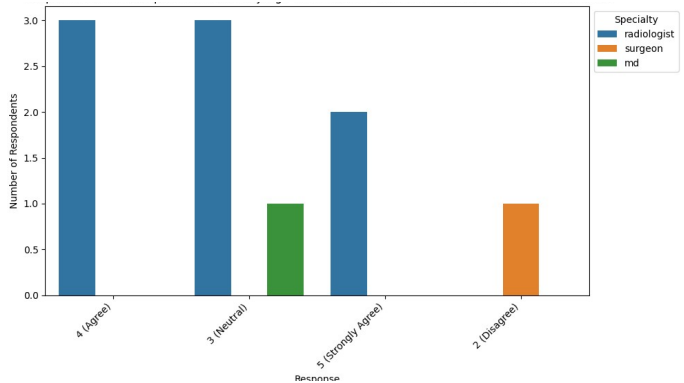
Significant portion of neutral responses ?

radiologists have a more favorable view of the XAI explanations in terms of trust, accuracy and completeness, while surgeon and the MD exhibit a wider range of opinions, with a trend towards neutrality or even disagreement on several aspects.

the XAI components may need adjustment to better fit to other medical experts' requirements and needs

This explanation lets me judge when I should trust or not trust the tool





Domain independent metrics with human in the loop

Trustworthy Explainability Acceptance : evaluate the Euclidean distances between explanations provided by an expert Xi and the explainability model Yi.

V. K. Venugopal, R. Takhar, S. Gupta, and V. Mahajan, ‘Clinical Explainability Failure (CEF) & Explainability Failure Ratio (EFR) – Changing the Way We Validate Classification Algorithms’, J. Med. Syst., vol. 46, no. 4, p. 20, Mar. 2022, doi: 10.1007/s10916-022-01806-2.

Kaur, D., Uslu, S., Durresi, A., Badve, S., & Dundar, M. (2021). Trustworthy Explainability Acceptance: A New Metric to Measure the Trustworthiness of Interpretable AI Medical Diagnostic Systems. In L. Barolli, K. Yim, & T. Enokido (Eds.), Complex, Intelligent and Software Intensive Systems (pp. 35–46). Springer International Publishing.

Healthcare metric with human in the loop

Clinical Explainability Failure and Explainability Failure Ratio : developed for saliency maps in healthcare applications but are applicable to other fields as well.

- Clinical Explainability Failure:
  - step1 : a model's output is compared to ground truth to see if it recognises the correct features.
  - step2. If the bounding box inserted on the saliency map does not match the ground truth and an expert fails to understand why, the model is deemed to have made a Clinical Explainability Failure.

- Explainability Failure Ratio is calculated as the amount of Explainability Failures divided by the total amount of explanations.

V. K. Venugopal, R. Takhar, S. Gupta, and V. Mahajan, 'Clinical Explainability Failure (CEF) & Explainability Failure Ratio (EFR) – Changing the Way We Validate Classification Algorithms', J. Med. Syst., vol. 46, no. 4, p. 20, Mar. 2022, doi: 10.1007/s10916-022-01806-2.

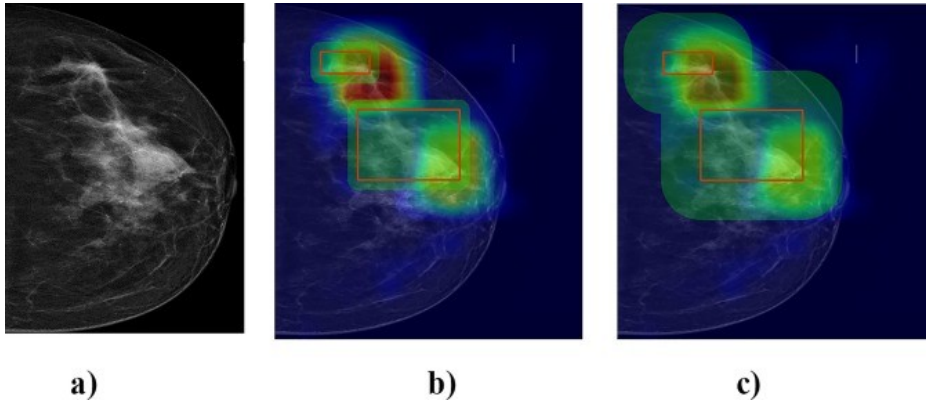
Pointing Game metric

Cerekci, Esma et al. **Quantitative evaluation of Saliency-Based Explainable artificial intelligence (XAI) methods in Deep Learning-Based mammogram analysis,** European Journal of Radiology, Volume 173, 111356

while saliency-based methods offer explainability to a certain degree, they often fall short in precisely delineating how DL models arrive at a decision in numerous instances.

necessity for researchers and clinicians alike to approach them with a clear understanding of their limitations.

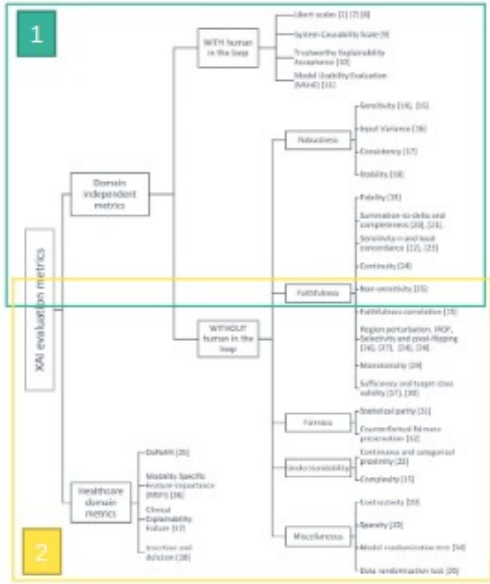
https://www.ejradiology.com/article/S0720-048X(24)00072-X/fulltext#secst090

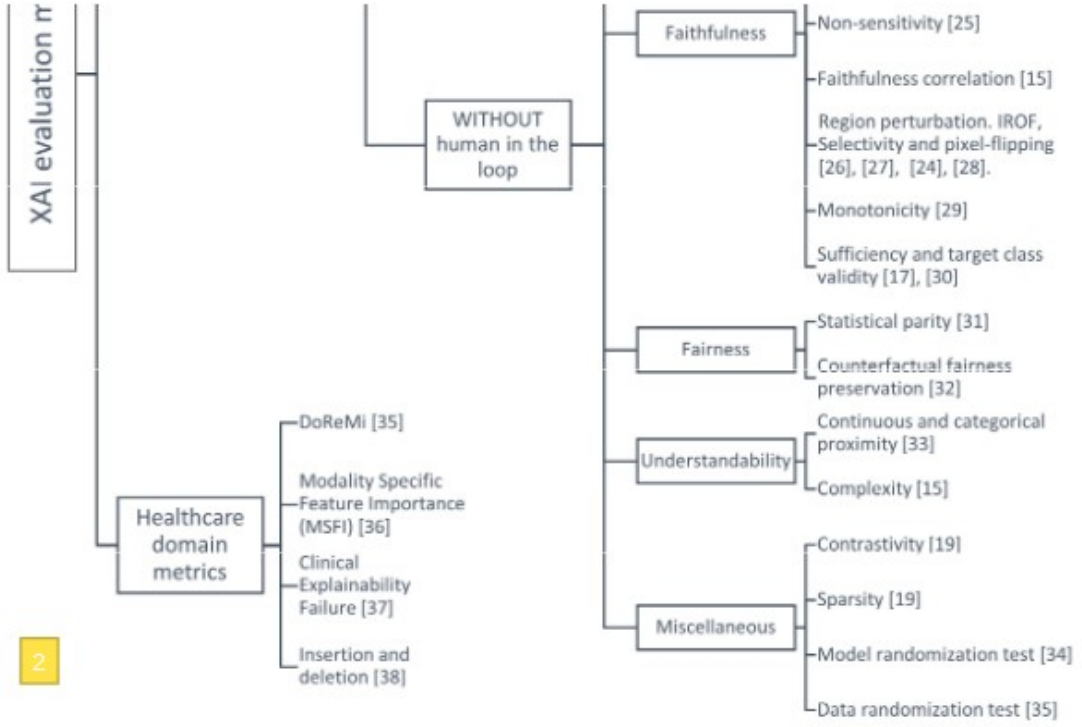


Pointing Game metric with different τ values. We illustrate the influence of τ parameter on pointing game score.

Original image is shown in Fig. a. Pointing game aims to evaluate the saliency map methods by focusing on a neighborhood around the peak value. We showcase the influence of varying τ (15,60) on an example case. The τ parameter is adjustable and gives a certain degree of offset with respect to the ground-truth boxes in calculating whether the peak value of the saliency map lies within the ground-truth bounding boxes. In this example image, the green rounded rectangle around the ground-truth boxes represents the area where if the peak value of the saliency map overlaps, it is accepted as hit. As the τ increases, this area increases, and the chance of a hit also increases. We used the default τ of 15 in this study. The ground-truth boxes (a), τ of 15 (b), and 60 (c) are shown.

XAI metrics evaluation





" When an explanation is not enough: An overview of evaluation metrics of explainable AI systems in the healthcare domain", E. Pietilä and P.A Moreno-Sanchez, Medicon 2023, paper accepted.

Domain independent metrics without human in the loop : Robustness

Explanation Robustness: how well the explanation holds when minor changes are made in the input.

If robust, not sensitive to noise and able to detect the underlying reasons for the predictions

else explaining the noise...

- **Consistency :** how probably instances with the same explanation are given the same prediction. indicates that explanations are not randomly generated.
- **Sensitivity** : how the explainability method reacts to small changes in the input. calculated as the distance between explanations for an instance and its perturbations.

Domain independent metrics without human in the loop : Faithfulness

Explanation faithfulness Necessary for a credible explanation: the right properties are explained and feature importances are correct. Very popular. Features are removed from the input in relevance order derived from the explanation and different calculations are conducted based on the output.

- **Fidelity** : the loss of accuracy of an AI model if features with saliency values higher than 0.01 are removed. Saliency map includes selected extremely relevant features?
  - Faithfulness Correlation: iteratively replaces a random subset of given attributions with a baseline value and then measuring the correlation between the sum of this attribution subset and the difference in function output
  - Faithfulness Estimate: computes the correlation between probability drops and attribution scores on various points
- **Completeness** : how much of the black box behavior is described in the explanation.
  - Check if Feature importances of explanation are correct by checking that the total attribution in an image explanation corresponds to the difference between prediction F at instance x and the determined baseline. This is a sanity check for explainers based on feature importance and desirably the value should be low.
- **Continuity** : quantifies how continuous an explanation function is. This is done by calculating the strongest variation over all inputs.

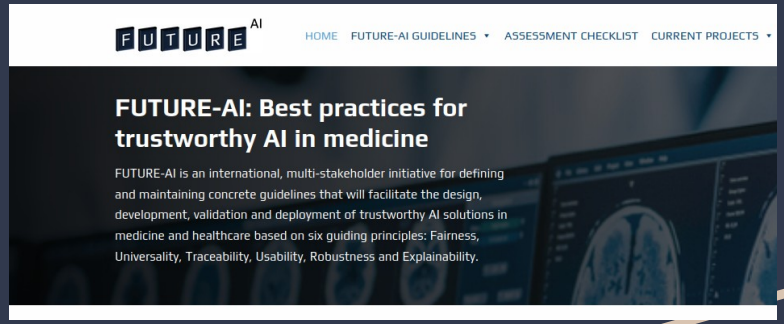
Domain independent metrics without human in the loop : understandability

Understandability. Even if the explanation was faithful and robust, it cannot be explainable if the user is uncapable of understanding it. Although understandability is often measured with users in the loop, there are some metrics to evaluate understandability without involving users.

- **Complexity** is a metric to check how many features of the input are included in the explanation. Although including all the features would result in a more accurate explanation, Bhatt et al. argue that it would make the explanation more difficult to understand.
- **Continuous proximity and categorical proximity** evaluate understandability of counterfactual explanations.
  - counterfactual that are similar but not identical to the original instance would be the most useful to the user and make the explanation most understandable.
  - Calculated as the negative vector distance between the instance and its counterfactual's features.

Guidelines and Recommendations

Future-AI



https://future-ai.eu/principle/explainability/

Recommendation Description

Explainability 1

Define explainability needs
At the design phase, it should be established if explainability is required for the AI tool. In this case, the specific requirements for explainability should be defined with representative experts and end-users, including
(i) the goal of the explanations (*e.g.* global description of the model's behaviour vs. local explanation of each AI decision),
(ii) the most suitable approach for AI explainability, and
(iii) the potential limitations to anticipate and monitor (*e.g.* over-reliance of the end-users on the AI decision).

Explainability 2

Evaluate explainability
The explainable AI methods should be evaluated, first quantitatively by using *in silico* methods to assess the **correctness** of the explanations, then qualitatively with end-users to assess their impact on **user satisfaction, confidence and clinical performance**.
The evaluations should also identify any **limitations** of the AI explanations (*e.g.* they are clinically incoherent or sensitive to noise or adversarial attacks, they unreasonably increase the confidence in the AI-generated results).

Short title	Question(s)	Example
Identifying explainable biomarkers	To increase clinical value, did you evaluate if the explainability methods enable to **identify variables or features that can serve as biomarkers**? Did you determine if the identified imaging biomarkers are previously known?	Yes, a qualitative analysis of attribution maps revealed that the model uses the skin outside the lesion for the diagnosis of pigmented actinic keratosis. Yes, this is an already known biomarker.

Short title	Question(s)	Example
Quantitative evaluation of explainability	Did you use some quantitative evaluation tests to determine if the explanations are **robust and trustworthy**?	Yes, I performed model randomisation tests, data randomisation tests, reproducibility tests and determined Area Over Perturbation Curve (AOPC) for the quantitative evaluation of the attribution maps. https://github.com/CristianCosci/AOPC_MoRF

Short title	Question(s)	Example
Qualitative evaluation of explainability	Did you perform some **qualitative** evaluation tests with clinicians?	Yes, the System Causability Scale (SCS) was used by my clinical collaborators to rate the explanations.

Short title	Question(s)	Example
Robustness of explainability against adversarial attacks	Did you evaluate **robustness to adversarial attacks**, by assessing if the explanations remain consistent when the input images are subjected to small input perturbations and noise?	Yes, I applied small input perturbations to the input image and added noise to generate new input images indistinguishable from the original. I found that the classifications did not change but the attribution maps highlighted very different areas of the tissues than before.

Short title	Question(s)	Stage	Example
Explainability in clinical practice	Did you evaluate the effect of explainability methods in clinical practice by performing a **collaborative human-AI study** in which the doctor performs clinical task using the AI tool with and without explanations? Did you identify any resulting **bias** from the introduction of the explainability methods?	6,7	Yes, the radiologists utilised the AI tool with and without the explanations for my diagnosis AI tool. The human-AI collaboration led to better detection of pathological cases when the explanations were also available. However, it also led to increased false positives as the attribution maps influenced the clinicians. The area highlighted by attribution maps was interpreted as diseased tissues.

(6) AI evaluation, and (7) AI deployment and monitoring.

Some State of Art and some facts

state of art 1 review

M. Nauta *et al.*, From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. arXiv, May 31, 2022. Accessed: Jun. 20, 2022. [Online]. Available: http://arxiv.org/abs/2201.08164.

not specifically in medicine

ACM Computing Surveys Volume 55Issue 13s Article No.: 295, 2

statistics on XAI evaluation:

- 33% only evaluated with anecdotal evidence
- 58% applied quantitative evaluation
- 22% evaluated with human subjects in a user study, of which 23% evaluated with domain experts, i.e. application grounded

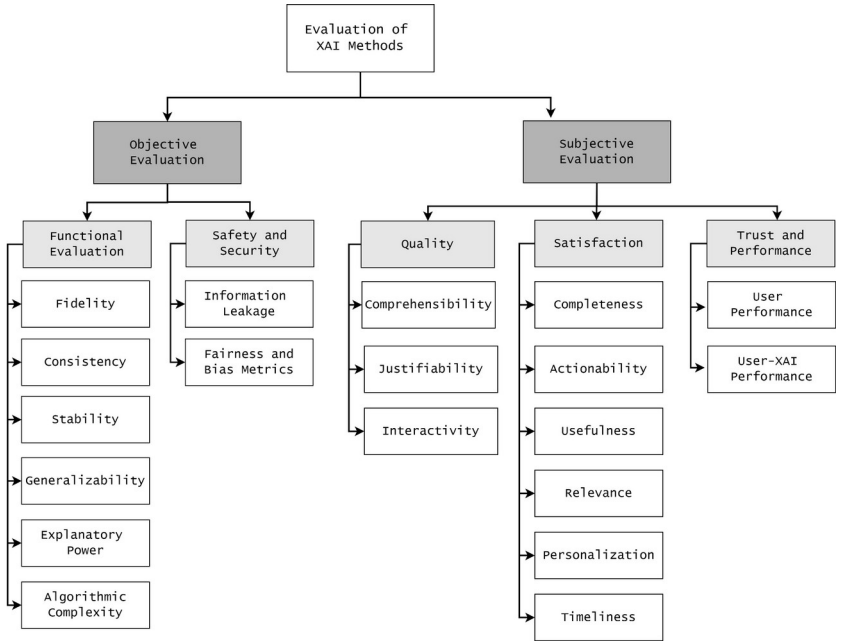
Groen AM, Kraan R, Amirkhan SF, Daams JG, Maas M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? Eur J Radiol. 2022 Dec;157:110592. doi: 10.1016/j.ejrad.2022.110592. Epub 2022 Nov 5. PMID: 36371947.

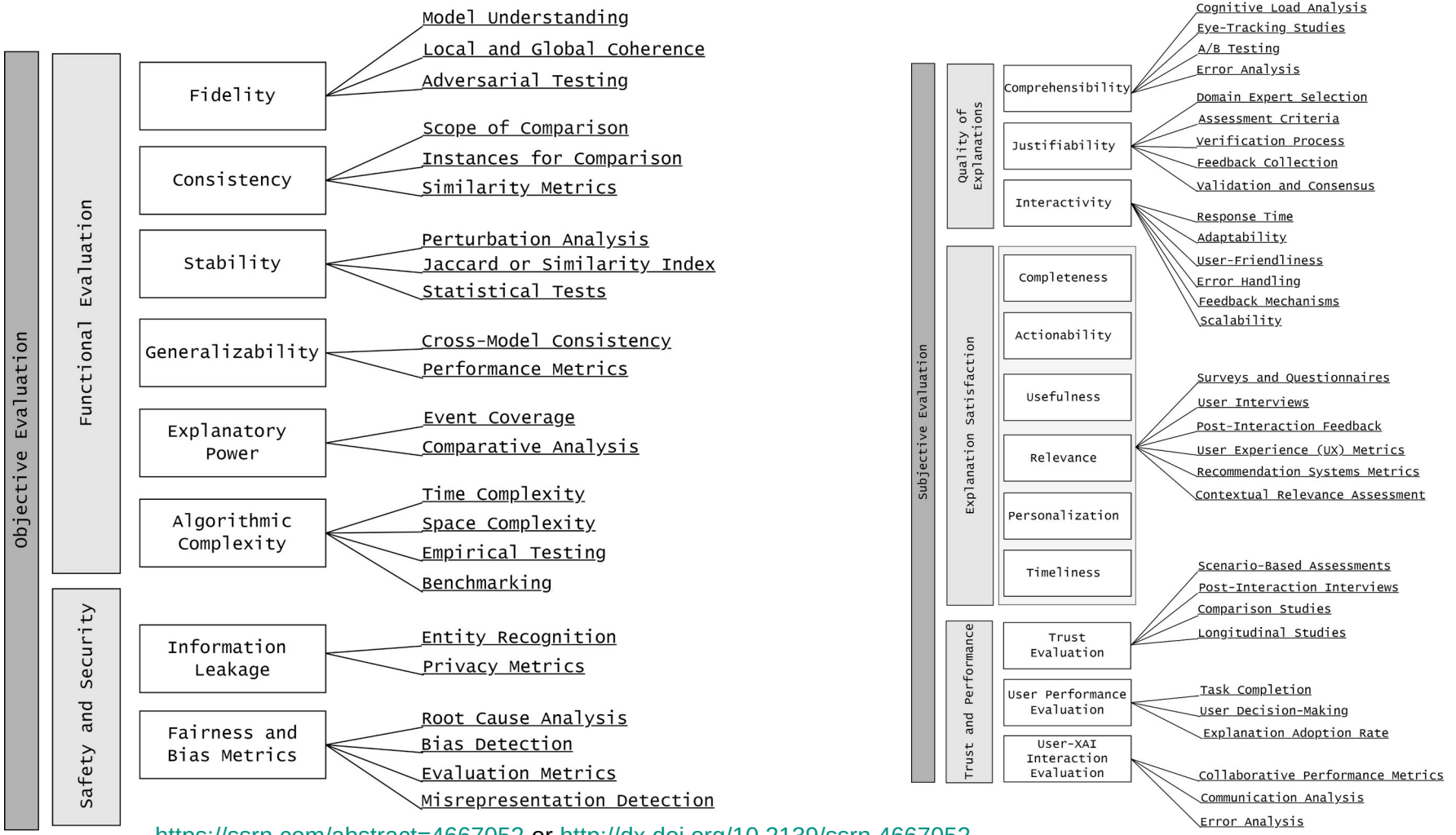
'Researchers are agnostic about the quality of the resulting explanations because XAI evaluation measures are lacking.'

Only 1 study used measures to evaluate the outcome of their explainable AI.

state of art 2 guidelines

Tekkesinoglu, Sule, Exploring Evaluation Methodologies for Explainable AI: Guidelines for Objective and Subjective Assessment (December 17, 2023). Available at SSRN: https://ssrn.com/abstract=4667052 or http://dx.doi.org/10.2139/ssrn.4667052





https://ssrn.com/abstract=4667052 or http://dx.doi.org/10.2139/ssrn.4667052

state of art 3 review - XAI in healthcare

to what extent the properties of explainability, i.e. **interpretability** (consisting of clarity and parsimony) and **fidelity** (consisting of completeness and soundness), are satisfied for model-based, attribution-based, and example-based explanations.

quantitative evaluation metrics, which are important for objective standardized evaluation, are still lacking for some properties (e.g. clarity) and types of explanations (e.g. example-based methods).

Aniek F. Markus, Jan A. Kors, Peter R. Rijnbeek,The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies, Journal of Biomedical Informatics,
Volume 113, 2021, 103655, ISSN 1532-0464,
https://doi.org/10.1016/j.jbi.2020.103655.
https://www.sciencedirect.com/science/article/pii/S1532046420302835#b0060

the goal of evaluation methods is twofold.

***allows a formal comparison of available explanation methods**. Many methods have been proposed, often with a similar goal, but it is unclear which one is to be preferred. When evaluating post-hoc explanations, the problem is there is no ground truth, as we do not know the real inner workings of the model.*

*offers a formal method to **assess if explainability is achieved in an application**. Here the focus lies on determining if the offered form of explainability achieves the defined objective.*

state of art 4 XAI methods in Healthcare

Essi Pietilä , METHODS AND METRICS FOR EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE DOMAIN

- Methods to evaluate XAI for healthcare applications with expert in the loop
- Metrics to evaluate XAI for healthcare applications without expert in the loop

54 metrics and methods were found for evaluating XAI

faithfulness of an explanation the most popular aspect of explainability - 22 metrics found

Many metrics are similar - particularly for faithfulness.

Involvement of experts in the evaluation loop varies greatly between studies. Including experts in evaluating healthcare XAI is often encouraged (be resource conscious!) .

A clear need for standardisation of XAI evaluation particularly in healthcare domain:
what to evaluate and with which methods and metrics.

https://trepo.tuni.fi/bitstream/10024/147475/2/PietilaEssi.pdf&ved=2ahUKEwjR8ejwxJ6GAxUg6AIHHbJcCaoQFnoECBoQAQ&usg=AOvVaw0c97suD6kt8mJXDNdGt-PH

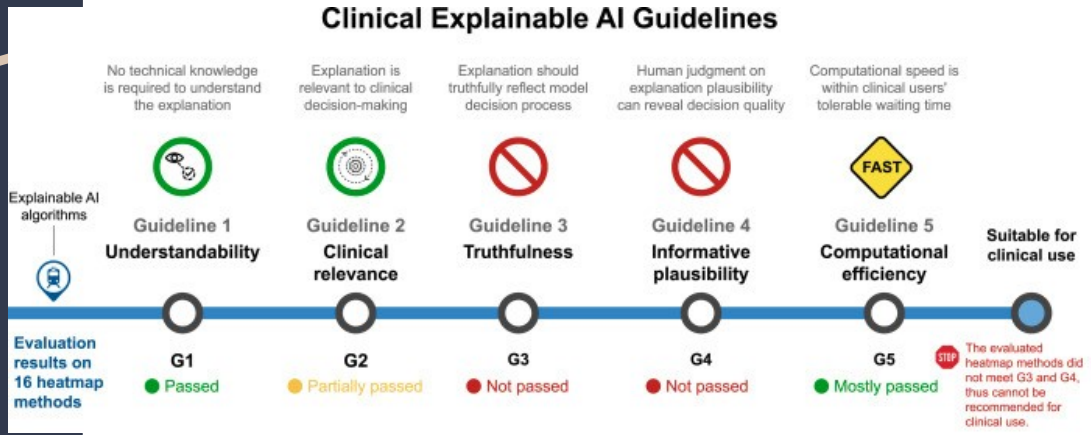
state of art 5
The Clinical Explainable AI Guidelines

Weina Jin, Xiaoxiao Li, Mostafa Fatehi, Ghassan Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, Volume 84, 2023, 102684, ISSN 1361-8415, https://doi.org/10.1016/j.media.2022.102684.

Clinical usability
G1: Understandability
Explanations should be easily understandable by clinical users without requiring technical knowledge.
G2: Clinical relevance
Explanation should be relevant to physicians' clinical decision-making pattern, and can support their clinical reasoning process.

Evaluation
G3: Truthfulness
Explanations should truthfully reflect the AI model decision process. This is the prerequisite for G4.
G4: Informative plausibility
Users' judgment on explanation plausibility may inform users about AI decision quality, including potential flaws or biases.

Operation
G5: Computational efficiency
The speed to generate an explanation should be within clinical users' tolerable waiting time on the given task.



Toolboxes

https://pypi.org/project/quantus/0.1.2/

https://github.com/amparore/leaf

https://github.com/SinaMohseni/Awesome-XAI-Evaluation

https://github.com/Oxid15/xai-benchmark

https://github.com/chus-chus/teex

https://pypi.org/project/xai-metrics/

Sum Up

Could it be That an Explanation Does Not Increase Trust Because the Explanations are Too Technical for People to Understand?
https://technologyandsociety.org/human-centricity-in-the-relationship-between-explainability-and-trust-in-ai/

XAI validation in the medical domain is evolving

Not much evidence yet

Many terms and metrics

Need for Standardisation

Need for linking of XAI evaluation to the other aspects of trustworthiness

Thanks!

- Ioanna Chouvarda
- Aristotle University of Thessaloniki, Greece
- ioannach@auth.gr

Questions

https://ahaslides.com/ONOX7