






Towards an Evaluation Framework for Explainable Artificial Intelligence Systems for Health and Well-Being

Esperança Amengual-Alcover¹^a, Antoni Jaume-i-Capó¹^b, Miquel Miró-Nicolau¹^c,
Gabriel Moyà-Alcover¹^d and Antonia Paniza-Fullana²^e

¹Department of Mathematics and Computer Science, University of the Balearic Islands,
Ctra. de Valldemossa, Km. 7.5, 07122 - Palma de Mallorca, Spain

²Department of Private Law, University of the Balearic Islands, Ctra. de Valldemossa,

{eamengual, antoni.jaume, miquel.miro, gabriel.moya, antonia.paniza}@uib.es

Keywords: Explainable Artificial Intelligence (XAI), Explainability, Health and Well-Being, XAI Evaluation Measurements, Evaluation Framework.


Abstract: The integration of Artificial Intelligence in the development of computer systems presents a new challenge: make intelligent systems explainable to humans. This is especially vital in the field of health and well-being, where transparency in decision support systems enables healthcare professionals to understand and trust automated decisions and predictions. To address this need, tools are required to guide the development of explainable AI systems. In this paper, we introduce an evaluation framework designed to support the development of explainable AI systems for health and well-being. Additionally, we present a case study that illustrates the application of the framework in practice. We believe that our framework can serve as a valuable tool not only for developing explainable AI systems in healthcare but also for any AI system that has a significant impact on individuals.


1 INTRODUCTION


The third wave of Artificial Intelligence (AI) systems is characterized by two key aspects: (1) technological advancements and diverse applications and (2) a human-centred approach (Xu 2019). While AI is achieving impressive results, these outcomes are often challenging for human users to interpret. To trust the behaviour of intelligent systems, especially in health and well-being, they need to clearly communicate the rationale behind their decisions and actions. Explainable Artificial Intelligence (XAI) aims to meet this need by prioritizing transparency, enabling AI systems to describe the reasoning behind their decisions and predictions.


A growing interest in XAI has been reflected in several scientific events (Adadi and Berrada 2018; Alonso, Castiello, and Mencar 2018; Anjomshoae et


al. 2019; Biran and Cotton 2017; Došilović, Brčić, and Hlupić 2018) and in the relevant increase of recent reviews about the topic (Abdul et al. 2018; Alonso, Castiello, and Mencar 2018; Anjomshoae et al. 2019; Chakraborti et al. 2017; Došilović, Brčić, and Hlupić 2018; Gilpin et al. 2018; Murdoch et al. 2019), particularly in the health and well-being areas (Mohseni, Zarei, and Ragan 2021; Tjoa and Guan 2021) XAI is emerging as a new discipline in need of standardized practices. The diverse goals, design strategies, and evaluation techniques used in XAI have resulted in a range of approaches for creating explainable systems (Mohseni, Zarei, and Ragan 2021). Murdoch et al. (Murdoch et al. 2019) propose a broad categorization of XAI methods into model-based and post-hoc techniques. However, as discussed in (Mohseni, Zarei, and Ragan 2021), achieving effective XAI design requires an integrated

^a <https://orcid.org/0000-0002-0699-6684>

^b <https://orcid.org/0000-0003-3312-5347>

^c <https://orcid.org/0000-0002-4092-6583>

^d <https://orcid.org/0000-0002-3412-5499>

^e <https://orcid.org/0000-0002-1302-9713>

approach that considers the dependencies between design goals and evaluation methods.

In this work, we present an evaluation framework that aims to guide the design of explainable AI systems for health and well-being, with an emphasis on legal and ethical issues. We illustrate our approach through a case study on medical image analysis, an area where we have previous experience. In healthcare, decision-making based solely on unexplainable predictions is insufficient to meet ethical and legal standards. Explainability helps to rationalize AI driven diagnoses, treatment plans, and disease predictions, enhancing understanding for both professionals and patients. Indeed, explainability is recognized as an essential ethical principle for AI systems, ensuring their transparency for end-users (Alcarazo 2022). This principle aligns with the European Parliament's "Report on Artificial Intelligence in a Digital Age" (VOSS, n.d.), which emphasizes transparency and explainability as foundational.

Our goal is to ensure that intelligent systems for medical image analysis adhere to these ethical standards and relevant regulations. Given the significant impact of AI-driven decisions on individuals, it is crucial to protect personal data and inform patients about how these systems are used. Furthermore, medical professionals must be able to comprehend the reasoning behind an AI system's conclusions to understand the logic guiding its predictions and recommendations.

2 PREVIOUS WORK

In (Mohseni, Zarei, and Ragan 2021), a generic framework for designing XAI systems is presented, offering multidisciplinary teams a high-level guideline for developing domain-specific XAI solutions. According to its authors, the framework's flexibility makes it broadly applicable, enabling customization to address specific needs across various fields. In our work, we build upon this framework to integrate design guidelines specifically tailored to meet the requirements of health and well-being applications. Figure 1 provides an overview of the original framework, which serves as the basis for our domain-focused extension. The layered structure links core design goals and evaluation priorities across different research communities, promoting multidisciplinary progress in the field of XAI systems. This structure supports the design steps, starting with the most external level (XAI System Objectives), then considering the needs of end users

at the intermediate level (Explainable Interfaces), and finally focusing on the interpretable algorithms at the most internal level (Interpretable Algorithms). The framework suggests iterative cycles of design and evaluation, enabling comprehensive consideration of both algorithmic and human-centred aspects.

In the case of our framework for XAI systems in health and well-being, the process begins with an existing AI system and a set of XAI methods previously applied for verification purposes. For this reason, our focus is on the Evaluation Pole, with the goal is of assessing the AI system and provide insights to improve explainability tasks within the Design Pole.

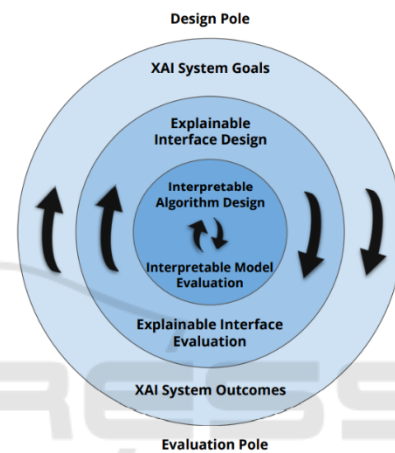


Figure 1: Generic XAI Design and Evaluation Framework (extracted from (Mohseni, Zarei, and Ragan 2021)).

3 XAI EVALUATION MEASURES

We propose to classify the different explanation qualities three main categories: Machine-Centred Features, Human-Centred Features, and Social-Centred Issues. These categories address different key aspects for assessing XAI systems.

3.1 Machine-Centred Features

Machine-Centred features focus on exclusively algorithmic aspects, independent of external evaluators.

Fidelity, defined as the correctness of a method in generating true explanations for model predictions (Mohseni, Zarei, and Ragan 2021), is the most extensively studied feature in this category. Fidelity evaluation methods can be divided into:

- **Synthetic Attribution Benchmarks (SABs):** These consist of datasets with ground truth

explanations, created under controlled scenarios. SABs help to identify incorrect methods but cannot confirm their correctness.

Various methodologies exist for generating these datasets, including those proposed by (Arias-Duart et al. 2022; Arras et al. 2017; Cortez and Embrechts 2013; Guidotti 2021; Mamalakis, Barnes, and Ebert-Uphoff 2022; Miró-Nicolau, Jaume-i-Capó, and Moyà-Alcover 2024a).

- **Post-Hoc Fidelity Metrics:** These metrics approximate fidelity in real-world scenarios where a ground truth explanation is absent. Several authors, including (Alvarez Melis and Jaakkola 2018; Bach et al. 2015; Samek et al. 2017; Rieger and Hansen 2020; Yeh et al. 2019), have proposed post-hoc fidelity metrics. However, these metrics have been criticized for unreliable results (Hedström et al. 2023; Miró-Nicolau, Jaume-i-Capó, and Moyà-Alcover 2025; Tomsett et al. 2020).

Robustness is defined as the expectation that minor changes in input data yield similar explanations (Alvarez Melis and Jaakkola 2018). Robustness metrics have been proposed by (Agarwal et al. 2022; Alvarez Melis and Jaakkola 2018; Dasgupta, Frost, and Moshkovitz 2022; Montavon, Samek, and Müller 2018; Yeh et al. 2019).

Complexity refers to the amount of variables used in an explanation. Its complementary feature, **sparsity**, ensures that only the truly predictive features contribute to the explanation (Chalasani et al. 2020).

Localisation test whether the explanation is focused on a specific region of interest. (Hedström et al. 2023).

Randomisation assesses how explanations degrade when data labels or model parameters are randomised, as explored by (Hedström et al. 2023).

3.2 Human-Centred Features

Human-Centred Features focus on subjective elements dependent on user interaction with XAI systems. These features, studied beyond AI, have been analysed in social and behavioural sciences (Hoffman et al. 2018; Miller 2019). Key features include **mental models**, **curiosity**, **reliability**, and **trust**, with trust being central to evaluating XAI systems (Barredo Arrieta et al. 2020; Miller 2019). Trust is critical in automation (Adams et al. 2003; Lee and See 2004; Mercado et al. 2016) and is often measured through various scales. (Jian, Bisantz, and Drury 2000) propose measuring trust through an 11-

items scale, which has become a de facto standard due to its wide use and influence on other scales.

User Trust is defined as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (Lee and See 2004). While trust is inherently subjective, measuring it objectively is desirable. (Mohseni, Zarei, and Ragan 2021) identify scales and interviews as subjective methods, while (Scharowski et al. 2022) advocate for behavioural measures. (Lai and Tan 2019) propose measuring trust by the frequency of user reliance on system predictions, finding that users trust correct predictions more. Furthermore, (Lai and Tan 2019; Miró-Nicolau et al. 2024) introduce a trust measure that integrates performance and trust data using a confusion matrix, resulting in four distinct measures. These measures, inspired by the well-established True Positive, True Negative, False Positive, and False Negative metrics from classification tasks, provide insights into the interplay between system performance and user trust, allowing for more complex measures, such as Precision and Recall.

3.3 Legal and Ethical Issues

Legal and ethical considerations are essential in ensuring transparency and traceability of data and operations within intelligent systems, especially for compliance with regulations and providing legal certainty. In the context of health-related data, it is crucial to address potential biases in algorithms.

According to the General Data Protection Regulation (GDPR) (Union 2016), personal data refers to any information that identifies or can identify a natural person, making privacy a key concern. The Spanish Data Protection Agency highlights the importance of identifying personal data processing, profiling, or decision making related to individuals, which mandates compliance to data protection laws.

While AI systems can use anonymous data, transparency and explicability of algorithms remain essential. Article 78 of the GDPR stipulates that developers should ensure compliance with data protection when designing, selecting, or using applications that process personal data. The principles of **privacy by design** and **privacy by default** (Article 25 of the GDPR) must be prioritized. Additionally, third parties, such as medical personnel, must understand the IA system, its algorithms, and outputs to prevent harm (Justa 2022). These requirements have significant ethical and legal implications, particularly regarding liability.

In 2024, the European Commission approved the AI act, regulating AI technologies within Europe ('Regulation (EU) 2024/1689 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) – European Sources Online' 2024). This law introduces a risk-based classification system, categorizing AI applications, as Unacceptable, High, Limited, or Minimal risk. Applications in the unacceptable risk category are prohibited, while high-risk and limited-risk applications must meet specific requirements, including human oversight and robustness. Thus, the need for transparency and reliable XAI remains paramount.

Furthermore, the European Commission guidelines for Trustworthy AI ('Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment | Shaping Europe's Digital Future' 2020) define seven requirements for reliable 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) societal and environmental well-being; and (7) accountability. These criteria must be evaluated throughout the AI system's lifecycle to ensure legal and ethical compliance.

4 XAIHEALTH

In the previous section, we analysed various metrics for evaluating XAI systems, highlighting the significant diversity of approaches used to assess different aspects of explainability. We classified these approaches into three categories: machine-centred, human-centred, and social-centred features. However, it's important to note that these evaluation aspects cannot be assessed simultaneously due to their interdependencies. For instance, if an explanation lacks fidelity to the underlying causes of an AI prediction, the user's trust in that explanation becomes irrelevant, as the prediction itself may be incorrect. Therefore, a comprehensive evaluation of XAI systems must be conducted sequentially.

The framework proposed by (Mohseni, Zarei, and Ragan 2021) requires adaptation for specific contexts, as noted in the introduction. In this section, we present a tailored adaptation of this framework specifically for the healthcare domain. Given that AI systems in healthcare are classified as high-risk under

the EU AI Act, they necessitate rigorous verification and monitoring. Our adaptation addresses the unique requirements and challenges associated with integrating XAI into these high-stakes environments.

To facilitate this, we propose a new evaluation framework, named XAIHealth, designed to effectively assess XAI systems within the context of health and well-being. Figure 2 illustrates the adaptation in relation to the foundational framework.

As shown in Figure 2, XAIHealth centres its approach on the Evaluation Pole tasks defined in the base framework (see Figure 1). Each layer, moving from the innermost to the outermost level, comprises two phases: machine-centred analysis and human-centred assessment. These phases are preceded by an initial pre-evaluation phase, which includes training the AI model and applying a XAI method. Legal and ethical considerations are cross-cutting elements that must be addressed throughout the development, evaluation and deployment of the system. Figure 3 illustrates the phases and process flow of the XAIHealth framework.

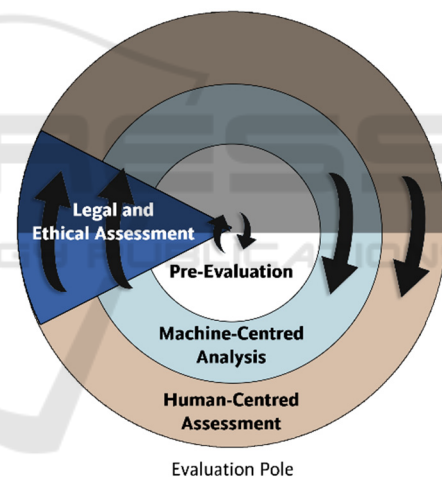


Figure 2: XAIHealth phases in the evaluation pole.

4.1 Legal and Ethical Assessment

In our framework, ethical and legal factors are integrated into every phase, emphasizing the importance of privacy, data management, and transparency for reliable AI systems in alignment with current legislation.

For privacy and data management, the most restrictive regulations should be prioritized, with European legislation (notable the GDPR) as a reference point due to its rigorous standards. According to Article 4.2 of the GDPR, both, profiling and decision-making about individuals are considered forms of data processing. Health data,

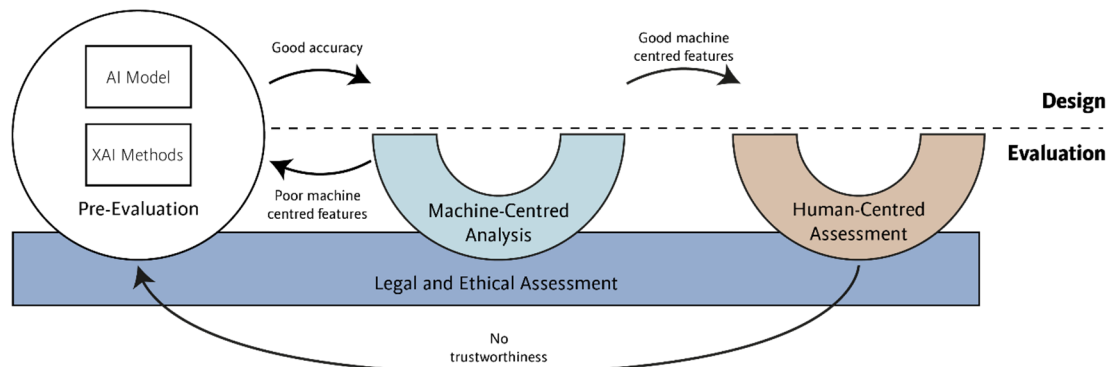


Figure 3: XAIHealth phases and process flow.

categorized as special personal data under Article 9 of the GDPR, requires additional protections. Either anonymous data should be used, or a legal basis must be established for processing. In our case, consent from affected individuals is a viable approach for ensuring compliance.

Transparency is addressed in Article 78 of the GDPR, which mandates that participants must be informed if AI systems will be used in decision-making processes that affect them.

To assess both legal compliance and ethical standards, we propose using the ALTAI (Assessment List for Trustworthy Artificial Intelligence) guidelines ('Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment | Shaping Europe's Digital Future' 2020), developed by an expert group commissioned by the European Commission (AI HLEG). These guidelines operationalize the Ethics Guidelines for Trustworthy AI and align with the principles of the EU AI act.

The ALTAI document identifies seven essential requirements for building trustworthy AI systems. Each requirement entails specific evaluation criteria based on the context. Below, we summarize these requirements:

- (1) **Human Agency and Oversight.** This ensures that AI systems respect user autonomy and informs users of AI involvement. Six key questions guide the evaluation of whether user autonomy is preserved and whether users are aware they are interacting with AI.
- (2) **Technical Robustness and Safety.** This requirement assesses the AI system's resilience to adversarial inputs, novel data, and cybersecurity risks. It focuses on robustness, particularly the system's ability to maintain performance against adversarial alterations (Goodfellow, Shlens, and Szegedy 2015), and varying input data.
- (3) **Privacy and Data Governance.** Focused on DGPR compliance, this requirement emphasizes

the proper handling and protection of data. Its primary goal is to ensure that data governance practices align with legal privacy standards.

- (4) **Transparency.** Transparency refers to how well AI system's processes can be understood. ALTAI specifies that transparency combines explainability with effective communication about the AI system's functions and limitations.
- (5) **Diversity, Non-Discrimination and Fairness.** This requirement addresses the minimization of biases and the promotion of inclusive design. It seeks to eliminate discriminatory elements by ensuring accessibility and universal design principles are considered.
- (6) **Societal and Environmental Well-Being.** This requirement evaluates the AI system's broader impact, including environmental sustainability, effects on employment and skills, societal influence, and democratic processes. Evaluations are task-dependent and aim to mitigate potential societal risk.
- (7) **Accountability.** This requirement focuses on establishing mechanisms for risk management and accountability throughout the system's lifecycle. It enables ongoing monitoring to detect potential errors and risk-prone behaviours.

Each of these requirements aligns with the phases of our proposed framework. As we present each phase, we will outline the relevant ALTAI requirement and discuss strategies for ensuring that the AI system meets these standards. This approach aims to integrate compliance with ethical and legal guidelines into each step of the framework.

4.2 Phase 0 – Pre-Evaluation

The Pre-Evaluation phase begins with an existing AI system that has demonstrated adequate efficacy. A set of XAI methods is selected and applied to this system

for verification/assessment purposes. The goal is to evaluate holistically the XAI system.

During this phase, three key requirements from the ALTAI guidelines must be addressed: **Privacy and Data governance**, **Diversity, Non-discrimination and Fairness**, and **Societal and Environmental Well-Being**. We will illustrate how each of these requirements is applied in our case study in the following subsection.

4.2.1 Application in Case Study

In the Pre-Evaluation phase, we assess an AI system designed to identify signs of pneumonia in chest x-ray images. Specifically, we utilized the AI model proposed in Miró-Nicolau et al. (Miró-Nicolau, Jaume-i-Capó, and Moyà-Alcover 2024a), which proposed a new approach to measure trust in AI systems within a healthcare context. With this experimental setup, the authors trained a well-known Convolutional Neural Network (CNN), ResNet18, introduced by (He et al. 2016). This AI system was trained using a supervised learning approach on a labelled dataset, which includes inputs paired with the correct outputs. The model was trained on 2048 x-ray images from *Hospital Universitari Son Espases* (HUSE), featuring cases of COVID-19 pneumonia and non-pneumonia cases. The system achieved an accuracy of 0.8, a measure of the ratio of correct predictions of total samples. To provide explanations, the GradCAM algorithm (Alvarez Melis and Jaakkola 2018) was applied to produce heatmaps showing which areas of the images were most influential in the model's predictions, giving insight into the model's decision-making process.

From a legal and ethical perspective, the following three requirements from ALTAI must be addressed in this phase:

- **Privacy and Data Governance.** Compliance with GDPR is a primary goal. In our case study, the Research Commission from *Hospital Universitari Son Espases* (HUSE) verified data compliance, ensuring all data was anonymized. Thus, we confirm that the Privacy and Data Governance requirement is satisfied.
- **Diversity, Non-Discrimination and Fairness.** This requirement, closely related to privacy and data governance, focuses on ensuring that anonymized x-ray images are free from bias. Accessibility and universal design were also considered in the GUI design process, even so these elements fall outside the scope of the evaluation framework. We conclude that the goal

of avoiding discrimination has been met in this case.

- **Societal and Environmental Well-Being.** The model's purpose is to identify COVID-19 pneumonia in patients, and in this context, three main factors must be evaluated: its environmental impact, implications for democracy, and broader societal influence. Firstly, it is clear that this application does not have any direct impact on democratic processes. Similarly, its effect on society is limited, as the system is designed as a diagnostic support tool to enhance medical practice rather than to change social structures.

In terms of environmental impact, the primary concern lies in the energy consumption involved, particularly during the models' training phase, which is generally the most resource-intensive aspect of the process. However, in this case, energy use has been minimized due to two factors: the adoption of a relatively small model (ResNet18) and the use of a modest dataset (2048 images). Thus, we assert that the model meets requirements for environmental sustainability.

4.3 Phase 1 – Machine-Centred Analysis

The primary objective of this phase is to evaluate the machine-centred features of the XAI system, focusing on algorithmic attributes that can be measured independently of the end-user. Fidelity is considered the most crucial metric in this category, as shown by its predominance in the state-of-the-art. However, calculating fidelity is challenging due, as (Hedström et al. 2023) note “since the evaluation function is applied to the results of the unverifiable explanation function, the evaluation outcome also renders unverifiable”. To address this, we consider necessary to adopt a validated post-hoc measurement approach, using only validated metrics to avoid unreliable evaluations.

In our analysis, we excluded unvalidated metrics, therefore the lack of verification in existing post-hoc fidelity metrics makes them unsuitable for real-world applications. (Hedström et al. 2023) proposed a meta-evaluation process that reviewed several machine-centred metrics, revealing both strengths and weaknesses. However, none of the ten metrics analysed achieved perfect results, meaning fidelity and similar features may not yet be reliably usable in real-world contexts.

Among the machine-centred features, robustness has also been largely studied. Miró-Nicolau et al. (Miró-Nicolau, Jaume-i-Capó, and Moyà-Alcover

2024b), developed a set of tests for assessing robustness metrics, initially applying them to the AVG-Sensitivity and MAX-Sensitivity by (Yeh et al. 2019). We extended these tests to include other robustness metrics, identifying Local Lipschitz Estimate (LLE) by (Alvarez Melis and Jaakkola 2018) as a reliable and practical option, being . LLE is the only metric that passed both robustness tests, making it our primary criterion for assessing explanation robustness. However, if additional metrics are developed and validates in the future, they can be incorporated into the framework without requiring further modifications.

This phase also addresses two ALTAI requirements: Technical Robustness and Safety and Transparency. If issues arise during this phase, it may be necessary to revisit the Pre-evaluation phase to determine whether the shortcomings stem from the AI model or the XAI method itself.

4.3.1 Application in Case Study

To evaluate the robustness of explanations in our case study, we applied the Local Lipschitz Estimate (LLE) proposed by (Alvarez Melis and Jaakkola 2018) to GradCAM generated explanations. The optimal LLE score is 0, representing maximum robustness, while 1 is the worst possible outcome. We established a flexible acceptance threshold, considering an explanation robust if its results fall within the top 10%. This was because the fact that even a perfect XAI algorithm may exhibit slight robustness limitations due to the underlying AI model itself, making a minor margin of error acceptable. We obtained a mean LLE value of 0,082 with a standard deviation of 0,108. This results show that the system's robustness falls within the accepted threshold, fulfilling the Technical Robustness and Safety requirement of the ALTAI guidelines. Additionally, Transparency – a core goal of any XAI approach and a key component in mitigating black-box limitations – is addressed by the GradCAM method, which offers an interpretable heatmap-based explanation of model predictions. These results confirm that the system is ready to proceed to the next evaluation phase.

4.4 Phase 2 - Human-Centred Assessment

The purpose of this phase is to evaluate human-centred features of the XAI method which are directly influenced by the end-user's experience. To account for this, an interface that displays both the AI

system's prediction and the accompanying explanation form the XAI method must be utilized.

This phase encompasses several aspects, with trust being a primary focus in the XAI field, as highlighted by Miller (Miller 2019). Trust can be assessed from two perspectives: as an attitude (the user's self-perception) or as a behaviour (following the AI system's advice), according to (Scharowski et al. 2022). These two approaches are clearly related to different evaluation methods: objective (behavioural) or subjective (attitudinal).

To measure trust effectively, we adopt pre-existing metrics reviewed in the previous section. We recommend the metric proposed by (Miró-Nicolau et al. 2024), which uses a behavioural approach that incorporates the AI system's prediction performance into the trust evaluation.

If the outcomes of this phase indicate issues, it may be necessary to revisit the pre-evaluation phase to identify whether issues stem from the interface design or broader XAI system limitations.

In this phase, the ALTAI requirement of Human Agency and Oversight must be addressed, primarily by ensuring users are informed at all times that they are interacting with an AI system. This requirement closely aligns with the focus of this phase, which emphasizes user interface design and clear information communications between the user and the AI system.

4.4.1 Application in Case Study

In our case study, which used x-ray images, we assessed whether radiologists and other specialists trusted the AI system's outputs. Results of this phase, along with the trust measurement, were published by Miró-Nicolau et al. (Miró-Nicolau et al. 2024). The design team developed an interface that presented the AI system's prediction and the XAI method-generated explanation simultaneously to the end-users. To simplify understanding, the explanations were refined by removing less significant pixels using various threshold values. Specialists then rated their level of trust in the combined diagnosis and explanation, with their responses assessed using (Miró-Nicolau et al. 2024) trust metric.

The Human Agency and Oversight ALTAI requirement is fully consistent with the interface design in this case study. Specifically, the interface clearly communicates that the AI model and XAI method are in use, fulfilling this ALTAI requirement.

The outcomes of this phase are summarized in Table 1. The table shows trust metrics combining classification accuracy with user trust. A complete

trust in the model would be reflected by a value of 1 across all three metrics. However, the results reveal that users did not fully trust the model, prompting a return to Phase 0 to improve the system's reliability. Thus, until trust is restored, this system is not suitable for real-world deployment.

Table 1: Trust results from (Miró-Nicolau et al. 2024).

Metric	User 1	User 2	Mean
Precision	0.1094	0.0156	0.0625
Recall	0.3333	0.0714	0.2022
F1-Score	0.1647	0.0256	0.0952

Despite the lower trust levels observed, these results demonstrate the effectiveness of our framework in identifying limitations within AI systems, ensuring that only trustworthy models proceed to real-world application.

4.5 Xai System Operation

It is essential to monitor health tools during public use to assess their long-term effects. Consequently, implementing a monitoring phase is critical to ensure that the XAI system can continuously provide services in compliance with the specified requirements. This need aligns with the seven requirements outlined in the ALTAI guidelines. Current legislation, specifically the EU AI Act, mandates that high-risk AI deployments, such as those related to health and well-being, undergo appropriate risk assessment and mitigation strategies. Furthermore, XAI systems in healthcare must be subject to review, approval, and ongoing monitoring by an institutional ethics committee. If necessary, corrections, modifications, and enhancements should be made to the deployed system.

5 CONCLUSIONS

In this paper, we present **XAIHealth**, a new evaluation framework specifically designed for XAI Systems for health and well-being. The framework is built by taking the general guidelines for a multidisciplinary approach to develop XAI systems, as detailed in (Mohseni, Zarei, and Ragan 2021).

The result is a structured evaluation framework that comprises two main phases: **Machine-Centred Analysis**, and **Human-Centred Assessment**. These phases are preceded by a **Pre-Evaluation** stage and conclude with the **XAI System Operation** phase. Iteration and feedback are integral throughout the

framework's phases, and legal and ethical considerations are addressed at every step of the evaluation process.

While our proposal primarily targets health and well-being applications, we believe that this framework could also be applicable to any XAI system that significantly influences human behaviour.

Future work will focus on two key areas. First, we will elaborate on the tasks necessary during the system operation phase, specifically concerning monitoring and improvement. Second, we aim to apply the framework to additional case studies to validate its effectiveness and identify potential enhancements.

ACKNOWLEDGEMENTS

This work is part of the Project PID2023-149079OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

REFERENCES

- Abdul, Ashraf, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanahalli. 2018. 'Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda'. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18. Montreal QC Canada: ACM. <https://doi.org/10.1145/3173574.3174156>.
- Adadi, Amina, and Mohammed Berrada. 2018. 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)'. *IEEE Access* 6:52138–60. <https://doi.org/10.1109/ACCESS.2018.2870052>.
- Adams, Barbara D, Lora E Bruyn, Sébastien Houde, and Kim Iwasa-Madge. 2003. 'TRUST IN AUTOMATED SYSTEMS LITERATURE REVIEW'.
- Agarwal, Chirag, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. 2022. 'Rethinking Stability for Attribution-Based Explanations'. <https://openreview.net/forum?id=BfxZAuWOg9>.
- Alcarazo, Lucía Ortiz de Zárate. 2022. 'Explicabilidad (de la inteligencia artificial)'. *EUNOMÍA. Revista en Cultura de la Legalidad*, no. 22 (March), 328–44. <https://doi.org/10.20318/eunomia.2022.6819>.
- Alonso, Jose M., Ciro Castiello, and Corrado Mencar. 2018. 'A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field'. *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, edited by Jesús Medina, Manuel Ojeda-Aciego, José Luis Verdegay,

- David A. Pelta, Inma P. Cabrera, Bernadette Bouchon-Meunier, and Ronald R. Yager, 3–15. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91473-2_1.
- Alvarez Melis, David, and Tommi Jaakkola. 2018. 'Towards Robust Interpretability with Self-Explaining Neural Networks'. *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.
- Anjomshoae, Sule, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. 'Explainable Agents and Robots: Results from a Systematic Literature Review'. *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 1078–88. AAMAS '19. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
- Arias-Duart, Anna, Ferran Parés, Dario Garcia-Gasulla, and Victor Giménez-Ábalos. 2022. 'Focus! Rating XAI Methods and Finding Biases'. *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. <https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882821>.
- Arras, Leila, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. 'Explaining Recurrent Neural Network Predictions in Sentiment Analysis'. *arXiv:1706.07206 [Cs, Stat]*, August. <http://arxiv.org/abs/1706.07206>.
- 'Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment | Shaping Europe's Digital Future'. 2020. 17 July 2020. <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>.
- Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. 'On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation'. *PLOS ONE* 10 (7): e0130140. <https://doi.org/10.1371/journal.pone.0130140>.
- Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrién Bannetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI'. *Information Fusion* 58 (June):82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>.
- Biran, Or, and Courtenay V. Cotton. 2017. 'Explanation and Justification in Machine Learning: A Survey Or'. <https://www.semanticscholar.org/paper/Explanation-and-Justification-in-Machine-Learning-%3A-Biran-Cotton/02e2e79a77d8aabc1af1900ac80ceebac20abde4>
- Chakraborti, Tathagata, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. 'Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy'. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 156–63. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2017/23>.
- Chalasanani, Prasad, Jiefeng Chen, Amrita Roy Chowdhury, Somesh Jha, and Xi Wu. 2020. 'Concise Explanations of Neural Networks Using Adversarial Training'.
- Cortez, Paulo, and Mark J. Embrechts. 2013. 'Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models'. *Information Sciences* 225 (March):1–17. <https://doi.org/10.1016/j.ins.2012.10.039>.
- Dasgupta, Sanjoy, Nave Frost, and Michal Moshkovitz. 2022. 'Framework for Evaluating Faithfulness of Local Explanations'. *Proceedings of the 39th International Conference on Machine Learning*, 4794–4815. PMLR. <https://proceedings.mlr.press/v162/dasgupta22a.html>.
- Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. 2018. 'Explainable Artificial Intelligence: A Survey'. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–15. <https://doi.org/10.23919/MIPRO.2018.8400040>.
- Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. 'Explaining Explanations: An Overview of Interpretability of Machine Learning'. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. <https://doi.org/10.1109/DSAA.2018.00018>.
- Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. 'Explaining and Harnessing Adversarial Examples'. *arXiv*. <https://doi.org/10.48550/arXiv.1412.6572>.
- Guidotti, Riccardo. 2021. 'Evaluating Local Explanation Methods on Ground Truth'. *Artificial Intelligence* 291 (February):103428. <https://doi.org/10.1016/j.artint.2020.103428>.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. 'Deep Residual Learning for Image Recognition'. 770–78. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.
- Hedström, Anna, Philine Lou Bommer, Kristoffer Knutsen Wickstrøm, Wojciech Samek, Sebastian Lapuschkin, and Marina MC Höhne. 2023. 'The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus'. *Transactions on Machine Learning Research*, February. <https://openreview.net/forum?id=j3FK00HyfU>.
- Hoffman, R., Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. 'Metrics for Explainable AI: Challenges and Prospects'. *arXiv Preprint arXiv:1812.04608*, December.
- Jian, Jiun-Yin, Ann M. Bisantz, and Colin G. Drury. 2000. 'Foundations for an Empirically Determined Scale of Trust in Automated Systems'. *International Journal of Cognitive Ergonomics* 4 (1): 53–71. https://doi.org/10.1207/S15327566IJCE0401_04.
- Justa, Segura Y. 2022. 'INTELIGENCIA ARTIFICIAL ÉTICA EN SANIDAD'. *DigitalIES*, 62.
- Lai, Vivian, and Chenhao Tan. 2019. 'On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception

- Detection'. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38. FAT* '19. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287590>.
- Lee, John, and Katrina See. 2004. 'Trust in Automation: Designing for Appropriate Reliance'. *Human Factors* 46 (February):50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>.
- Mamalakis, Antonios, Elizabeth A. Barnes, and Imme Ebert-Uphoff. 2022. 'Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience'. *Artificial Intelligence for the Earth Systems* 1 (4). <https://doi.org/10.1175/AIES-D-22-0012.1>.
- Mercado, Joseph E., Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. 'Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management'. *Human Factors* 58 (3): 401–15. <https://doi.org/10.1177/0018720815621206>.
- Miller, Tim. 2019. 'Explanation in Artificial Intelligence: Insights from the Social Sciences'. *Artificial Intelligence* 267 (February):1–38. <https://doi.org/10.1016/j.artint.2018.07.007>.
- Miró-Nicolau, Miquel, Antoni Jaume-i-Capó, and Gabriel Moyà-Alcover. 2024a. 'Assessing Fidelity in XAI Post-Hoc Techniques: A Comparative Study with Ground Truth Explanations Datasets'. *Artificial Intelligence* 335 (October):104179. <https://doi.org/10.1016/j.artint.2024.104179>. 2024b. 'Meta-Evaluating Stability Measures: MAX-Sensitivity and AVG-Sensitivity'. *Explainable Artificial Intelligence*, edited by Luca Longo, Sebastian Lapuschkin, and Christin Seifert, 356–69. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63787-2_18. 2025. 'A Comprehensive Study on Fidelity Metrics for XAI'. *Information Processing & Management* 62 (1): 103900. <https://doi.org/10.1016/j.ipm.2024.103900>.
- Miró-Nicolau, Miquel, Gabriel Moyà-Alcover, Antoni Jaume-i-Capó, Manuel González-Hidalgo, Maria Gemma Sempere Campello, and Juan Antonio Palmer Sancho. 2024. 'To Trust or Not to Trust: Towards a Novel Approach to Measure Trust for XAI Systems'. arXiv. <https://doi.org/10.48550/arXiv.2405.05766>.
- Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. 2021. 'A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems'. *ACM Transactions on Interactive Intelligent Systems* 11 (3–4): 1–45. <https://doi.org/10.1145/3387166>.
- Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. 2018. 'Methods for Interpreting and Understanding Deep Neural Networks'. *Digital Signal Processing* 73 (February):1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>.
- Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. 'Interpretable Machine Learning: Definitions, Methods, and Applications'. *Proceedings of the National Academy of Sciences* 116 (44): 22071–80. <https://doi.org/10.1073/pnas.1900654116>.
- 'Regulation (EU) 2024/1689 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) – European Sources Online'. 2024. 2024. <https://www.europeansources.info/record/proposal-for-a-regulation-laying-down-harmonised-rules-on-artificial-intelligence-artificial-intelligence-act-and-amending-certain-union-legislative-acts/>.
- Rieger, Laura, and Lars Kai Hansen. 2020. 'IROF: A Low Resource Evaluation Metric for Explanation Methods: Workshop AI for Affordable Healthcare at ICLR 2020'. *Proceedings of the Workshop AI for Affordable Healthcare at ICLR 2020*.
- Samek, Wojciech, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. 'Evaluating the Visualization of What a Deep Neural Network Has Learned'. *IEEE Transactions on Neural Networks and Learning Systems* 28 (11): 2660–73. <https://doi.org/10.1109/TNNLS.2016.2599820>.
- Scharowski, Nicolas, Sebastian A. C. Perrig, Nick von Felten, and Florian Brühlmann. 2022. 'Trust and Reliance in XAI -- Distinguishing Between Attitudinal and Behavioral Measures'. arXiv. <http://arxiv.org/abs/2203.12318>.
- Tjoa, Erico, and Cuntai Guan. 2021. 'A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI'. *IEEE Transactions on Neural Networks and Learning Systems* 32 (11): 4793–4813. <https://doi.org/10.1109/TNNLS.2020.3027314>.
- Tomsett, Richard, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. 2020. 'Sanity Checks for Saliency Metrics'. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (04): 6021–29. <https://doi.org/10.1609/aaai.v34i04.6064>.
- Union, Publications Office of the European. 2016. 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance)'. Website. Publications Office of the European Union. 27 April 2016. <http://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1>.
- VOSS, Axel. n.d. 'REPORT on Artificial Intelligence in a Digital Age | A9-0088/2022 | European Parliament'. Accessed 17 October 2022. https://www.europarl.europa.eu/doceo/document/A-9-2022-0088_EN.html.
- Xu, Wei. 2019. 'Toward Human-Centered AI: A Perspective from Human-Computer Interaction'. *Interactions* 26 (June):42–46. <https://doi.org/10.1145/3328485>.

Yeh, Chih-Kuan, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. 2019. 'On the (in)Fidelity and Sensitivity of Explanations'. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 10967–78. 984. Red Hook, NY, USA: Curran Associates Inc.

