

Review

Trustworthy AI Guidelines in Biomedical Decision-Making Applications: A Scoping Review

Marçal Mora-Cantallops ^{*,†}, Elena García-Barriocanal [†] and Miguel-Ángel Sicilia [†]

Computer Science Department, Universidad de Alcalá, 28801 Madrid, Spain; elena.garcia@uah.es (E.G.-B.); msicilia@uah.es (M.-Á.S.)

* Correspondence: marcal.mora@uah.es

† These authors contributed equally to this work.

Abstract: Recently proposed legal frameworks for Artificial Intelligence (AI) depart from some frameworks of concepts regarding ethical and trustworthy AI that provide the technical grounding for safety and risk. This is especially important in high-risk applications, such as those involved in decision-making support systems in the biomedical domain. Frameworks for trustworthy AI span diverse requirements, including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, fairness, and societal and environmental impact. Researchers and practitioners who aim to transition experimental AI models and software to the market as medical devices or to use them in actual medical practice face the challenge of deploying processes, best practices, and controls that are conducive to complying with trustworthy AI requirements. While checklists and general guidelines have been proposed for that aim, a gap exists between the frameworks and the actual practices. This paper reports the first scoping review on the topic that is specific to decision-making systems in the biomedical domain and attempts to consolidate existing practices as they appear in the academic literature on the subject.

Keywords: artificial intelligence (AI); trustworthy AI; AI regulation



Citation: Mora-Cantallops, M.; García-Barriocanal, E.; Sicilia, M.-Á. Trustworthy AI Guidelines in Biomedical Decision-Making Applications: A Scoping Review. *Big Data Cogn. Comput.* **2024**, *8*, 73. <https://doi.org/10.3390/bdcc8070073>

Received: 2 June 2024

Revised: 24 June 2024

Accepted: 26 June 2024

Published: 1 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The swift advancement of Artificial Intelligence (AI) is delivering considerable economic and social advantages across various sectors; its application to healthcare holds significant transformative potential, from enhancing drug discovery to improving the delivery of healthcare services. The scope of AI applications in healthcare is vast, including medical image analysis, automated genetic data analysis, disease prediction, medical robotics, telemedicine, syndromic surveillance, and virtual consultations. In summary, Artificial Intelligence in medicine has the potential to be a key catalyst in enhancing diagnostics and treatments and boosting overall efficiency to address the (ever-growing) demand for healthcare services.

Despite the promise of AI in healthcare, significant challenges and risks persist [1]. The AI hype often overshadows actual scientific progress, particularly in validation and readiness for clinical implementation. Flawed algorithms, underscoring the need for rigorous debugging, validation, and regulatory scrutiny, have all the potential for harm and medical malpractice. Moreover, AI systems can exacerbate existing healthcare inequities and possess inherent biases due to unrepresentative datasets. The inherent opaqueness of deep learning models and their “black box” nature raise further concerns about transparency and explainability, essential for patient care compliance with regulations like the EU’s General Data Protection Regulation (GDPR). The privacy and security of data are also critical concerns, with risks of hacking and data breaches posing significant barriers to the adoption of AI in healthcare. Another critical concern remains whether AI will foster greater inclusivity and fairness in healthcare or exacerbate existing disparities and

inequities. These challenges necessitate new models of health data ownership and robust security measures to protect patient information and ensure the responsible use of AI in medicine, with tailored regulatory frameworks to navigate the socioethical dimensions of AI applications.

Therefore, as AI becomes more integrated in these areas, it is increasingly crucial that these systems are deemed trustworthy. This is because any violation of trust in such pervasive systems can have profound societal repercussions and impacts. In this regard, in 2019, the European Commission High-Level Expert Group (HLEG) on AI presented the report titled “Ethics Guidelines for Trustworthy Artificial Intelligence” [2]. European Union lawmakers also reached a political agreement on the draft of the Artificial Intelligence (AI) Act in December 2023. The drafted AI act, proposed by the European Commission in April 2021, sets a common framework for the use and supply of AI systems in the EU on a “risk-based approach”. With it becoming legislation, a wide range of high-risk AI systems that can have a detrimental impact on people’s health, safety, or fundamental rights will be authorized but subject to a set of requirements and obligations to gain access to the EU market. These efforts are not exclusive to Europe though; other geographies are also following the same trend, such as the U.S. Food and Drug Administration (FDA) approving the integration of AI technologies into the medical sector when they consider ethical issues such as privacy, transparency, safety, and accountability [3].

However, while numerous principles and standards aimed at guiding the development of ethical and reliable AI systems exist, these often struggle to make the leap from theory to real-world application, as Goirand et al. [4] concluded in a previous review. Many of the guidelines for trustworthy AI are high-level and abstract, lacking specific, actionable steps that developers and organizations can follow. Furthermore, implementing the principles of trustworthy AI—such as transparency, fairness, and accountability—into complex AI systems presents significant technical challenges, from resource-intensive needs to increased complexity. There may also be a gap in understanding or expertise among AI developers and engineers regarding ethical implications and the means to implement trustworthy AI principles, while effective implementation of trustworthy AI guidelines requires the involvement of diverse stakeholders (such as end-users, domain experts, and regulators) who are not always included in the development loop.

In order to assess how researchers and practitioners are translating the HLEG trustworthy AI requirements into practice, we conduct this scoping review on decision-making systems in the biomedical domain. The aim is to identify and categorize the range of existing approaches and technologies, highlighting variations and commonalities in how decision-making systems are implemented across different biomedical contexts while highlighting those areas lacking sufficient research or development to guide future studies and technological advancements to address these deficiencies.

Thus, the questions posed by this scoping review are as follows:

1. What are the existing practices that implement trustworthy AI principles in medical research?
2. What are the principles from the ethics guidelines that are addressed by these practices?
3. What are the common approaches found in the existing literature?
4. What gaps prevent the coverage of all principles and, therefore, system trustworthiness?

The rest of this paper is structured as follows: Section 2 describes the methodology, criteria, and classification work that were followed. Afterward, Section 3 summarizes our findings according to the review conducted. Finally, Section 4 provides answers to the posed questions and highlights the strengths and weaknesses found in the existing translations of the trustworthy AI principles to medical practice and decision systems.

2. Materials and Methods

2.1. Data Sources and Systematic Searches

This study used the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Extension for Scoping Reviews (PRISMA-ScR) statement [5]. This recent approach

is appropriate for scoping reviews due to its emphasis on systematic, transparent, and comprehensive reporting, its flexibility to accommodate a broad range of evidence, and its utility in identifying research gaps and informing future research agendas, and it is widely used in the field.

After some pilot tests and reviewing the terms used in relevant publications, the exact search syntax was customized for the PUBMED, MEDLINE, and Web of Science databases, which are considered in this study due to their topic. For instance, due to the pilot tests, the term “reliable” was discarded from the search as, even though some studies might use it as a substitute term for trustworthy AI, and it is also a very common term in medical practices. Thus, the final search syntax included search terms related to the target domains (“medical” or “biomedical”), the focus of the review (“trustworthy”, which also implies a certain degree of purposefulness), and the related techniques (“language model”, “machine learning”, “deep learning”, “Artificial Intelligence” or “AI”). Searches were also limited to studies published in the English language. The final query could be reproduced as follows, searching both in the title and in the abstract fields:

(AI OR “Artificial Intelligence” OR “deep learning” OR “machine learning” OR “language model”) AND (medical OR biomedical) AND (trustworthy)

Since the ethics guidelines for trustworthy AI were presented in 2019, that year was chosen as the starting point of the review. The cut-off date was set to 10 April 2024 (the date of the final queries). Review articles, editorials, and abstract-only types were excluded. Finally, all three researchers manually examined the contents of all remaining articles to identify those that contained or described actual practices involving trustworthy and responsible AI in the medical domain.

Analyzed articles were characterized by the following criteria: number of citations, authors, title, year of publication, source titles, and trustworthy AI requirements covered. The first author’s affiliation defined the country of origin if there was more than one affiliation.

Finally, it is important to note that, due to its properties, this study did not need to obtain the approval of an ethics committee or institutional review board.

2.2. Inclusion and Exclusion Criteria

The following criteria were considered for this study and the inclusion/exclusion of the obtained papers:

- Papers had to be written in English.
- Papers had to be published in a journal or a conference proceeding.
- Papers had to be explicit and intentional about their trustworthiness objectives.
- Papers had to describe, totally or partially, a practical case where one or multiple trustworthy AI principles were covered or answered.
- Papers had to be published between 1 January 2019 and 10 April 2024.

On the other hand, studies outside the scope of the study would be as follows:

- Papers written in languages other than English.
- Papers that are descriptive or prescriptive in nature or propose guidelines or theoretical frameworks without associated use cases or ways to implement them in practice. Thus, review articles were also excluded.

2.3. Identification and Selection of Studies

After the initial search, a total of 263 results were found, with 113 from PUBMED and 150 from Web of Science. Ninety-five duplicate records were removed before screening. The remaining 168 documents were screened by reading the abstracts, and 86 were excluded based on predetermined criteria. Subsequently, 82 studies were identified for complete retrieval. Two of these could not be retrieved, resulting in a final set of 80 studies for analysis. They were examined to confirm their suitability for inclusion in the review. Articles were fully accessed and read, their general information (authors, countries, citations, and year of publication) tabulated, and their practices, if any, linked to the corresponding requirement.

Furthermore, they were summarized together with a list of the techniques and tools they described or used. Following a detailed analysis and consensus among the researchers, 30 additional exclusions were made, leaving a final corpus of 50 documents for further examination (see Figure 1). This was complemented with a narrative synthesis, grouping the identified practices by requirement covered. The results are synthesized in the following section, with the identified strengths and gaps in the discussion further below.

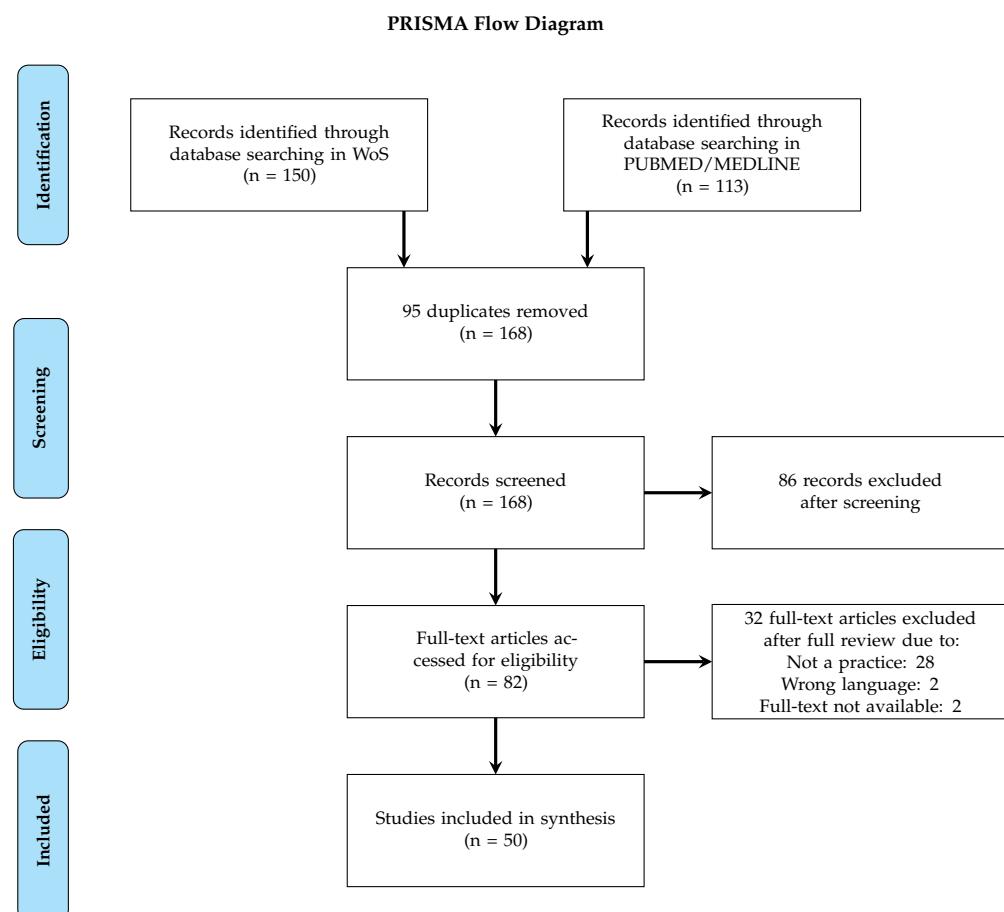


Figure 1. PRISMA flow diagram of the article selection.

2.4. Limitations

We employed an extensive search strategy adhering to a strict process for screening and evaluating articles. However, initiatives from the private sector, which are often not available through public literature searches, may be underrepresented in this review. The transversal nature of the topic resulted in a heterogeneous set of study data that did not always fit neatly into the systematic scoping review process, making screening and eligibility assessments challenging and data collection and evaluation complex. Utilizing three reviewers and agreement mechanisms helped mitigate potential biases and gaps in data collection. Additionally, due to the language limitations of English only, countries like China, Korea, and Japan, which are significant contributors to AI development in healthcare, might be under-represented in our review.

3. Results

3.1. Overview

In recent years, the interest in trustworthy Artificial Intelligence (AI) within the medical and biomedical sectors has seen a significant surge (see Figure 2). Note that the data for 2024 only include the first 100 days of the year, so they show a similar trend as the previous years once extrapolated. This heightened attention is driven by the expanding integration

of AI technologies in critical areas such as diagnostics, treatment planning, and patient management [3,6,7]. As AI systems increasingly influence decisions that directly impact patient outcomes, the imperative for these systems to be reliable, safe, and ethical has become more pronounced. This growing focus is also a response to concerns over privacy, bias, and transparency, which are particularly sensitive in healthcare due to the potential for direct effects on patient well-being. Consequently, the medical research community is betting heavily on frameworks and guidelines that ensure AI systems are not only effective but also align with the requirements for medical ethics and patient safety [4,8].

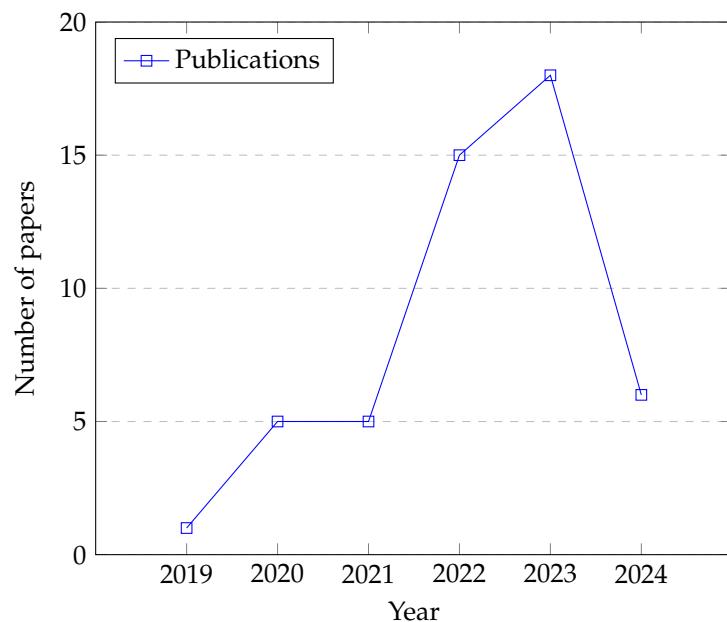


Figure 2. Number of publications per year

However, these concerns and theoretical calls to action do not reflect a proportional volume of trustworthy AI practices. Our search and posterior filters resulted in an arguably small set of 50 papers that purposely engage with trustworthiness in their research or model design. Furthermore, most works are European ($n = 20$, representing 40% of the included works), which might be related to the efforts conducted by the European Commission in this sense. North America (USA and Canada) is represented by 13 works (26%), while Asia (with India on top) follows closely at 12 (24%). Saudi Arabia and Australia contribute two works (4%) each, and a single study comes from an African institution (Egypt, 2%).

Regarding the medical scopes or specialties involved, the great majority of practices can be classified into two broad categories: 13 articles (26%) are labeled as “bioinformatics”, considering their general approach to any healthcare data science flow (e.g., concerning general model security against attacks or privacy-preserving architectures), while 17 (34%) belong to the “medical imaging” models (e.g., image classifiers), which seem to be the most widely adopted AI approach in healthcare. The rest (20 works in total, 40%) are almost evenly divided among 14 different medical specialties, such as epidemiology, oncology, or cardiology, as well as ICU and emergency management.

The results for each requirement (as described by the HLEG) (Table 1) are detailed in the following subsections and constitute the core of the analysis. However, the sixth requirement, which focuses on societal and environmental well-being, is not included in the analysis. This requirement stresses the importance of AI technologies making positive contributions to society, especially in healthcare. It is assumed that all the included publications adhere to this requirement, as their models, frameworks, and practices are intended to improve patient outcomes, make healthcare more accessible, and reduce inequalities. These efforts aim to benefit all segments of society and advance individual care

and public health goals. Furthermore, while we argue that all included publications address societal well-being, none were found that address environmental impact in any form.

Table 1. The seven requirements for trustworthy AI.

Requirement	Description
Human Agency and Oversight	AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights, while maintaining appropriate human oversight and intervention capabilities.
Technical Robustness and Safety	AI systems should be resilient and secure, with reliable performance and fallback plans in case of issues. This includes managing risks related to cybersecurity and ensuring accuracy, reliability, and reproducibility.
Privacy and Data Governance	AI systems must ensure that personal data are protected, incorporating mechanisms for data protection, consent, and privacy preservation. Proper data governance practices must be in place to manage data quality and integrity.
Transparency	AI systems should operate transparently, with traceable processes and understandable outcomes. This involves clear communication about the system's capabilities, limitations, and decision-making processes.
Diversity, Nondiscrimination, and Fairness	AI systems should promote inclusion and diversity, preventing bias and ensuring that they do not discriminate against any individuals or groups. They must be designed and tested to ensure fair outcomes.
Environmental and Societal Well-Being	AI systems should consider their environmental impact, promoting sustainability and ensuring that they contribute positively to societal well-being and the common good.
Accountability	There must be mechanisms to ensure accountability for AI systems and their outcomes, with clear guidelines for responsibility and mechanisms for action in cases of harm or negative impacts.

The following subsections are a breakdown of the results summarized in Table 2 by requirement.

Table 2. Summary of included studies in regard to the related requirement. HUM. (Human Agency), TEC. (Technical Robustness and Safety), PRI. (Privacy and Data Governance), TRA. (Transparency), DIV. (Diversity, Nondiscrimination, and Fairness), and ACC. (Accountability).

Publication	HUM.	TEC.	PRI.	TRA.	DIV.	ACC.
Alves et al. [9]	N	Y	Y	N	N	N
Moradi and Samwald [10]	N	Y	N	N	N	N
Alves et al. [9]	N	Y	N	Y	N	Y
Moradi and Samwald [10]	Y	Y	N	Y	N	Y
Ma et al. [11]	Y	Y	N	Y	Y	N

Table 2. Cont.

Publication	HUM.	TEC.	PRI.	TRA.	DIV.	ACC.
Khanna et al. [12]	Y	Y	N	Y	N	N
Fidon et al. [13]	N	Y	N	N	Y	Y
Nambiar et al. [14]	N	N	N	Y	N	N
Rashid et al. [15]	Y	Y	Y	Y	Y	N
Kumar et al. [16]	N	N	Y	Y	N	N
Salahuddin et al. [17]	N	N	N	Y	N	N
Zicari et al. [18]	Y	Y	Y	Y	Y	Y
Bruckert et al. [19]	Y	N	N	Y	N	Y
Ma et al. [20]	Y	N	N	Y	N	N
Imboden et al. [21]	N	Y	N	N	N	N
Karim et al. [22]	N	Y	N	N	N	N
Mu et al. [23]	N	Y	Y	Y	N	N
Kamal et al. [24]	Y	N	N	Y	N	Y
Hassan et al. [25]	N	N	N	Y	N	N
Tasnim et al. [26]	N	N	N	Y	N	N
El-Rashidy et al. [27]	N	N	N	N	N	N
Prifti et al. [28]	N	N	N	Y	N	N
Miao et al. [29]	N	N	N	N	N	N
Kumar et al. [30]	Y	N	N	Y	Y	N
Vijayvargiya et al. [31]	N	N	N	N	N	N
Pintelas et al. [32]	N	N	N	Y	N	N
Wang et al. [33]	N	N	N	Y	N	N
Lugan et al. [34]	N	Y	Y	N	N	N
Shukla et al. [35]	N	Y	N	N	N	N
Bassiouny et al. [36]	Y	N	N	Y	N	N
Jiang et al. [37]	N	Y	Y	N	N	N
Abdelfattah et al. [38]	N	N	Y	N	N	N
De Paolis Kaluza et al. [39]	N	N	N	N	Y	N
Aboutalebi et al. [40]	Y	N	N	Y	Y	N
Uzunova et al. [41]	N	N	N	Y	N	N
Lu et al. [42]	Y	Y	N	Y	Y	N
Chen et al. [43]	N	N	N	Y	N	Y
Araujo et al. [44]	Y	N	N	Y	N	N
Malik et al. [45]	N	N	Y	N	N	N
Zerka et al. [46]	N	N	Y	N	N	N
Saleem et al. [47]	N	N	N	Y	N	N
El Houda et al. [48]	N	Y	Y	N	N	N
Stenwig et al. [49]	N	N	N	Y	N	N
Ogbomo-Harmitt et al. [50]	N	N	N	Y	N	N
Alzubaidi et al. [51]	N	Y	N	Y	N	N
El-Sappagh et al. [52]	N	N	N	Y	N	N
Gundersen and Baeroe [53]	N	N	N	N	Y	N
Mukhopadhyay [54]	N	Y	N	Y	Y	N
Alamro et al. [55]	N	Y	N	N	N	N
Soni et al. [56]	N	Y	N	N	N	N

3.2. Human Agency and Oversight

The principle of human agency and oversight in trustworthy AI emphasizes the necessity of maintaining human control over AI systems. This ensures that AI actions are reversible and that systems are designed and operated in a manner that respects user autonomy and allows meaningful human intervention and oversight when needed. A total of 13 studies (26%) were categorized as including practices that cover requirements from this category.

However, it is worth noting that all 13 studies focus on explainability, which implies that their ethical considerations on human agency and oversight are a byproduct of their interpretability efforts. For instance, Cho et al. [57] includes a section devoted to assessing their model explainability (an explicit inclusion of trustworthiness efforts that is missing from most other works) and claim in their discussion that they “employed a global interpretable ML models [sic] to construct a decision support system, especially for making critical medical decisions” and how it “can help healthcare professionals understand which factors contribute most significantly to the predicted outcomes”, stressing this idea of a second opinion system [44], while others directly acknowledge the need of expert oversight for critical decisions, as they leave “the ultimate control of diagnosis to the clinician” for human-in-command AI [36]. This mention of aided decision making is significantly subtle in most works ([11,12,15,20,24,30]), which gives this principle a footnote that is closer to a contribution derived from the previous work in explainability than a clear intention of practically integrating a human in the loop.

A few exceptions have been found, and they are highlighted here as relevant practices for integrated and intentional human agency and oversight:

- Aboutalebi et al. [40] develop an explainability-driven framework for machine learning models with two human-guided phases; a first one called “clinician-guided design” phase, where the dataset is preprocessed using explainable AI and domain expert input, and a second one named “explainability-driven design refinement”, where they employ explainability methods not only for transparency but also “to gain a deeper understanding of the clinical validity of the decisions made” along an expert clinician, using such insights to refine the model iteratively. However, such human agency is limited to building the model and does not describe or design a control loop for the ongoing operation of the model.
- Similarly, Lu et al. [42] include a step in their proposed workflow where medical doctors “label the differential diagnosis features with medical guidelines and professional knowledge”; this is supposed to black list meaningless features extracted from electronic health records (EHR). The authors claim to “reduce workloads of clinicians in human-in-loop data mining” as they use oversight features instead of full predictions.
- Bruckert et al. [19] highlight the challenge for “human decision makers to develop trust, which is much needed in life-changing decision tasks” and answer such a challenge with an integrated framework aimed at medical diagnosis. Their inclusion of interactive ML and human-in-the-loop learning ideas enables them to integrate human expert knowledge into ML models so “that humans and machines act as companions within a critical decision task”, and, thus, represents a significant step towards trustworthy and human-overseen expert decision systems.
- Finally, the work “On Assessing Trustworthy AI in Healthcare” by Zicari et al. [18] presents a general translation of the AI HLEG trustworthy AI guidelines to the practice in the healthcare domain, and that includes human agency and oversight. In this regard, they delve into the challenges posed by this requirement. In particular, they illustrate with a practical case the issue of determining the appropriate level of human involvement in the decision-making process, how the AI system might reduce human agent agency and autonomous decision making (which are often critical in emergency call management, for instance), the balance between both actors and, most importantly, the need for an informed decision-making criteria.

3.3. Technical Robustness and Safety

The technical robustness and safety principle emphasizes that AI systems must be resilient and secure. They should function correctly, consistently, and safely under all conditions, with a particular focus on avoiding unintended harm. This has a double side: first, it includes ensuring resilience to attacks and security challenges, maintaining data integrity, and being able to reliably handle errors or inconsistencies during all phases of AI system lifecycles. Second, the AI system should also be accurate and reliable regarding the problem it aims to solve; reproducibility (e.g., being able to obtain a consistent result from the system) is also relevant here.

Therefore, the 21 works (42%) found that include practices related to technical robustness and safety can be divided into two blocks and are described below.

3.3.1. Safety, Including Accuracy and Reliability

Most of the labeled practices involve accuracy, safety, or reliability (19). Obtaining reasonable accuracy while ensuring the trustworthiness of the AI system results, often, in a challenge. The impact of data curation [9] and model calibration [58] is stressed here. Most of the practices, however, can be classified as basic; these include testing multiple models (e.g., in [12]), enhancing techniques such as transfer learning, and common practices in the domain such as cross-validation, ablation studies, and trust score computation [11]). In most cases, these efforts can be summarized as balancing the best possible accuracy but with an arguably explainable model, but studies rarely include fail-safe mechanisms [13] in the design.

While mentions of the safety and reliability of the model abound, the concepts are often found to be mixed with the idea of accuracy. Safety is a broader concept that includes not only the correct functioning of the AI but also ensuring that its operation does not lead to harmful outcomes, while accuracy is a more specific aspect focused on the correctness of task-specific outcomes. Both are crucial for building trust in AI systems, but safety addresses trust from the perspective of harm prevention and overall risk management, whereas accuracy builds trust through the system's performance consistency and the reliability of its outputs, which is the shared aim among the analyzed works.

3.3.2. Security and Resilience to Attacks

Another subdivision is required at this point, as it is necessary to differentiate between security practices related to access and intrusions (e.g., cybersecurity) and resilience to attacks on the models (e.g., inputting malicious data to the model to worsen its performance). Three works ([23,34,37]) that proposed distributed learning for privacy-preserving ML also include security practices to prevent unauthorized access to the workflow. This is complemented by the more technical and specific security practices described in Alamro et al. [55] to enable “IoT devices in the healthcare sector to transmit medical data securely and detect intrusions in the system” and in Soni et al. [56] regarding user authentication.

On the other hand, two relevant examples of robustness practices against attacks have been identified. Moradi and Samwald [10] detail extensive research on improving the robustness of biomedical neural language models against adversarial attacks, highlighting the importance of creating AI systems that can resist manipulations and erroneous inputs, thus ensuring safety and reliability in medical contexts. To do so, they apply various adversarial attack methods such as HotFlip, DeepWordBug, TextBugger, and TextFooler, revealing the vulnerabilities of deep neural NLP models to textual adversaries in the biomedical domain. Similarly, Karim et al. [22] applied equivalent tools (FGSM and DeepFool) to generate adversarial samples (and conduct adversarial retraining) to increase the robustness of their cancer prediction model against malicious data.

3.4. Privacy and Data Governance

The principle of privacy and data governance in trustworthy AI stresses the protection of personal data to ensure privacy throughout an AI system's lifecycle. This involves

implementing robust data handling and security measures and ensuring data integrity and confidentiality, aligned with existing privacy regulations. In spite of 11 studies (22%) being labeled as including practices related to privacy and data governance, a vast majority of them (10) focus on privacy over practices on data governance, which, in turn, is addressed in only a small set (4) of the analyzed studies.

The privacy nature of most medical datasets makes it difficult for clinicians and healthcare service providers to share their sensitive data [48]. For instance, while EHR systems, wearable devices, and medical testing provide “a fertile environment for powerful detection, classification, modeling, monitoring, and control applications in medicine” [15], using such information requires both a privacy-preserving treatment and in-depth data curation [9]. Therefore, obtaining high-quality data to build accurate and reliable models remains problematic owing to substantive legal and ethical constraints in making clinically relevant research data available offsite; even though recent developments such as distributed learning open new possibilities, they unfortunately tend to suffer from a lack of transparency, which undermines trust in what data are used for the analysis [46]. To answer these challenges, most of the identified practices aim to balance the trade-offs between the privacy, accuracy, and transparency aspects of trustworthy AI [18]. Such trade-offs could be assessed with the proposed privacy leakage, interpretability, and transferability metrics described in Kumar et al. [16].

The most common approach in this regard seems to be the combination of ML or DL models with Federated Learning (FL) and blockchain technologies to enable privacy-preserving and distributed learning among multiple clinician collaborators ([23,34,45,46,48]). Their underlying practices are as follows:

- Use distributed learning to keep control over the data. This implies that the data can stay in the owning organization, which should have privacy controls in place without exposing such protected data to the exterior. It is a privacy-preserving approach that also implies a certain degree of data governance (even though it may not cover the full spectrum of the requirement).
- Use agreement techniques for federated/distributed learning to ensure the integrity and robustness of the model, which mainly refers to the technical robustness and safety principles but also points to a data/model quality control.
- Finally, blockchain techniques are suggested to keep the model confidential (with the required level of privacy or access) and also for traceability purposes—which relates to the transparency principle.

Notice how all three “ethical tensions” [18] (privacy, accuracy–quality, and transparency–traceability) are intertwined in these approaches.

Other approaches include cryptographic methods [38] and edge-based privacy-preserving frameworks for IoT (Internet of Things) environments [37], which have arguably similar characteristics but are adjusted for particular scenarios.

3.5. Transparency

The transparency principle remarks the importance of making AI systems and their decision-making processes understandable and traceable. Furthermore, it also calls for clear communication about AI system capabilities and limitations, ensuring that all relevant information is accessible to users for informed decision making. This principle supports accountability and fosters trust among stakeholders by promoting openness in AI operations. In the current analysis, this is significantly the larger category, with 31 papers (62% of the total corpus) that include practices related to the principle. However, virtually all of them target explainability, while only a minor fraction include the idea of traceability (5) or the relevancy of proper communication (2).

The main practice in this regard is to add an additional layer of explainability to a developed method, often applied to explain the output of the obtained ML model, to add a certain degree of human understanding to it; most authors claim that this makes the model “more transparent by explaining why a case receives its prediction and the contributions of

the features to the forecast” [26]. Other studies, such as [52], devote large sections to the explainability of every single layer in the model, a practice that results in a much deeper understanding of the model reasoning. Such information is often used to allow the medical doctor to use it as a second-opinion system [24,44].

In any case, most studies opt for this approach of postanalysis, where the model is already built and trained when queried for explainability. Many tools exist for that purpose; the most commonly found, by far, are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), but many others exist and are found applied in the literature, such as Grad-CAM heat maps, Axiom-based Grad-CAM (A-G-CAM), Attention Branch Network (ABN), layer-wise relevance propagation (LRP), Guided Backpropagation, Perturbations, Occlusions, Explain Like I’m 5 (ELI5), Qlattice, adaptive neuro-fuzzy inference system (ANFIS), pixel density analysis (PDA), and other self-developed methods.

However, as common as these practices might be in the existing literature, they also raise some concerns. Studies that compared some of these methods in their practices found inconsistencies among them when explaining the same results. Chen et al. [43], for instance, found that their “team-based Shapley approach is more consistent than LIME”. Nambiar et al. [14] applied both SHAP and LIME to their model and found that “although SHAP and LIME facilitate model interpretation, the choice between them depends on the specific use case, the nature of the model, and the desired level of explanation”. They found SHAP to be more stable and consistent but computationally intensive, while LIME was more sensitive to perturbations and lighter computationally. In any case, they noticed that “different XAI models could provide different interpretations as well”. In their review of multiple interpretability methods, Salahuddin et al. [17] conclude that “post hoc interpretability methods are only approximations and do not accurately depict the correct model behavior and therefore compromise trust in the explanations”, so they should be used with caution. The complementary case seems to be true too; Stenwig et al. [49] developed four different ML models with similar performance but found that they were widely inconsistent regarding their identification of relevant features when interpreting them using SHAP.

A different approach that can help overcome the pitfalls of post hoc interpretability is to introduce interpretability during the design process. This might require additional data and domain expertise, but it could significantly help stakeholders and developers to build explainable models. Bassiouny et al. [36], for instance, instead of training a model to obtain a diagnostic prediction, train the model to extract seven meaningful features (as defined by experts) that can be easily associated with a specific pathological lung condition; the clinician receives the tagged images and, thus, reaches the final conclusion or diagnoses. This practice both provides natural explainability and introduces a human in the loop. Another related practice is to use tangible features or to create interpretable factors, such as lines, vertices, contours, and the area size of objects in an image [32]. Even though this practice seems to be the most promising, especially in medical imaging, interpretability for image segmentation tasks is still in the early stages of research, while the most common interpretability methodologies have originally been proposed and applied to (general) image classification tasks [47].

3.6. Diversity, Nondiscrimination, and Fairness

The principle of diversity, nondiscrimination, and fairness in trustworthy AI aims to prevent unfair bias by promoting inclusivity; it requires AI systems to be accessible and to deliver equitable results across diverse user groups, ensuring that no particular group is systematically disadvantaged. Similarly, the principle of stakeholder participation in trustworthy AI emphasizes the importance of involving stakeholders throughout the AI system’s lifecycle to ensure that the development and deployment of AI technologies are aligned with society’s diverse values, needs, and expectations, including those affected by its outcomes. A total of 10 works (20%) that include practices related to this requirement were identified.

Half of them (5) include actions related to avoiding unfair bias, which can take many forms [18]. Despite that, only two integrate such actions in their models: Ma et al. [11] discuss some of the common methods used for handling imbalanced data and opt for implementing AUC maximization in their proposed “Trustworthy Deep Learning Framework for Medical Image Analysis” framework. Rashid et al. [15] included adversarial debiasing in their model and found an existing trade-off between the debiasing model and the privacy mechanism, highlighting the ethical conflict between privacy and fair ML. On the other hand, De Paolis Kaluza et al. [39] recognize how “data biases are a known impediment to the development of trustworthy machine learning models and their application to many biomedical problems” and proceed to develop a statistical test for the presence of the general form of bias in labeled data, computing the distance between corresponding class-conditional distributions in labeled and unlabeled data. Real biomedical data are used to test the procedure. Similarly, Mukhopadhyay [54] introduces a new information-theoretic learning framework, called admissible machine learning, and algorithmic risk-management tools (InfoGram, L-features, ALFA-testing) intended to redesign ML methods “to be regulation compliant, while maintaining good prediction accuracy”, highlighting the ethical conflict that raises between fairness and accuracy. They also warn against spurious bias, alerting practitioners about some of the flaws of current fairness criteria.

The other half (5) of the practices can be related to stakeholder participation. As AI systems can raise ethical and societal concerns from direct stakeholders, such as patients in healthcare, and from indirect stakeholders, such as politicians or general media, and include a vast array of topics, it is critical to include them in medical workflows. In this sense, the four models for medical AI described in Gundersen and Baeroe [53] are helpful in framing the different efforts found in the literature:

1. The Ordinary Evidence Model “implies a clear division of labor between the designers of algorithms and medical doctors who are to apply the algorithms in clinical practice”. Even though this is the most commonly found model, the authors claim that due to the accountability and opacity issues inherent to this approach, such “reliance on AI in medicine challenges and even disrupts the professional accountability upon which the ordinary evidence model rests”.
2. The Ethical Design Model addresses ethical concerns in the design process. This approach was found in a few works that involve some relevant stakeholders (e.g., medical doctors), partially in the model design [42] or deployment [13]. However, as Gundersen and Baeroe [53] argue, this approach might be insufficient since stakeholder engagement and ethical deliberation are required both in design and in use.
3. The Collaborative Model states that “collaboration and mutual engagement between medical doctors and AI designers are required to align algorithms with medical expertise, bioethics, and medical ethics”. The idea is to use expert input into AI systems (to calibrate them, make them accessible, and increase the user’s trust in the system) while also helping AI designers understand and interpret the outputs appropriately according to medical practice. Aboutalebi et al. [40] follow this approach, integrating clinicians in both the design and refinement phases. On the other hand, Kumar et al. [30] integrate usability and expert evaluation with doctors (or final users) as a final step in the model implementation to see if it is considered trustable in terms of decision making by the direct stakeholders. This approach might be limited, but it is arguably the only practice related to accessibility that has been found.
4. The Public Deliberation Model involves other relevant stakeholders besides AI designers, bioethicists, and medical experts; policymakers and the general public are included in this group for broader discussions about the transformative impact of a new AI technology and/or societal impact. No practices were found that involved this approach.

3.7. Accountability

The accountability principle for trustworthy AI demands that mechanisms are in place to ensure responsibility and accountability for AI systems and their outcomes. This includes auditability and the ability to report and address negative impacts that may arise from AI systems. It must be noted that it is not the same as transparency, although it is undoubtedly helpful to build accountable systems.

Most of the identified work with practices that can be linked to accountability (7 in total, 14%) can be labeled as risk management (6) in nature. In medical applications, out-of-domain (OOD) data are a huge challenge, as high accuracy in training often does not translate to similar performance with OOD testing data. Different factors that may affect the distribution of clinical data include variability in data collection parameters, differences between patients, and rare data classes, which are less uncommon here (using a different imaging machine, for instance, could dramatically impact performance, becoming dangerous to patients). Stolte et al. [58] propose a loss function to improve calibration, “ensuring that the model makes less risky errors even when incorrect” and, thus, reducing risk. A common practice found in other works [13,19,43,57] is to focus on enhanced explainability (either by providing human-understandable explanations, human-identifiable features, or fail-safe mechanisms) that are provided as a set of tools and information to final users (e.g., clinicians or medical doctors) to either take a decision or to review arguably risky cases that might fall out of the model’s domain. These practices can be considered as methods to partially mitigate the risk associated with medical decisions based on AI systems.

On the other hand, only a few practices (3) can be associated with the purpose of auditability and, as discussed before, they are mostly extensions of the explainability efforts. For instance, Bruckert et al. [19] state that their practical framework “emphasizes the need for systems that allow for bidirectional information exchange between humans and machines, supporting accountability by making it easier for users to understand and question AI decisions”, while Kamal et al. [24] use tools such as SP-LIME to analyze the processing of a particular decision and show the associated features so “clinicians/medical experts and patients can easily understand the decision-making process”.

4. Discussion

In this scoping review, a review of existing decision-making systems in the biomedical field and a consolidation of existing practices related to the principles of reliable AI has been carried out. Even though our search resulted in 168 nonduplicate articles, only a fraction, 50, were ultimately included. The attrition of articles between screening, eligibility, and inclusion also hints at the gap between recommendations, theoretical frameworks, and actual practices. This lack of more studies could be attributed to the relatively recent embrace of Artificial Intelligence in healthcare, which has dramatically changed the medical landscape [59] but is still in the early stages, as the integration of human and Artificial Intelligence (AI) for medicine has barely begun [1], let alone the refined application and understanding of the trustworthy AI principles. What follows is a summary of the review’s main findings and the most relevant gaps identified in the existing literature.

4.1. Explainability Is Key

Explainability is highly relevant as it bridges the gap between AI decisions and human understanding, which is crucial for acceptance and ethical application in sensitive fields like healthcare. Thus, a large majority of the identified practices incorporate explainability as an essential layer to their models and AI systems, focusing on being able to understand (or interpret) the outcomes of the ML models. This is achieved by using multiple tools to explain the contributions of specific features to a prediction, aiming to enhance user comprehension and trust. The most popular tools for explainability after model training include SHAP and LIME, with others like Grad-CAM heat maps and various proprietary methods also being utilized. These tools allow for the analysis of how different features influence model predictions.

However, comparisons among different explainability tools (like SHAP vs. LIME) reveal inconsistencies in explanations for the same outcomes, which can undermine trust and reliability and require further exploration or the adoption of other methods, as many post hoc interpretability methods do not fully capture the true behavior of models, potentially misrepresenting how decisions are made. Integrating the interpretability in the models in the design phase (e.g., with explainable features) could prevent such caveats, although research in this regard is still scarce. Furthermore, current research often focuses on general models, with limited exploration into domain-specific applications like image segmentation within medical imaging, where tailored interpretability is crucial.

4.2. Efforts on Model Accuracy

Both safety and accuracy are crucial for building trust in AI systems. Safety ensures that the AI operates without causing harm, encompassing comprehensive risk management, while accuracy pertains to the correctness of task-specific outcomes, enhancing the system's reliability and performance consistency. The extracted practices in technical robustness and safety are mostly related to achieving reasonable accuracy without compromising the AI system's trustworthiness. This balance appears to be one of the main ethical challenges for implementing AI systems in healthcare. Techniques like model calibration, data curation, cross-validation, and transfer learning are common. However, fail-safe mechanisms, although vital, are less frequently mentioned or implemented.

Although a significant number of studies focus on accuracy and performance, there is a significant need for more research into integrating fail-safe mechanisms directly into the AI design process, which could prevent harmful outcomes before they occur. This includes specific safety mechanisms that should go beyond error handling and include proactive strategies to prevent harm in various operational environments.

On the other hand, while some studies address cybersecurity and resilience against data manipulation, this area requires further differentiation and exploration. Practices must focus on protecting against unauthorized access and enhancing resilience against data tampering that could degrade the model's performance. There is indeed a need for more practical implementations of security measures that ensure data integrity and confidentiality in real-world applications, especially in sensitive domains like healthcare.

4.3. Privacy Preservation

From the recovered studies, it becomes clear that personal data protection is considered crucial, especially in the medical field, where data sensitivity is high. Effective data governance and privacy measures ensure that personal information is handled securely, maintaining both integrity and confidentiality throughout the data lifecycle; most practices, however, favor privacy over data governance, most likely due to its criticality and immediacy. There is a noticeable gap in comprehensive data governance practices that go beyond privacy to include aspects like data quality control, data usage policies, and lifecycle management.

Furthermore, while many practices aim to balance privacy with accuracy and transparency, the methods to assess these trade-offs, such as privacy leakage and interpretability metrics, require further exploration and validation. On the technical side, although Federated Learning and blockchain are prominent, there is room for more innovative approaches that enhance data privacy without compromising data utility or transparency. While distributed learning offers a way to maintain control over data, ensuring robust privacy and data integrity in these settings remains a challenge. More detailed and specific practices must be developed to enhance trust in these technologies. Cryptographic methods and edge-based frameworks also show a certain promise, but more case studies and practical implementations would be needed to establish their effectiveness in real-world scenarios.

Therefore, while significant advancements have been made in privacy-preserving AI practices, particularly using Federated Learning and blockchain, the field requires more

comprehensive approaches to data governance and innovative solutions to balance privacy with other critical aspects of trustworthy AI.

4.4. Testimonial Human Agency and Stakeholder Integration

A representative sample of studies indirectly supports human agency and oversight through their explainability efforts, as these practices often enhance the transparency of AI systems, enabling healthcare professionals to understand and potentially intervene in AI decisions, particularly for critical medical decisions. However, while there are significant efforts to integrate human agency and oversight within AI systems in healthcare, significant gaps remain in ensuring that these systems are designed and operated in ways that truly respect user autonomy and provide meaningful opportunities for human intervention.

While initial model design often includes human input, ongoing human oversight during the operational phase is less commonly addressed. This gap suggests a need for systems and practices that allow continuous human interaction and control, not just during the development phase but throughout the lifecycle of the AI system. Additionally, as mentioned before, many studies emphasize explainability as a tool for oversight but do not fully integrate mechanisms that ensure human control over AI actions. There is a need for clear strategies and practical methods that maintain human control, especially in making critical healthcare decisions. In this sense, there is a critical need to define and implement optimal levels of human involvement that balance the benefits of AI automation with the irreplaceable nuances of human judgment, particularly in life-changing or emergency medical decisions.

While some studies involve healthcare professionals in the data preprocessing or feature selection phases, fewer studies address how these experts can actively participate in ongoing AI decision-making processes. Practical frameworks are needed that detail the roles of human professionals in regular AI operations and integrate as many stakeholders as possible in the process. In this sense, although several studies claim to implement human-in-the-loop systems, the depth and impact of human involvement often remain vague. Future research should aim to define and quantify the impact of human participation on AI outcomes, ensuring it is meaningful and not just a supplementary check.

4.5. Missing Techniques

Most of the analyzed works apply either machine learning or deep learning models to a dataset and, post hoc, add an explainability layer aiming to understand the reasoning applied by the model. It is particularly relevant, though, to notice the absence of techniques that are specifically designed to fill that gap from their own design. For instance, causal ML goes beyond correlation in data to understand the underlying cause-and-effect relationships. In healthcare, understanding these causal relationships could be crucial for determining the effectiveness of treatments and interventions on health outcomes. However, only a single and recent study [23] uses or even mentions it. Replacing black-box methods with causal methods could potentially enhance the trustworthiness of these systems by understanding the causal pathways, as healthcare providers, for instance, could make more informed decisions about patient care, leading to personalized treatment plans based on the likely effectiveness of different options for individual patients. Causal ML also helps identify true causal relationships, reducing the risk of adopting ineffective or harmful medical practices based on incorrect data interpretations, which are critical in this context.

On the other hand, and although privacy-preserving ML seems to be a relevant topic, none of the works discuss any anonymization techniques or practices, which stands out, as healthcare providers are bound by strict regulations and anonymization facilitates compliance with these laws by ensuring that data used for training AI systems does not compromise patient privacy, also enabling sharing information across institutions for research without violating privacy laws and promoting collaboration and the pooling of data resources, which is essential for developing robust AI models.

4.6. Legal Loopholes

Several critical points emerge from the perspective of legal consequences, implications, and gaps in the application of the accountability principle for trustworthy AI in healthcare. The principle of accountability requires that AI systems be developed and implemented in a manner where responsibility for outcomes can be clearly assigned, which is particularly significant in healthcare due to the potential for life-altering consequences. There is a need for clear mechanisms for accountability to ensure that when AI systems are used in healthcare, liability can be determined in cases of malpractice or harm. This involves, among others, understanding which party (the software developer, the healthcare provider, or another entity) is responsible when an AI-driven decision leads to a negative outcome. However, it is also important to stress that healthcare is a highly regulated sector, and AI systems must comply with existing medical device regulations and data protection laws. Accountability practices such as audit trails and transparency reports would be essential to demonstrate compliance with these regulations.

The current practices, however, fail to adequately establish how responsibility is assigned between AI developers, healthcare providers, and other stakeholders. This lack of clarity can complicate legal proceedings and patient recourse in case of an AI-induced error. Furthermore, while auditability is mentioned in a few works, the specific standards and methodologies for auditing AI in healthcare seem to be still underdeveloped, potentially leading to unchecked deployment of AI systems. Furthermore, despite the prevalent focus on risk management, such as developing AI systems that make fewer risky errors, the legal frameworks to support these practices are not always clear. Finally, the connection between explainability and accountability highlights a gap in legal requirements for explainable AI. While explainable AI practices are often developed to enhance user trust and understanding, their role in fulfilling legal standards for accountability needs more explicit recognition and integration. A collaborative effort between policymakers, legal experts, AI developers, and healthcare providers is thus required to ensure that AI systems are not only technically sound and ethically aligned but also legally compliant and accountable.

4.7. Lack of Holistic Approaches

Holistic approaches are essential for developing and implementing AI systems that are not only technically proficient but also ethically sound, legally compliant, and widely accepted by all stakeholders, ensuring that the benefits of AI in healthcare are realized safely and effectively. However, most of the practices found in the existing literature consider a single requirement (besides societal well-being) in their practice (23, 46% of the total practices). Only one paper proposes a holistic, practical process to expose “the complex trade-offs and the necessity for such thorough human review when tackling sociotechnical applications of AI in healthcare” [18].

Thus, this raises a significant gap between trustworthy AI requirements—that are general—and actual practices—that often focus on a single requirement. This is a potential risk, as healthcare is a complex field where decisions can have life-altering consequences. A holistic approach would ensure that all aspects of AI implementation, from technical accuracy to ethical implications, are considered, minimizing potential risks to patients and improving overall safety in turn. A holistic assessment ensures that AI solutions are compatible across different platforms and can be seamlessly integrated without disrupting existing workflows or compromising data security, too.

The idea of an end-to-end approach brings further benefits in the form of being able to include diverse stakeholders, including patients, clinicians, administrators, and regulators, ensuring that the AI systems meet the needs and expectations of all parties involved and maintain human-centric values, and helps build systems that are robust and adaptable to changes, whether they are technological advancements or shifts in healthcare policies, while helping ensure that these technologies adhere to ethical norms such as fairness, nondiscrimination, and respect for patient autonomy, thus upholding the trust and integrity of healthcare services. A holistic approach to assessing AI also ensures compliance

with all applicable laws and regulations, avoiding legal repercussions and ensuring that the AI systems are lawful and fit for purpose, also facilitating the integration of accountability mechanisms throughout the AI system's lifecycle, from design and development to deployment and postdeployment monitoring.

4.8. Future Work

Future work should consider establishing a comprehensive framework covering the specific needs of trustworthy AI for healthcare, considering the particular restrictions and risks that exist in this scope. There is a noticeable need for more reliable explainability methods and practical integration of explainability during the design process, rather than relying on post hoc interpretability tools that have shown inconsistencies and unreliability. From a robustness point of view, there is a general lack of comprehensive safety and security strategies, including fail-safe mechanisms and robust measures against data manipulation, essential for patient safety; this links with a clear underdevelopment of data governance practices beyond privacy, such as data quality control, data usage policies, and lifecycle management. Additionally, environmental impact is a requirement that has not been covered by any practice and, thus, requires further investigation.

Furthermore, future work should consider not only whether these approaches to address trustworthy AI principles are included but also the quality and appropriateness of such practices. In this sense, such work should also be used as a starting point to guide researchers, practitioners, and policymakers in implementing trustworthy AI principles in medical research.

5. Conclusions

Although a growing interest and engagement with trustworthy Artificial Intelligence within the medical and biomedical sectors has been identified (RQ1), demonstrated by the increasing use of AI in critical areas like diagnostics and patient management, the translation of theoretical frameworks into actual trustworthy AI practices remains limited, as evidenced by a relatively small number of dedicated papers. Most studies originate from Europe, likely influenced by regulatory efforts there, with significant contributions from North America and Asia as well. The analysis shows that while bioinformatics and medical imaging are the primary focus areas, there is broad involvement across various medical specialties. However, there remains a general gap in applying these practices across different regions and specialties, emphasizing the need for a more comprehensive integration of trustworthy AI principles in healthcare practices globally.

The scoping review highlights that while interest in trustworthy AI in healthcare is growing, the actual implementation of its principles varies significantly across different requirements (RQ2). While the detailed approaches and commonalities (RQ3) and gaps (RQ4) can be found in Sections 3 and 4, the highlights are presented here:

- Most identified practices focus on explainability, aiming to make AI systems more understandable to healthcare professionals, which is crucial for ethical and informed decision-making. However, the review points out inconsistencies among various explainability tools, suggesting a need for more reliable methods.
- Regarding technical robustness and safety, the practices mostly focus on achieving a balance between system accuracy and trustworthiness, with less emphasis on integrating fail-safe mechanisms and specific security measures against data manipulation. This indicates a gap in comprehensive safety and security strategies essential for patient safety.
- Privacy and data governance are also predominantly focused on privacy aspects rather than comprehensive data governance. The common use of Federated Learning and blockchain suggests a focus on innovative privacy-preserving technologies, yet comprehensive data management practices remain underdeveloped.
- Regarding human agency and oversight, the review notes that most practices support human control indirectly through explainability efforts rather than through direct

mechanisms that ensure meaningful human oversight throughout the AI lifecycle. This points to a need for more explicit integration of human control in the operational use of AI.

Overall, the review reflects a fragmented implementation of trustworthy AI principles in healthcare, with significant advancements in some areas but notable gaps in others, such as security, comprehensive data governance, and effective integration of human oversight, raising similar concerns as found in previous review efforts [4,59]. This suggests a need for a more holistic approach to incorporating these principles to realize the full benefits of trustworthy AI in healthcare.

Author Contributions: Conceptualization, M.-Á.S.; data curation, M.M.-C.; formal analysis, M.M.-C.; investigation, M.M.-C. and E.G.-B.; methodology, M.M.-C.; resources, M.-Á.S.; supervision, M.-Á.S.; validation, M.M.-C., E.G.-B. and M.-Á.S.; writing—original draft, M.M.-C.; writing—review and editing, E.G.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [[CrossRef](#)] [[PubMed](#)]
2. Directorate-General for Communications Networks; Content and Technology (European Commission); Grupa Ekspertów Wysokiego Szczebla ds. Sztucznej Inteligencji. *Ethics Guidelines for Trustworthy AI*; Publications Office: Brussels, Belgium, 2019. [[CrossRef](#)]
3. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [[CrossRef](#)] [[PubMed](#)]
4. Goirand, M.; Austin, E.; Clay-Williams, R. Implementing ethics in healthcare AI-based applications: A scoping review. *Sci. Eng. Ethics* **2021**, *27*, 61. [[CrossRef](#)] [[PubMed](#)]
5. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.; Horsley, T.; Weeks, L.; et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [[CrossRef](#)] [[PubMed](#)]
6. Rong, G.; Mendez, A.; Assi, E.B.; Zhao, B.; Sawan, M. Artificial intelligence in healthcare: Review and prediction case studies. *Engineering* **2020**, *6*, 291–301. [[CrossRef](#)]
7. Silcox, C.; Zimlichmann, E.; Huber, K.; Rowen, N.; Saunders, R.; McClellan, M.; Kahn, C.N.; Salzberg, C.A.; Bates, D.W. The potential for artificial intelligence to transform healthcare: Perspectives from international health leaders. *NPJ Digit. Med.* **2024**, *7*, 88. [[CrossRef](#)] [[PubMed](#)]
8. Federico, C.A.; Trotsyuk, A.A. Biomedical Data Science, Artificial Intelligence, and Ethics: Navigating Challenges in the Face of Explosive Growth. *Annu. Rev. Biomed. Data Sci.* **2024**, *7*. [[CrossRef](#)] [[PubMed](#)]
9. Alves, V.M.; Auerbach, S.S.; Kleinstreuer, N.; Rooney, J.P.; Muratov, E.N.; Rusyn, I.; Tropsha, A.; Schmitt, C. Models Out: The Impact of Data Quality on the Reliability of Artificial Intelligence Models as Alternatives to Animal Testing. *Altern. Lab. Anim.* **2021**, *49*, 73–82. [[CrossRef](#)] [[PubMed](#)]
10. Moradi, M.; Samwald, M. Improving the robustness and accuracy of biomedical language models through adversarial training. *J. Biomed. Inf.* **2022**, *132*, 104114. [[CrossRef](#)]
11. Ma, K.; He, S.; Sinha, G.; Ebadi, A.; Florea, A.; Tremblay, S.; Wong, A.; Xi, P. Towards Building a Trustworthy Deep Learning Framework for Medical Image Analysis. *Sensors* **2023**, *23*, 8122. [[CrossRef](#)]
12. Khanna, V.V.; Chadaga, K.; Sampathila, N.; Prabhu, S.; Bhandage, V.; Hegde, G.K. A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome. *Appl. Syst. Innov.* **2023**, *6*, 32. [[CrossRef](#)]
13. Fidon, L.; Aertsen, M.; Kofler, F.; Bink, A.; David, A.L.; Deprest, T.; Emam, D.; Guffens, F.; Jakab, A.; Kasprian, G.; et al. A Dempster-Shafer Approach to Trustworthy AI with Application to Fetal Brain MRI Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 3784–3795. [[CrossRef](#)] [[PubMed](#)]
14. Nambiar, A.; Harikrishnaa, S.; Sharanprasad, S. Model-agnostic explainable artificial intelligence tools for severity prediction and symptom analysis on Indian COVID-19 data. *Front. Artif. Intell.* **2023**, *6*, 1272506. [[CrossRef](#)] [[PubMed](#)]
15. Rashid, M.M.; Askari, M.R.; Chen, C.; Liang, Y.; Shu, K.; Cinar, A. Artificial Intelligence Algorithms for Treatment of Diabetes. *Algorithms* **2022**, *15*, 299. [[CrossRef](#)]
16. Kumar, M.; Moser, B.A.; Fischer, L.; Freudenthaler, B. An Information Theoretic Approach to Privacy-Preserving Interpretable and Transferable Learning. *Algorithms* **2023**, *16*, 450. [[CrossRef](#)]
17. Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [[CrossRef](#)] [[PubMed](#)]

18. Zicari, R.V.; Brusseau, J.; Blomberg, S.N.; Christensen, H.C.; Coffee, M.; Ganapini, M.B.; Gerke, S.; Gilbert, T.K.; Hickman, E.; Hildt, E.; et al. On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Front. Hum. Dyn.* **2021**, *3*, 673104. [[CrossRef](#)]
19. Bruckert, S.; Finzel, B.; Schmid, U. The Next Generation of Medical Decision Support: A Roadmap toward Transparent Expert Companions. *Front. Artif. Intell.* **2020**, *3*, 507973. [[CrossRef](#)]
20. Ma, J.; Schneider, L.; Lapuschkin, S.; Achtibat, R.; Duchrau, M.; Krois, J.; Schwendicke, F.; Samek, W. Towards Trustworthy AI in Dentistry. *J. Dent. Res.* **2022**, *101*, 1263–1268. [[CrossRef](#)]
21. Imboden, S.; Liu, X.; Payne, M.C.; Hsieh, C.J.; Lin, N.Y.C. Trustworthy in silico cell labeling via ensemble-based image translation. *Biophys. Rep.* **2023**, *3*, 100133. [[CrossRef](#)]
22. Karim, M.R.; Islam, T.; Lange, C.; Rebholz-Schuhmann, D.; Decker, S. Adversary-Aware Multimodal Neural Networks for Cancer Susceptibility Prediction from Multiomics Data. *IEEE Access* **2022**, *10*, 54386–54409. [[CrossRef](#)]
23. Mu, J.; Kadoch, M.; Yuan, T.; Lv, W.; Liu, Q.; Li, B. Explainable Federated Medical Image Analysis through Causal Learning and Blockchain. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 3206–3218. [[CrossRef](#)] [[PubMed](#)]
24. Kamal, M.S.; Dey, N.; Chowdhury, L.; Hasan, S.I.; Santosh, K.C. Explainable AI for Glaucoma Prediction Analysis to Understand Risk Factors in Treatment Planning. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 3171613. [[CrossRef](#)]
25. Hassan, M.M.; Alqahtani, S.A.; Alrakhami, M.S.; Elhendi, A.Z. Transparent and Accurate COVID-19 Diagnosis: Integrating Explainable AI with Advanced Deep Learning in CT Imaging. *CMES-Comput. Model. Eng. Sci.* **2024**, *139*. [[CrossRef](#)]
26. Tasnim, N.; Al Mamun, S.; Shahidul Islam, M.; Kaiser, M.S.; Mahmud, M. Explainable Mortality Prediction Model for Congestive Heart Failure with Nature-Based Feature Selection Method. *Appl. Sci.* **2023**, *13*, 6138. [[CrossRef](#)]
27. El-Rashidy, N.; Sedik, A.; Siam, A.I.; Ali, Z.H. An efficient edge/cloud medical system for rapid detection of level of consciousness in emergency medicine based on explainable machine learning models. *Neural Comput. Appl.* **2023**, *35*, 10695–10716. [[CrossRef](#)] [[PubMed](#)]
28. Prifti, E.; Chevaleyre, Y.; Hanczar, B.; Belda, E.; Danchin, A.; Clement, K.; Zucker, J.D. Interpretable and accurate prediction models for metagenomics data. *Gigascience* **2020**, *9*, giaa010. [[CrossRef](#)]
29. Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Radhakrishnan, Y.; Cheungpasitporn, W. Chain of Thought Utilization in Large Language Models and Application in Nephrology. *Medicina* **2024**, *60*, 148. [[CrossRef](#)]
30. Kumar, A.; Manikandan, R.; Kose, U.; Gupta, D.; Satapathy, S.C. Doctor’s Dilemma: Evaluating an Explainable Subtractive Spatial Lightweight Convolutional Neural Network for Brain Tumor Diagnosis. *ACM Trans. Multimed. Comput. Commun. Appl.* **2021**, *17*, 1–26. [[CrossRef](#)]
31. Vijayvargiya, A.; Singh, P.; Kumar, R.; Dey, N. Hardware Implementation for Lower Limb Surface EMG Measurement and Analysis Using Explainable AI for Activity Recognition. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–9. [[CrossRef](#)]
32. Pintelas, E.; Livieris, I.E.; Pintelas, P. Explainable Feature Extraction and Prediction Framework for 3D Image Recognition Applied to Pneumonia Detection. *Electronics* **2023**, *12*, 2663. [[CrossRef](#)]
33. Wang, Z.; Samsten, I.; Kougia, V.; Papapetrou, P. Style-transfer counterfactual explanations: An application to mortality prevention of ICU patients. *Artif. Intell. Med.* **2023**, *135*, 102457. [[CrossRef](#)]
34. Lugan, S.; Desbordes, P.; Brion, E.; Tormo, L.X.R.; Legay, A.; Macq, B. Secure Architectures Implementing Trusted Coalitions for Blockchain Distributed Learning (TCLearn). *IEEE Access* **2019**, *7*, 181789–181799. [[CrossRef](#)]
35. Shukla, S.; Birla, L.; Gupta, A.K.; Gupta, P. Trustworthy Medical Image Segmentation with improved performance for in-distribution samples. *Neural Netw.* **2023**, *166*, 127–136. [[CrossRef](#)]
36. Bassiouny, R.; Mohamed, A.; Umapathy, K.; Khan, N.; IEEE. An Interpretable Object Detection-Based Model for the Diagnosis of Neonatal Lung Diseases Using Ultrasound Images. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico City, Mexico, 1–5 November 2021. [[CrossRef](#)]
37. Jiang, R.; Chazot, P.; Pavese, N.; Crookes, D.; Bouridane, A.; Celebi, M.E. Private Facial Prediagnosis as an Edge Service for Parkinson’s DBS Treatment Valuation. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2703–2713. [[CrossRef](#)] [[PubMed](#)]
38. Abdelfattah, S.; Baza, M.; Mahmoud, M.; Fouada, M.M.; Abualsaoud, K.; Yaacoub, E.; Alsabaan, M.; Guizani, M. Lightweight Multi-Class Support Vector Machine-Based Medical Diagnosis System with Privacy Preservation. *Sensors* **2023**, *23*, 9033. [[CrossRef](#)] [[PubMed](#)]
39. De Paolis Kaluza, M.C.; Jain, S.; Radivojac, P. An Approach to Identifying and Quantifying Bias in Biomedical Data. In Proceedings of the Pacific Symposium on Biocomputing 2023, Kohala Coast, HI, USA, 3–7 January 2023; Volume 28, pp. 311–322.
40. Aboutalebi, H.; Pavlova, M.; Shafiee, M.J.; Florea, A.; Hryniowski, A.; Wong, A. COVID-Net Biochem: An explainability-driven framework to building machine learning models for predicting survival and kidney injury of COVID-19 patients from clinical and biochemistry data. *Sci. Rep.* **2023**, *13*, 17001. [[CrossRef](#)] [[PubMed](#)]
41. Uzunova, H.; Ehrhardt, J.; Kepp, T.; Handels, H. Interpretable Explanations of Black Box Classifiers Applied on Medical Images by Meaningful Perturbations Using Variational Autoencoders. In *Medical Imaging 2019: Image Processing*; SPIE: Bellingham, WA, USA, 2019. [[CrossRef](#)]
42. Lu, K.; Tong, Y.; Yu, S.; Lin, Y.; Yang, Y.; Xu, H.; Li, Y.; Yu, S. Building a trustworthy AI differential diagnosis application for Crohn’s disease and intestinal tuberculosis. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 160. [[CrossRef](#)]
43. Chen, Y.; Aleman, D.M.; Purdie, T.G.; McIntosh, C. Understanding machine learning classifier decisions in automated radiotherapy quality assurance. *Phys. Med. Biol.* **2022**, *67*, 025001. [[CrossRef](#)]

44. Araujo, T.; Aresta, G.; Mendonca, L.; Penas, S.; Maia, C.; Carneiro, A.; Maria Mendonca, A.; Campilho, A. DR | GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Med. Image Anal.* **2020**, *63*, 101715. [[CrossRef](#)]
45. Malik, H.; Anees, T.; Naeem, A.; Naqvi, R.A.; Loh, W.K. Blockchain-Federated and Deep-Learning-Based Ensembling of Capsule Network with Incremental Extreme Learning Machines for Classification of COVID-19 Using CT Scans. *Bioengineering* **2023**, *10*, 203. [[CrossRef](#)] [[PubMed](#)]
46. Zerka, F.; Urovi, V.; Vaidyanathan, A.; Barakat, S.; Leijenaar, R.T.H.; Walsh, S.; Gabrani-Juma, H.; Miraglio, B.; Woodruff, H.C.; Dumontier, M.; et al. Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM). *IEEE Access* **2020**, *8*, 183939–183951. [[CrossRef](#)]
47. Saleem, H.; Shahid, A.R.; Raza, B. Visual interpretability in 3D brain tumor segmentation network. *Comput. Biol. Med.* **2021**, *133*, 104410. [[CrossRef](#)] [[PubMed](#)]
48. El Houda, Z.A.; Hafid, A.S.; Khoukhi, L.; Brik, B. When Collaborative Federated Learning Meets Blockchain to Preserve Privacy in Healthcare. *IEEE Trans. Netw. Sci. Eng.* **2023**, *10*, 2455–2465. [[CrossRef](#)]
49. Stenwig, E.; Salvi, G.; Rossi, P.S.; Skjaervold, N.K. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Med. Res. Methodol.* **2022**, *22*, 53. [[CrossRef](#)] [[PubMed](#)]
50. Ogbomo-Harmitt, S.; Muffoletto, M.; Zeidan, A.; Qureshi, A.; King, A.P.; Aslanidi, O. Exploring interpretability in deep learning prediction of successful ablation therapy for atrial fibrillation. *Front. Physiol.* **2023**, *14*, 1054401. [[CrossRef](#)] [[PubMed](#)]
51. Alzubaidi, L.; Salhi, A.; A Fadhel, M.; Bai, J.; Hollman, F.; Italia, K.; Pareyon, R.; Albahri, A.S.; Ouyang, C.; Santamaria, J.; et al. Trustworthy deep learning framework for the detection of abnormalities in X-ray shoulder images. *PLoS ONE* **2024**, *19*, e0299545. [[CrossRef](#)] [[PubMed](#)]
52. El-Sappagh, S.; Alonso, J.M.; Islam, S.M.R.; Sultan, A.M.; Kwak, K.S. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci. Rep.* **2021**, *11*, 2660. [[CrossRef](#)] [[PubMed](#)]
53. Gundersen, T.; Baeroe, K. The Future Ethics of Artificial Intelligence in Medicine: Making Sense of Collaborative Models. *Sci. Eng. Ethics* **2022**, *28*, 17. [[CrossRef](#)]
54. Mukhopadhyay, S. InfoGram and admissible machine learning. *Mach. Learn.* **2022**, *111*, 205–242. [[CrossRef](#)]
55. Alamro, H.; Marzouk, R.; Alruwais, N.; Negm, N.; Aljameel, S.S.; Khalid, M.; Hamza, M.A.; Alsaid, M.I. Modeling of Blockchain Assisted Intrusion Detection on IoT Healthcare System Using Ant Lion Optimizer with Hybrid Deep Learning. *IEEE Access* **2023**, *11*, 82199–82207. [[CrossRef](#)]
56. Soni, P.; Pradhan, J.; Pal, A.K.; Islam, S.K.H. Cybersecurity Attack-Resilience Authentication Mechanism for Intelligent Healthcare System. *IEEE Trans. Ind. Inform.* **2023**, *19*, 830–840. [[CrossRef](#)]
57. Cho, K.H.; Kim, E.S.; Kim, J.W.; Yun, C.H.; Jang, J.W.; Kasani, P.H.; Jo, H.S. Comparative effectiveness of explainable machine learning approaches for extrauterine growth restriction classification in preterm infants using longitudinal data. *Front. Med.* **2023**, *10*, 1166743. [[CrossRef](#)] [[PubMed](#)]
58. Stolte, S.E.; Volle, K.; Indahlastari, A.; Albizu, A.; Woods, A.J.; Brink, K.; Hale, M.; Fang, R. DOMINO: Domain-aware loss for deep learning calibration. *Softw. Impacts* **2023**, *15*, 100478. [[CrossRef](#)]
59. Albahri, A.S.; Duhamim, A.M.; Fadhel, M.A.; Alnoor, A.; Baqer, N.S.; Alzubaidi, L.; Albahri, O.S.; Alamoodi, A.H.; Bai, J.; Salhi, A.; et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf. Fusion* **2023**, *96*, 156–191. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.