Trustworthy AI in Medical Image Analysis: A Unified Perspective Built on Robustness and Layers of Trust*

Maria A. Zuluaga[a,b,*], Ivana Išgum[c,d,e,f] and Meritxell Bach Cuadra[g,h]

[a] EURECOM, Sophia Antipolis, France
[b] School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom
[c] Department of Radiology, Mayo Clinic, Rocherster, US
[d] Department of Biomedical Engineering and Physics, Amsterdam UMC, Amsterdam, The Netherlands
[e] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[f] Department of Radiology and Nuclear Medicine, Amsterdam UMC, Amsterdam, The Netherlands
[g] CIBM Center for Biomedical Imaging, Lausanne, Switzerland
[h] Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

ARTICLE INFO

Keywords:
Trustworthy AI in medical imaging
Technical robustness
Quality Control
Uncertainty Quantification
Explainability

ABSTRACT

Trustworthy AI is critical for effectively adopting AI systems in medical imaging and broader healthcare contexts. While the Trustworthy AI framework defines seven core principles —ranging from technical robustness to societal well-being— these are often addressed in isolation, lacking a coherent integration strategy. In this perspective paper, we propose a unified, layered framework that organizes these principles across three tiers of increasing trust: core operations, feedback, and explainability. Each layer aligns with the fundamental components of an AI system—input data, model, and outputs, integrating the different principles and offering a structured path toward increasing levels of trust. Central to our framework is technical robustness, positioned as a cross-cutting enabler that intertwines with the other trust principles across all layers. Through this lens, we review recent advances in trustworthy AI techniques in medical imaging and highlight persistent challenges. and future research directions for building trustworthy AI systems in medical imaging.

1. Introduction

Establishing trust is essential for the safe and effective use of AI in medical imaging. Trust arises when a system demonstrates reliability and predictability—even under unexpected conditions. *Trustworthy AI* provides a structured approach to ensuring that an AI system is worthy of being trusted based on evidence of its stated requirements. It ensures that user and stakeholder expectations are met in a verifiable way [39].

Since the introduction of the term [23], research on trustworthy AI for medical imaging applications has been highly active, evolving rapidly from a conceptual framework into a dynamic area of research and development. Grounded in the seven core principles of trustworthiness first outlined by the High-Level Expert Group on AI from the European Commission [23], this active development has led to significant technical advancements and focused research on essential aspects of AI development [37] and their application to medical imaging.

Within the seven principles, *technical robustness* refers to consistent performance and system resilience to unexpected and challenging conditions, minimizing failures that may have a negative impact. *Transparency* refers to the understandability and visibility of an AI system's inner workings, predictions, and limitations. It includes explainability, interpretability, and traceability to ensure that decisions can be audited. *Human oversight* supports human autonomy. In medical imaging applications, this involves clinical validation, model drift monitoring, and AI-human interactions, through approaches such as human-in-the-loop (HITL), human-on-the-loop (HOTL), and human-in-command (HIC) to maintain the user's control. *Diversity and fairness* mitigate bias and ensure inclusive performance. *Privacy and data governance* address robust data quality, integrity, and legitimate access controls. *Accountability and responsibility* assign liability throughout the AI lifecycle. *Societal well-being* refers to AI's positive societal impact.

These core principles of trustworthy AI are interconnected rather than isolated [23]. For instance, reliability is essential not just for technical robustness but also for ensuring effective human oversight. In the same way, robustness directly influences data governance, as maintaining high-quality data is fundamental to a system's resilience. Despite the close interplay among the core principles, most research in the field tends to be conducted in isolation [37], focusing exclusively on one principle at a time. There is a lack of unifying frameworks toward the shared goal of trustworthiness within an AI pipeline.

In this perspective paper, we propose a general unifying framework that organizes these principles into a layered architecture, where each tier represents an increasing level

* This work has been partially funded by TRAIN (ANR-22-FAI1-0003-02) and the Hasler Foundation Responsible AI program (MSxplain).
** We acknowledge access to the facilities and expertise of the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by CHUV, UNIL, EPFL, UNIGE and HUG.
*Corresponding author
maria.zuluaga@eurecom.fr (M.A. Zuluaga)
ORCID(s): 0000-0001-7511-2910 (M.A. Zuluaga); 0000-0003-1869-5034 (I. Išgum); 0000-0003-2730-4285 (M. Bach Cuadra)

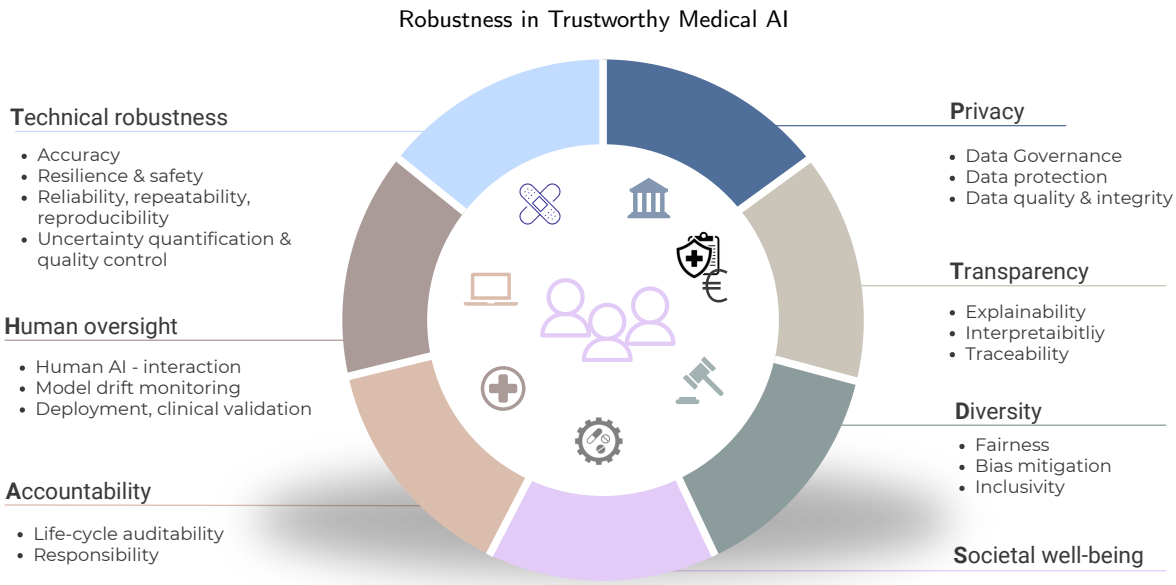

Figure 1: Overview of the seven core principles and characteristics of trustworthy AI [23] and the representation of involved stakeholders (patients, developers, healthcare professionals, healthcare industry, lawyers, insurers, and governments).

of trust. Each layer addresses a distinct subset of trustworthiness principles, illustrating how they collectively contribute to a system's trustworthiness. Central to our framework is the notion of robustness, which we position as a foundational property that permeates all layers. By embedding robustness throughout the trust hierarchy, we underscore its crucial role in meeting both technical and ethical requirements for AI systems in medical imaging. In the following sections, we introduce the components of the proposed unifying framework and illustrate their implementation with concrete examples from medical imaging applications.

2. A Unifying Perspective of Trustworthy AI in Medical Imaging

We propose a unified, layered view of Trustworthy AI that integrates its seven principles across the architecture of AI pipelines. Each layer corresponds to increasing levels of trust, grounded in the three core components of an AI system: input data, the model, and the output (Figure 2). Central to our framework is technical robustness, which underpins every layer as a foundational requirement for trust.

The *core* layer establishes the system's foundation by focusing on the quality and integrity of training data and initial model design. Here, trust is anchored in representative, diverse, and well-annotated datasets, which are critical for mitigating bias and enabling generalization across varying populations, imaging devices, and protocols. Robustness at this level is essential, as it ensures accuracy, reliability, and consistency. A robust *core* design intertwines with principles of *privacy and data governance* to ensure data integrity, representativeness, and compliance. It also encompasses *fairness* by enabling consistent performance across diverse populations through a data and model design that effectively addresses bias.

The *feedback* layer facilitates *human oversight* by providing information that allows user participation in the AI system's decision-making process through the HICL, HOCL, or HIC mechanisms. This includes tools and mechanisms for inspecting and monitoring inputs, models, and outputs, enabling users to provide feedback, correct errors, and refine a system's performance. The design of an AI system that operates in a symbiotic relationship with its users ensures safe operation, thereby enhancing its robustness. This layer also supports *accountability* by delivering information about the status, performance, or condition of the *core* layer during system operation.

The *explainable* layer addresses *transparency* by making the system's logic interpretable to all stakeholders. It builds on information from lower layers to provide clear explanations of decisions, uncertainties, and limitations. This interpretability broadens stakeholder engagement and feedback, enhances system validation, and fosters broader trust. Overall, a system that covers all layers is highly trustworthy and contributes to *societal well-being*.

In the following, we cover each tier, provide examples of recent trustworthiness research in medical imaging, and illustrate how the trustworthiness of an AI system's inputs, models, and outputs is addressed.

3. The *Core* Tier

The *core* tier establishes the foundation for a robust system by focusing on the interplay between input data, the AI model, and its outputs. Building trust through robustness begins with data quality. Hence, high-quality, diverse datasets that reflect relevant populations are essential (Sec 3.1). For the AI model, this tier addresses design and training to ensure it can effectively, fairly, and reliably manage data variability (Sec 3.2). In supervised learning

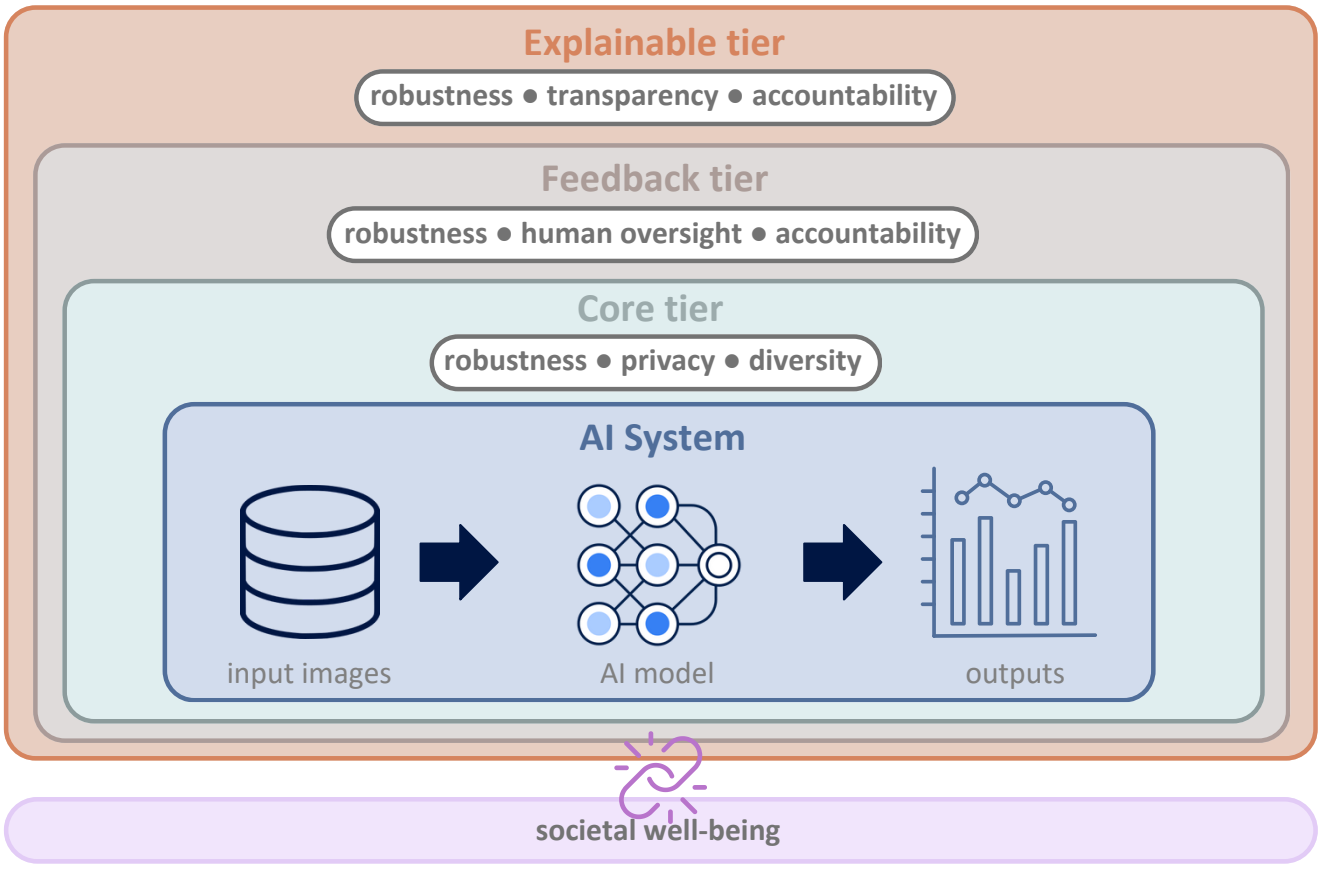

Figure 2: A unified view of the trustworthy AI framework and principles featuring three layers of trust: core operations, feedback, and explainability. A trustworthy AI system that covers all layers inherently contributes to the principle of societal well-being.

settings, this requires high-quality labels (Sec. 3.3). In learning scenarios without labels, such as unsupervised or self-supervised settings, the quality of the outputs is indirectly ensured by the input data. Similarly, in reinforcement learning, although there is no concept of an output, ensuring that the environment and agents are representative and of high quality is analogous to managing the input training data in a supervised learning framework.

3.1. Input Data: Ensuring Heterogeneity

An AI system's generalization and reliable performance rely on high-quality, diverse, and representative training data [57]. While gathering high-quality, diverse datasets seems straightforward, it is challenging and costly. Failing to obtain high-quality, diverse data leads to biased models, harming *robustness* and *fairness*. To circumvent this, there are techniques to automatically verify image quality and to boost training data heterogeneity.

Input image quality control (QC) verifies that the images fulfill the semantic content requirements of the target problem. This includes aspects such as the absence of artifacts or noise [49] or an adequate field of view [35], which are crucial, particularly when leveraging multi-centric data [55].

Domain randomization (DR) enhances data diversity by leveraging data augmentation and image synthesis techniques. Rather than applying manual transformations or optimizing generative models, DR expands the synthetic data distribution by randomizing the "generator" parameters. Initially proposed for robotics, DR has been applied to brain image segmentation, demonstrating excellent accuracy and robustness [4].

Advancements in generative modeling have significantly improved synthetic image quality, enabling high-quality, anatomically plausible datasets, including pathological variations [15]. However, these methods suffer from hallucinations [29], risking model reliability. Therefore, synthetically generated data, like real data, requires verification and curation before usage, for which image QC techniques can be employed.

3.2. Models: Handling Data Heterogeneity

Deployed AI systems may encounter data that has drifted from the training distribution (*domain shift*) [50]. Deep learning models struggle with significant domain shifts [29]. To mitigate the risk of failure, models must be designed to handle these gaps.

State-of-the-art approaches involve learning domain-invariant feature representations. Unlike image translation, these techniques build an intermediate representation [25] adaptable to downstream tasks [38] without domain translation. This approach is similar to foundation models (FMs), some recently developed for medical imaging [46]. Within

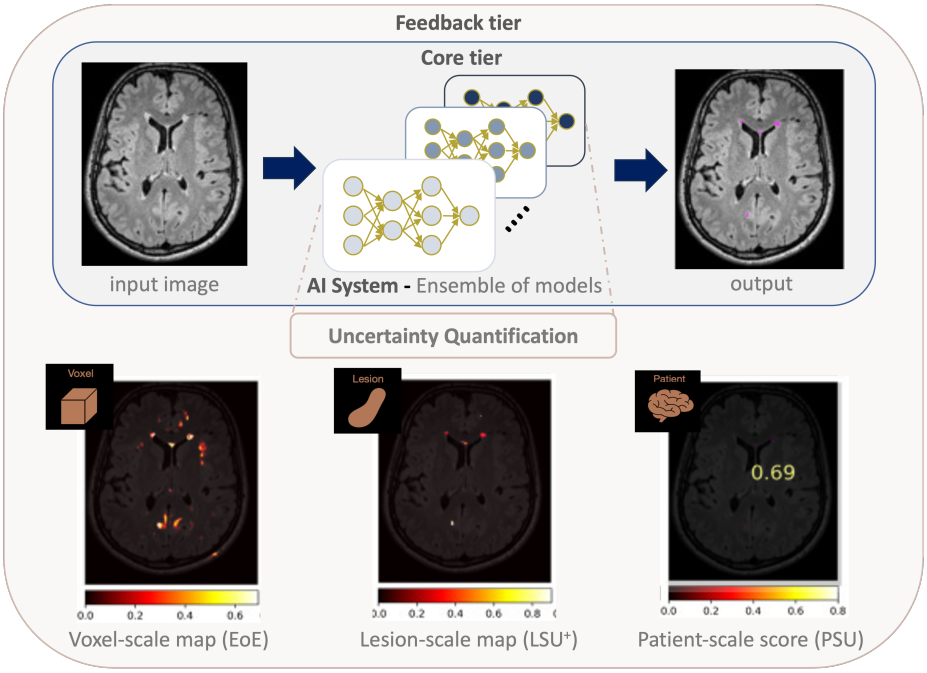

Figure 3: UQ in the *feedback* tier at different anatomical scales: example of Multiple Sclerosis lesion segmentation in magnetic resonance imaging using an ensemble of models (*core* tier). Adapted from [42].

a collaborative learning setup [19], these techniques enable robustness while preserving data *privacy*.

3.3. Outputs: High-Quality Labels

High-quality labels are paramount yet often overlooked [18]. Annotating experts are susceptible to errors due to fatigue or mistakes. They may also follow different annotation guidelines, introducing biases. These inaccuracies impact AI model reliability. Without high-quality labels, achieving *technical robustness* is challenging for clinical applications. If annotations are flawed, the model's performance will be compromised, leading to potentially incorrect diagnoses or treatment recommendations. Beyond training, labels are crucial for evaluation, serving as "ground truth." Inaccurate labels lead to misleading results, hinder progress, and affect the reproducibility and repeatability of findings.

Recent studies highlight the need for improved data curation and quality control to ensure the reliability of AI models. [52] demonstrates that annotation companies produce higher-quality labels than crowdsourcing. [54] analyzes public medical imaging datasets, revealing limitations in data quality and data governance.

4. The *Feedback* Tier

The *feedback* layer introduces human oversight and interaction, enabling bidirectional communication between AI systems and users. This tier provides continuous monitoring and evaluation mechanisms that enable feedback and contribute to *accountability*. The involvement of users in decision-making (*human oversight*) increases system robustness, as the final decision often rests with the expert user. Specifically, it addresses input data quality (Sec 4.1), model performance monitoring (Sec 4.2), and output quality control (Sec 4.3).

4.1. Input Data: Ensuring Quality

At the *feedback* layer, input image QC ensures data used by the AI system meets specifications for robust *operation*. Image QC verifies that data meet an AI system's operational specifications, including semantic content and the absence of missing data (e.g., in an imaging modality). It also assesses distribution shifts relative to the training data, a key step in maintaining reliability and safety. The latter is typically framed as an out-of-distribution (OOD) detection problem, benefiting from many statistical and machine-learning techniques [17].

In principle, any non-compliant image could be discarded or corrected. While discarding may be acceptable during training (*core* tier), correction is likely required once an AI system is deployed. Correction techniques include missing-modality imputation [59], image-quality transfer to bridge the gap between low-quality images at inference and higher-quality training sets [13], and data harmonization [40] and translation [24] to reduce data variability.

4.2. Models: Identifying Reliability with Uncertainty Quantification

Uncertainty quantification (UQ) assesses the reliability of model predictions and elucidates conditions under which a model may be incorrect. This information contributes to interpretability (by revealing the confidence level behind a model's predictions), supports model validation (by highlighting when and where a model's predictions may be

unreliable), and ultimately, improves technical robustness by enabling the system to recognize uncertain inputs or model decisions, thus facilitating model drift monitoring and informed decision-making. y Uncertainty arises from aleatoric uncertainty (data uncertainty), inherent in data due to noise or ambiguous labeling, and epistemic uncertainty (model uncertainty), from lack of knowledge when encountering unseen data or OOD data. Methods for UQ in deep learning can be classified by the type of uncertainty they address [31]. Some focus on aleatoric uncertainty through softmax calibration or test-time augmentations. Others address both uncertainties through Bayesian principles (e.g., Bayesian Neural Networks, Monte Carlo Dropout, or deep ensembles). More recent approaches include evidential deep learning and conformal prediction.

UQ operates at different levels (Figure 3), ranging from voxel- to organ, lesion, and subject-level (*human oversight*). While voxel-level UQ is common in segmentation, the resulting uncertainty maps can be overwhelming. Recent research focuses on higher-level estimates, aggregating voxel-level uncertainty or developing structural metrics, which better align with healthcare professionals' focus [42]. Subject-level UQ is valuable for patient analysis, as it incorporates input QC and helps identify poor predictions (Sec. 4.3).

Assessing the quality of UQ is crucial for clinical effectiveness. To that end, model calibration verifies the relationship between predicted probabilities and error rates [51], whereas error retention curves evaluate how uncertainty measures correlate with model performance [31].

4.3. Outputs: Quality Control

Ensuring robustness at the output stage involves mechanisms to evaluate prediction accuracy and identify errors. The task, denoted *quality control* (QC), has been done manually through visual inspection, followed by sample correction or removal. Frequently, identified poor-quality output samples may indicate model drift.

Because output QC is tedious, it has evolved into an automated process in which another system inspects outputs to assess their correctness, enabling the processing of large volumes of data. A straightforward approach to QC is UQ (Sec 4.2) [61], where UQ estimates are mapped to the output. This is commonly used in image classification [9], reconstruction, and synthesis [62]. In segmentation, QC is formulated as a classification or regression task. In classification, a model flags segmentations as good or bad. In regression, the QC model predicts a performance score, often the Dice score [3]. However, recent approaches have introduced QC-specific metrics based on performance variance [28, 48]. These assess variability among predictions, associating low variance with high quality, while circumventing the need for annotated data and correlating with the Dice score [28].

5. The *Explainable* Tier

The *explainable* tier enhances *transparency*, *human oversight*, and *accountability* by providing insights into an AI system's inner workings, enabling stakeholders to understand model decisions [6]. Explainability methods are typically classified according to the type of explanation (e.g., visual, example-based, textual, concept-based) [8] or to when they are applied (pre-modelling, in-modelling, post-modelling) [56]. Here, we analyze them through the elements of an AI system. Input-level methods assess feature influence (Sec 5.1); model-level methods examine layer influence (Sec 5.2); and output-level methods interpret decisions by analyzing predictions (Sec 5.3).

5.1. Input-Level Explanations

We consider input-level methods as those focusing on understanding how image characteristics influence the model output. Mainly, these include gradient-, perturbation-, and concept-based explanations.

Gradient-based methods, such as Saliency Maps or Class Activation Maps (CAMs), and their variants, such as Grad-CAMs, use the model's gradients to assess individual feature contributions to the output. Pixels with high positive/negative gradients drive/detract from the model's decisions. A strength of gradient-based methods is that they provide local explanations and may reveal unintended cues [14] by identifying pixels impacting the output, they are efficient, and can be applied to pre-trained models. Although these models show voxels affecting the output, they do not provide explanations and are sensitive to input changes [6].

Perturbation-based methods examine how the changes in the input impact the output. For example, occlusion-sensitivity methods occlude parts of the input to reveal their importance for the output, or modify parts to identify features that create biases [8]. Hence, like gradient-based, perturbation-based methods also verify the location of the abnormality, revealing unintended cues [2]. However, they are sensitive to the choice of masking, and their high computational cost hampers scalability [2]. Similarly, multi-instance learning-based explanations use bags of samples (e.g., image patches), gaining insights into the role played by the patches.

Finally, concept-based explanations quantify the influence of high-level (semantic) concepts, thereby connecting human and model reasoning. However, a model's concept may not fully align with a human concept or concept definition [45]. While concept definition is prone to inducing biases, automatic concept-based explanations mitigate it by semantically segmenting and grouping similar images [6].

5.2. Model-Level Explanations

We define model-level explainability methods as techniques that analyze how a model processes inputs across various network layers. The most common approaches are discussed below.

Feature visualization allows users to interpret neuron activations and reconstruct inputs, helping them understand what the model has learned. While they may also verify the location of the abnormality [10], they do not explain the reasons for a decision [8].

In layer-wise relevance propagation (LRP), backpropagation is used to compute relevance scores that show how

neurons and layers contribute to a prediction, allowing to generate fine-grained saliency maps that are well suited for small lesions [20]. However, while robust to small changes in a model's input, it is harder to interpret than heatmaps [8]. Moreover, LRPs may be unstable or inconsistent in their explanations [5]. In contrast, attention mechanisms highlight key areas of an input image through assigned attention weights. While the weights displayed in attention maps have been suggested to reveal the most relevant inputs at each layer, several studies showed that attention does not provide meaningful explanations [8]. These weights are displayed in attention maps, revealing the most important inputs at each layer [8].

5.3. Explainability at the Output Level

We define output-level explainability methods as techniques that interpret model decisions based on their predictions. Key methods include example-based approaches and textual explanations.

Example-based methods provide explanations by drawing on the most similar prior cases or counterfactual examples [16]. Similar example-based approaches identify the most similar prototypes to the given test case using image similarity (prototype-based methods) or distances in the semantically meaningful latent space (distance-based methods). The prototypes may be real patient images, image parts, or synthetic examples. These methods are intuitive but raise privacy and ethical concerns because they rely on learned prototypes. To preserve privacy, different approaches may be implemented, including the use of generative models to generate synthetic examples or latent information or scores, but not the specific example images [21, 43]. In addition, similar example-based methods are susceptible to spurious correlations and depend on the embedding quality [8].

In contrast to example-based methods, counterfactual explanation methods synthesize images that are highly similar to the test image yet yield a different prediction. Originally conceived for classification tasks, they have been extended to regression tasks [22]. Unlike example-based methods that use learned prototypes, counterfactual examples achieve explainability by modifying the test image and therefore do not affect patient privacy. By showing image locations that change the prediction, they enable the verification of the location of the abnormality [8]. However, the images they create may be unrealistic [63]. Lastly, textual explanations rely on natural language, in addition to using images. Specifically, image captioning methods in which specific words describe the image's visual content can provide insight into the rationale for model decisions [36]. However, the captions may not reflect features the model uses, and these methods lack localization of the explanation and are limited in explaining subtle findings [1].

6. Discussion and Perspectives

This work introduces a unified, three-layer perspective on Trustworthy AI to structure the seven core principles, as first proposed in [23], that warrant trust of an AI system. Importantly, the framework does not attempt to achieve trust itself. Trust is an inherently subjective feature that varies across stakeholders and contexts [53]. Instead, each layer translates high-level trustworthy-AI principles into concrete, actionable requirements: privacy, data governance, and bias-mitigation measures at the *core*; human oversight and monitoring in the *feedback* tier; and interpretable, transparent outputs in the *explainable* tier, all grounded in technical robustness. By organizing these elements into a coherent layered architecture, the framework provides a practical way to operationalize trustworthiness in medical imaging systems, while recognizing that the ultimate decision to grant trust remains beyond the scope of this work.

To illustrate the implementation of the seven core principles of AI within the proposed unifying layered framework, our work also presents concrete examples from medical imaging applications that showcase how the field has advanced in recent years. As such, our approach complements broader reviews [37, 39] and is compatible with other guidelines [32]. Indeed, by grounding our work on the seminal work on Trustworthy AI [23], we consider that our framework is sufficiently general to encompass the differing definitions of the core principles observed in the literature.

Finally, while the field has made remarkable strides in recent years, multiple challenges remain. Below, we identify and discuss three of these challenges, which point towards future research directions.

Ensuring Technical Robustness in Foundation Models. FMs offer a paradigm shift characterized by their massive scale, training on diverse datasets, and ability to be adapted to a wide range of downstream tasks. However, their "black box" nature, trained on often unknown datasets, lacks *transparency* regarding the data distribution. This hinders robustness, as understanding data characteristics is crucial for identifying biases.

While FMs may improve *core* performance [46], their opacity makes extracting *feedback* information or providing *explainable* explanations difficult. This limits evaluation and trust in healthcare applications. As FMs and large biomedical AI systems [47] become central, research should focus on developing novel *feedback* methods for monitoring black-box models with limited knowledge of input data statistics, an area that has received little attention so far [58].

Trustworthy for All. Explainable for Whom? While trustworthiness principles are objective, their perception varies. Developers prioritize performance and *feedback* information. Clinicians prefer feedback resembling real-world practice. Patients focus on understanding AI's contribution to diagnosis and *privacy*. Regulators prioritize *transparency* and *fairness* [7]. *Explainable* techniques have improved communication, but they remain too technical [30].

Language offers the flexibility to tailor explanations to diverse stakeholders. Large language models (LLMs) represent a promising new paradigm for achieving transparency in medical imaging applications. Preliminary work shows

LLMs can generate textual image descriptions [11], suggesting their capacity to bridge the gap between complex data and human understanding. A simpler yet related approach has been explored in medical imaging through concept bottlenecks, which provide textual feedback to sonographers during ultrasound image acquisition [34]. The challenge here lies in transforming *feedback* information into meaningful, audience-specific explanations.

Evaluation Frameworks. Evaluation is crucial in achieving the trustworthiness of any AI system. Evaluations should be repeatable and standardized. While medical imaging challenges promote standardized performance evaluation [12], similar standardization is needed for trustworthiness metrics, particularly at the *feedback* and *explainable* levels.

Trustworthy AI is a young field. As a result, the development of metrics, standardized benchmarks, and protocols for assessing its underlying principles is still ongoing, with some preliminary efforts reported in the literature. The MICCAI Quantification of Uncertainties in Biomedical Image Quantification Challenge [33] established a benchmark for algorithms that generate uncertainty estimates based on six medical image segmentation tasks. The Fairness Benchmark for Medical Imaging FMs [26] proposed a standardized evaluation process to assess the fairness of FMs across nine segmentation and eleven classification tasks. The open benchmarking framework for failure detection in medical image segmentation [61] evaluated failure-detection methodologies across five image segmentation tasks. Finally, the benchmark of trustworthiness in medical vision language models (VLMs) [60] assessed trustworthiness of VLMs across five dimensions: trustfulness, fairness, safety, privacy, and robustness, making it the most comprehensive trustworthiness benchmark to date.

Nonetheless, several principles remain to be covered. In particular, those that are somewhat intangible and therefore difficult to quantify, especially within the *explainable* tier. Explainability is a highly subjective concept. The "goodness" of an explanation may vary across different stakeholders. What constitutes a clear and understandable explanation for a clinician may differ substantially for a patient or a regulator. Although some works have begun to formalize criteria for assessing explainable AI techniques [27], these efforts are still in the preliminary stages. The lack of standardized evaluation frameworks may explain why the deployment of AI systems that implement the *explainable* tier remains limited, whereas *core*- and *feedback*-compliant AI systems are now being clinically validated [41] and commercialized [44].

Finally, perhaps the major challenge for standardized evaluation may arise from the need to integrate the different principles within a single AI system. In line with the evolution of research and development on trustworthy AI, principles are evaluated in isolation. However, it is well known that the orchestration of the principles entails trade-offs (e.g., improved fairness degrades performance) that cannot be quantified when they are assessed in isolation. Hence, it is crucial to advance in the definition of evaluation frameworks and protocols that account for these interactions and their implications, enabling stakeholders to identify such trade-offs and make decisions based on their priorities.

References

[1] D.-R. Beddiar, M. Oussalah, and T. Seppänen, "Automatic captioning for medical imaging (MIC): a rapid review of literature," *Artificial intelligence review*, vol. 56, no. 5, pp. 4019–4076, 2023.

[2] D. Bhati, F. Neha, and M. Amiruzzaman, "A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging," *Journal of Imaging*, vol. 10, no. 10, p. 239, 2024.

[3] ∗B.. Billot, C. Magdamo, Y. Cheng, S. E. Arnold, S. Das, and J. E. Iglesias, "Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets," *Proceedings of the National Academy of Sciences*, vol. 120, no. 9, Feb. 2023.

Exemplifies an AI system that implements the core and feedback layers effectively, demonstrating superior robustness while providing feedback to the user about the system's performance in brain parcellation.

[4] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias, "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, p. 102789, May 2023.

[5] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification," *Frontiers in aging neuroscience*, vol. 11, p. 194, 2019.

[6] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, "Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches," *European journal of radiology*, vol. 162, p. 110786, 2023.

[7] R. Bouderhem, "Shaping the future of ai in healthcare through ethics and governance," *Humanities and Social Sciences Communications*, vol. 11, no. 1, Mar. 2024.

[8] C. Brás, H. Montenegro, L. Y. Cai, V. Corbetta, Y. Huo, W. Silva, J. S. Cardoso, B. A. Landman, and I. Išgum, "Chapter 16 - explainable ai for medical image analysis," in *Trustworthy AI in Medical Imaging*, ser. The MICCAI Society book Series, M. Lorenzi and M. A. Zuluaga, Eds. Academic Press, 2025, pp. 347–366.

[9] A. Chen, Y. Li, W. Qian, K. Morse, C. Miao, and M. Huai, *Modeling and Understanding Uncertainty in Medical Image Classification*. Springer Nature Switzerland, 2024, pp. 557–567.

[10] B. D. De Vos, J. M. Wolterink, T. Leiner, P. A. De Jong, N. Lessmann, and I. Išgum, "Direct automatic coronary calcium scoring in cardiac and chest CT," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2127–2138, 2019.

[11] L. Dunlap, Y. Zhang, X. Wang, R. Zhong, T. Darrell, J. Steinhardt, J. E. Gonzalez, and S. Yeung-Levy, "Describing differences in image sets with natural language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 199–24 208.

[12] M. Eisenmann, A. Reinke, V. Weru, M. D. Tizabi, F. Isensee, T. J. Adler, S. Ali, V. Andrearczyk, M. Aubreville, U. Baid, S. Bakas, N. Balu, S. Bano, J. Bernal, S. Bodenstedt, A. Casella, V. Cheplygina, M. Daum, M. de Bruijne, A. Depeursinge, R. Dorent, J. Egger, D. G. Ellis, S. Engelhardt, M. Ganz, N. M. Ghatwary, G. Girard, P. Godau, A. Gupta, L. Hansen, K. Harada, M. P. Heinrich, N. Heller, A. Hering, A. Huaulmé, P. Jannin, A. E. Kavur, O. Kodym, M. Kozubek, J. Li, H. B. Li, J. Ma, C. Martín-Isla, B. H. Menze, J. A. Noble, V. Oreiller, N. Padoy, S. Pati, K. Payette, T. Rädsch, J. Rafael-Patino, V. S. Bawa, S. Speidel, C. H. Sudre, K. M. H. van Wijnen, M. Wagner, D. Wei, A. Yamlahi, M. H. Yap, C. Yuan, M. Zenk, A. Zia, D. Zimmerer, D. B. Aydogan, B. Bhattarai, L. Bloch, R. Brüngel, J. Cho, C. Choi, Q. Dou, I. Ezhov, C. M. Friedrich, C. Fuller, R. R. Gaire, A. Galdran, Á. García-Faura, M. Grammatikopoulou, S. Hong,

M. Jahanifar, I. Jang, A. Kadkhodamohammadi, I. Kang, F. Kofler, S. Kondo, H. J. Kuijf, M. Li, M. Luu, T. Martincic, P. Morais, M. A. Naser, B. Oliveira, D. Owen, S. Pang, J. Park, S. Park, S. Plotka, É. Puybareau, N. M. Rajpoot, K. Ryu, N. Saeed, A. Shephard, P. Shi, D. Stepec, R. Subedi, G. Tochon, H. R. Torres, H. Urien, J. L. Vilaça, K. A. Wahid, H. Wang, J. Wang, L. Wang, X. Wang, B. Wiestler, M. Wodzinski, F. Xia, J. Xie, Z. Xiong, S. Yang, Y. Yang, Z. Zhao, K. H. Maier-Hein, P. F. Jäger, A. Kopp-Schneider, and L. Maier-Hein, "Why is the winner the best?" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 19 955–19 966.

[13] A. K. Eldaly, M. Figini, and D. C. Alexander, "Alternative learning paradigms for image quality transfer," *Machine Learning for Biomedical Imaging*, vol. 2, pp. 2195–2222, 2024. [Online]. Available: https://melba-journal.org/2024:027

[14] M. Ennab and H. Mcheick, "Advancing AI interpretability in medical imaging: a comparative analysis of pixel-level interpretability and Grad-CAM models," *Machine Learning and Knowledge Extraction*, vol. 7, no. 1, p. 12, 2025.

[15] V. Fernandez, W. H. L. Pinaya, P. Borges, M. S. Graham, P.-D. Tudosiu, T. Vercauteren, and M. J. Cardoso, "Generating multi-pathological and multi-modal images and labels for brain MRI," *Medical Image Analysis*, vol. 97, p. 103278, Oct. 2024.

[16] M. Fontes, J. D. S. De Almeida, and A. Cunha, "Application of example-based explainable artificial intelligence (xai) for analysis and interpretation of medical imaging: a systematic review," *IEEE Access*, vol. 12, pp. 26 419–26 427, 2024.

[17] M. Fuchs, A. N. Angelopoulos, M. Paschali, C. Baumgartner, and A. Mukhopadhyay, "Navigating the unknown: out-of-distribution detection for medical imaging," in *Trustworthy AI in Medical Imaging*. Elsevier, 2025, pp. 73–99.

[18] M. Galanty, D. Luitse, S. H. Noteboom, P. Croon, A. P. Vlaar, T. Poell, C. I. Sanchez, T. Blanke, and I. Išgum, "Assessing the documentation of publicly available medical image and signal datasets and their impact on bias using the beamrad tool," *Scientific Reports*, vol. 14, no. 1, Dec. 2024.

[19] F. Galati, R. Cortese, F. Prados, M. Lorenzi, and M. A. Zuluaga, "Federated multi-centric image segmentation with uneven label distribution," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer Nature Switzerland, 2024, pp. 350–360.

[20] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller, and A. Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," *Scientific reports*, vol. 10, no. 1, p. 6423, 2020.

[21] K. A. Hasenstab, L. Hahn, N. Chao, and A. Hsiao, "Simulating clinical features on chest radiographs for medical image exploration and CNN explainability using a style-based generative adversarial autoencoder," *Scientific Reports*, vol. 14, p. 24427, 2024.

[22] L. S. Hesse, N. K. Dinsdale, and A. I. Namburete, "Prototype learning for explainable brain age prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7903–7913.

[23] High-level expert group on artificial intelligence, EU Comission, "Ethics guidelines for trustworthy AI," 2019. [Online]. Available: https://ec.europa.eu/digital-single

[24] ∗∗J.E.. Iglesias, B. Billot, Y. Balbastre, C. Magdamo, S. E. Arnold, S. Das, B. L. Edlow, D. C. Alexander, P. Golland, and B. Fischl, "Synthsr: A public ai tool to turn heterogeneous clinical brain scans into high-resolution t1-weighted images for 3d morphometry," *Science Advances*, vol. 9, no. 5, Feb. 2023.

> Featured in Stanford's 2024 AI Index Report, this work converts highly heterogeneous brain scans into high-resolution T1-weighted images, thus addressing image quality variability that limits the analysis of large volumes of data acquired within the clinical routine for advanced research.

[25] W. Ji and A. C. S. Chung, " Diffusion-based Domain Adaptation for Medical Image Segmentation using Stochastic Step Alignment ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15008. Springer Nature Switzerland, October 2024.

[26] R. Jin, Z. Xu, Y. Zhong, Q. Yao, Q. Dou, S. K. Zhou, and X. Li, "Fairmedfm: Fairness benchmarking for medical imaging foundation models," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024.

[27] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, "Guidelines and evaluation of clinical explainable ai in medical image analysis," *Medical Image Analysis*, vol. 84, p. 102684, Feb. 2023.

[28] J. Kalkhof and A. Mukhopadhyay, "M3d-nca: Robust 3d segmentation with built-in quality control," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 169–178.

[29] S. Kim, C. Jin, T. Diethe, M. Figini, H. F. Tregidgo, A. Mullokandov, P. Teare, and D. C. Alexander, "Tackling structural hallucination in image translation with local diffusion," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–103.

[30] N. K. Kollerup, S. S. Johansen, M. G. Tolsgaard, M. Lønborg Friis, M. B. Skov, and N. van Berkel, "Clinical needs and preferences for AI-based explanations in clinical simulation training," *Behaviour & Information Technology*, vol. 0, no. 0, pp. 1–21, 2024. [Online]. Available: https://doi.org/10.1080/0144929X.2024.2334852

[31] ∗B.. Lambert, F. Forbes, S. Doyle, H. Dehaene, and M. Dojat, "Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis," *Artificial Intelligence in Medicine*, vol. 150, p. 102830, Apr. 2024.

> State-of-the-art review on uncertainty quantification (UQ) techniques, proposing a taxonomy for UQ methods based on the resolution of the quantification (voxel-, organ-, lesion-, or image/subject-level), as presented in this work.

[32] ∗K.. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, C. Cintas, C. P. Langlotz, E. Weicken, F. W. Asselbergs, F. Prior, G. S. Collins, G. Kaissis, G. Tsakou, I. Buvat, J. Kalpathy-Cramer, J. Mongan, J. A. Schnabel, K. Kushibar, K. Riklund, K. Marias, L. M. Amugongo, L. A. Fromont, L. Maier-Hein, L. Cerdá-Alberich, L. Martí-Bonmatí, M. J. Cardoso, M. Bobowicz, M. Shabani, M. Tsiknakis, M. A. Zuluaga, M.-C. Fritzsche, M. Camacho, M. G. Linguraru, M. Wenzel, M. De Bruijne, M. G. Tolsgaard, M. Goisauf, M. Cano Abadía, N. Papanikolaou, N. Lazrak, O. Pujol, R. Osuala, S. Napel, S. Colantonio, S. Joshi, S. Klein, S. Aussó, W. A. Rogers, Z. Salahuddin, and M. P. A. Starmans, "Future-ai: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *BMJ*, p. e081554, Feb. 2025.

> A set of consensus guidelines by more than 100 international experts on implementing and deploying trustworthy AI systems for healthcare applications. It complements this work by providing specific advice on the methodology to build an AI system.

[33] H. B. Li, F. Navarro, I. Ezhov, A. Bayat, D. Das, F. Kofler, S. Shit, D. Waldmannstetter, J. C. Paetzold, X. Hu, *et al.*, "Qubiq: Uncertainty quantification for biomedical image segmentation challenge," *arXiv preprint arXiv:2405.18435*, 2024.

[34] ∗M.. Lin, A. Feragen, Z. Bashir, M. G. Tolsgaard, and A. N. Christensen, "I saw, i conceived, i concluded: Progressive concepts as bottlenecks," Nov. 2022.

> This work explores the usage of concept bottlenecks as a mechanism to provide feedback to sonographers when acquiring images transparently, i.e., using a language they understand. This work shows a promising direction

towards using natural language to improve explainability across all stakeholders..

[35] M. Lin, J. Ambsdorf, E. P. F. Sejer, Z. Bashir, C. K. Wong, P. Pegios, A. Raheli, M. B. S. Svendsen, M. Nielsen, M. G. Tolsgaard, A. N. Christensen, and A. Feragen, "Learning semantic image quality for fetal ultrasound from noisy ranking annotation," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–5.

[36] Y. Lin, K. Lai, and W. Chang, "Skin medical image captioning using multi-label classification and siamese network," *IEEE Access*, vol. 11, pp. 23 447–23 454, 2023.

[37] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang, "Trustworthy AI: A Computational Perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, Nov. 2022. [Online]. Available: https://doi.org/10.1145/3546872

[38] P. Liu, O. Puonti, X. Hu, D. C. Alexander, and J. E. Iglesias, "Brain-id: Learning contrast-agnostic anatomical representations for brain imaging," in *European Conference on Computer Vision*. Springer, 2024, pp. 322–340.

[39] M. Lorenzi and M. A. Zuluaga, Eds., *Trustworthy AI in Medical Imaging*, 1st ed., ser. The MICCAI Society Book Series. Chantilly: Elsevier Science & Technology, 2024.

[40] C. Marzi, M. Giannelli, A. Barucci, C. Tessa, M. Mascalchi, and S. Diciotti, "Efficacy of MRI data harmonization in the age of machine learning: a multicenter study across 36 datasets," *Scientific Data*, vol. 11, no. 1, Jan. 2024.

[41] J. Merkow, F. J. Dorfner, X. Yang, A. Ersoy, G. Dasegowda, M. Kalra, M. P. Lungren, C. P. Bridge, and I. Tarapov, "Scalable drift monitoring in medical imaging ai," *arXiv preprint arXiv:2410.13174*, 2024.

[42] N. Molchanova, V. Raina, A. Malinin, F. L. Rosa, A. Depeursinge, M. Gales, C. Granziera, H. Müller, M. Graziani, and M. B. Cuadra, "Structural-based uncertainty in deep learning across anatomical scales: Analysis in white matter lesion segmentation," *Computers in Biology and Medicine*, vol. 184, p. 109336, Jan. 2025.

[43] H. Montenegro, W. Silva, and J. S. Cardoso, "Privacy-preserving generative adversarial network for case-based explainability in medical image analysis," *IEEE Access*, vol. 9, pp. 148 037–148 047, 2021.

[44] A. Y. Ng, C. J. Oberije, É. Ambrózay, E. Szabó, O. Serfőző, E. Karpati, G. Fox, B. Glocker, E. A. Morris, G. Forrai, *et al.*, "Prospective implementation of ai-assisted screen reading to improve early detection of breast cancer," *Nature Medicine*, vol. 29, no. 12, pp. 3044–3049, 2023.

[45] A. Nicolson, L. Schut, J. A. Noble, and Y. Gal, "Explaining explainability: Recommendations for effective use of concept activation vectors," *Transactions on Machine Learning Research*, vol. 2025, 2025.

[46] M. Paschali, Z. Chen, L. Blankemeier, M. Varma, A. Youssef, C. Bluethgen, C. Langlotz, S. Gatidis, and A. Chaudhari, "Foundation models in radiology: What, how, why, and why not," *Radiology*, vol. 314, no. 2, Feb. 2025.

[47] F. Pérez-García, H. Sharma, S. Bond-Taylor, K. Bouzid, V. Salvatelli, M. Ilse, S. Bannur, D. C. Castro, A. Schwaighofer, M. P. Lungren, M. T. Wetscherek, N. Codella, S. L. Hyland, J. Alvarez-Valle, and O. Oktay, "Exploring scalable medical image encoders beyond text supervision," *Nature Machine Intelligence*, vol. 7, no. 1, pp. 119–130, Jan. 2025.

[48] A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. M. Patel, "Ambiguous medical image segmentation using diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 536–11 546.

[49] D. Ravi, F. Barkhof, D. C. Alexander, L. Puglisi, G. J. Parker, and A. Eshaghi, "An efficient semi-supervised quality control system trained using physics-based MRI-artefact generators and adversarial training," *Medical Image Analysis*, vol. 91, p. 103033, Jan. 2024.

[50] J. Richiardi, V. Ravano, N. Molchanova, P. M. Gordaliza, T. Kober, and M. B. Cuadra, "Domain shift, domain adaptation, and generalization: A focus on MRI," in *Trustworthy AI in Medical Imaging*. Elsevier, 2025, pp. 127–151.

[51] M. Roschewitz, G. Khara, J. Yearsley, N. Sharma, J. J. James, É. Ambrózay, A. Heroux, P. Kecskemethy, T. Rijken, and B. Glocker, "Automatic correction of performance drift under acquisition shift in medical image classification," *Nature Communications*, vol. 14, no. 1, p. 6608, Oct. 2023.

[52] T. Rädsch, A. Reinke, V. Weru, M. D. Tizabi, N. Heller, F. Isensee, A. Kopp-Schneider, and L. Maier-Hein, *Quality Assured: Rethinking Annotation Strategies in Imaging AI*. Springer Nature Switzerland, Oct. 2024, pp. 52–69.

[53] M. Sagona, T. Dai, M. Macis, and M. Darden, "Trust in ai-assisted health systems and ai's trust in humans," *npj Health Systems*, vol. 2, no. 1, p. 10, 2025.

[54] A. J. Sánchez, N.-R. Avlona, D. Juodelyte, T. Sourget, C. Vang-Larsen, A. Rogers, H. D. Zajac, and V. Cheplygina, "Copycats: the many lives of a publicly available medical imaging dataset," in *Advances in Neural Information Processing Systems 38 (NeurIPS 2024): Datasets and Benchmarks Track*, 2024.

[55] T. Sanchez, O. Esteban, Y. Gomez, A. Pron, M. Koob, V. Dunet, N. Girard, A. Jakab, E. Eixarch, G. Auzias, and M. Bach Cuadra, "FetMRQC: A robust quality control system for multi-centric fetal brain MRI," *Medical Image Analysis*, vol. 97, p. 103282, 2024.

[56] A. Saranya and R. Subhashini, "A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends," *Decision analytics journal*, vol. 7, p. 100230, 2023.

[57] D. Schwabe, K. Becker, M. Seyferth, A. Klaß, and T. Schaeffter, "The metric-framework for assessing data quality for trustworthy ai in medicine: a systematic review," *npj Digital Medicine*, vol. 7, no. 1, Aug. 2024.

[58] M. Sun, W. Yan, P. Abbeel, and I. Mordatch, "Quantifying uncertainty in foundation models via ensembles," in *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*, 2022.

[59] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 878–15 887.

[60] P. Xia, Z. Chen, J. Tian, Y. Gong, R. Hou, Y. Xu, Z. Wu, Z. Fan, Y. Zhou, K. Zhu, *et al.*, "Cares: A comprehensive benchmark of trustworthiness in medical vision language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 140 334–140 365, 2024.

This work presents the most complete benchmark to assess trustworthiness reported in the literature so far. It has a narrow focus on medical vision language models, but could serve as a reference starting point to develop standardized evaluation frameworks for trustworthiness

[61] M. Zenk, D. Zimmerer, F. Isensee, J. Traub, T. Norajitra, P. F. Jäger, and K. Maier-Hein, "Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation," *Medical Image Analysis*, vol. 101, p. 103392, Apr. 2025.

[62] J. Zhang, S. Bi, and V. Fung, "On the quantification of image reconstruction uncertainty without training data," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2072–2081.

[63] Y. Zhu, L. Zhang, C. Sainsbury, F. Dong, J. MacLay, D. J. Lowe, and X. Ye, "Counterfactual medical images generation for lung disease diagnosis using probabilistic causal models and active learning," *IEEE Access*, 2025.