systems

MDPI

Article

Risk Analysis of Artificial Intelligence in Medicine with a Multilayer Concept of System Order

Negin Moghadasi [1,*], Rupa S. Valdez [1], Misagh Piran [2], Negar Moghaddasi [3], Igor Linkov [4], Thomas L. Polmateer [1], Davis C. Loose [1] and James H. Lambert [1]

[1] Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22903, USA; lambert@virginia.edu (J.H.L.)
[2] Department of Radiology, Nuclear Medicine and Molecular Imaging, Heart and Diabetes Center North-Rhine Westphalia, Ruhr University of Bochum, 44801 Bochum, Germany; mpiran@hdz-nrw.de
[3] Department of Dentistry, Western University of Health Sciences, Pomona, CA 91766, USA; ne.moghaddasijahromi@westernu.edu
[4] Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213, USA; igor.linkov@usace.army.milf
* Correspondence: nm2fs@virginia.edu

Abstract: Artificial intelligence (AI) is advancing across technology domains including healthcare, commerce, the economy, the environment, cybersecurity, transportation, etc. AI will transform healthcare systems, bringing profound changes to diagnosis, treatment, patient care, data, medicines, devices, etc. However, AI in healthcare introduces entirely new categories of risk for assessment, management, and communication. For this topic, the framing of conventional risk and decision analyses is ongoing. This paper introduces a method to quantify risk as the disruption of the order of AI initiatives in healthcare systems, aiming to find the scenarios that are most and least disruptive to system order. This novel approach addresses scenarios that bring about a re-ordering of initiatives in each of the following three characteristic layers: purpose, structure, and function. In each layer, the following model elements are identified: 1. Typical research and development initiatives in healthcare. 2. The ordering criteria of the initiatives. 3. Emergent conditions and scenarios that could influence the ordering of the AI initiatives. This approach is a manifold accounting of the scenarios that could contribute to the risk associated with AI in healthcare. Recognizing the context-specific nature of risks and highlighting the role of human in the loop, this study identifies scenario *s.06*—non-interpretable AI and lack of human–AI communications—as the most disruptive across all three layers of healthcare systems. This finding suggests that AI transparency solutions primarily target domain experts, a reasonable inclination given the significance of "high-stakes" AI systems, particularly in healthcare. Future work should connect this approach with decision analysis and quantifying the value of information. Future work will explore the disruptions of system order in additional layers of the healthcare system, including the *environment*, *boundary*, *interconnections*, *workforce*, *facilities*, *supply chains*, and others.

Keywords: risk management; risk communication; interpretable and explainable AI; systems engineering; scenario-based preferences

Citation: Moghadasi, N.; Valdez, R.S.; Piran, M.; Moghaddasi, N.; Linkov, I.; Polmateer, T.L.; Loose, D.C.; Lambert, J.H. Risk Analysis of Artificial Intelligence in Medicine with a Multilayer Concept of System Order. *Systems* **2024**, *12*, 47. https://doi.org/10.3390/systems12020047

Academic Editor: Ed Pohl

Received: 8 November 2023
Revised: 9 January 2024
Accepted: 30 January 2024
Published: 1 February 2024

1. Introduction

System engineering plays a vital role in informing the design of systems that can effectively respond to unprecedented and unimagined disruptions. Risk, safety, security, trust, and resilience programs are implemented to address the scope, allocation of resources, and evaluation of these complex systems. The conventional risk definition is a hallmark of medical statistics and epidemiology, as mentioned by [1], and the concept of risk as a disruptive event is expressed in other contexts, e.g., in immunity [2], or more general, in technology [3]. Studies by [4–6] focus on addressing the challenges associated with

risk management [7–9], safety assurance, security measures, and resilience strategies within these systems. By incorporating system modeling and engineering approaches, organizations can better understand and navigate the evolving priority orders of complex systems, enabling them to adapt and respond effectively to disruptions and ensure the robustness and effectiveness of their operations. With the continuous progress of science in the healthcare and medical sectors, there is an increasing need to enhance services provided to users. This growing demand has led to the adoption of advanced technologies, including artificial intelligence (AI) [10], to meet the surge in requirements. AI has revolutionized healthcare by advancing the state of the art for diagnoses [11–13], treatments, disease prevention, and surgical devices. AI valuation in the European healthcare market exceeds USD 1.15 billion in 2020 and is expected to grow more than 44.2% through 2027 [14]. AI in healthcare has the potential to significantly improve outcomes [15] and reduce procedure time and costs [16].

Utilization of AI in healthcare faces many challenges and risks. There is a particular concern regarding risks related to applications of AI and machine learning. AI should be valid, reliable, safe, fair [17], unbiased [18], secure, resilient, explainable, interpretable [19], accountable, and transparent [20–22].

The National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (NIST AI RMF), published in 2023, addresses risks in designing, developing, using, and evaluating AI systems and products [22]. The framework discusses the requirements for trustworthy AI applications [23,24]. NIST proposes aspects of trustworthy AI systems and describes how these systems need to be responsive to multiple criteria in order to address AI risks [22]. NIST AI RMF states that trustworthy AI is *safe*, *secure and resilient*, *explainable and interpretable*, *privacy enhanced*, *fair (with harmful bias managed)*, *accountable and transparent*, *valid and reliable* [22].

The NIST framework provides guidance for addressing risks in the design, development, use, and evaluation of AI systems [25] to ensure their trustworthiness. However, this paper identifies the need for further risk analysis to facilitate the widespread adoption of the NIST framework by organizations.

In order to complement existing systems models of *purpose*, *structure*, and *function*, there is a need for system modeling that focuses on evolving priority orders of complex systems. These priority orders encompass various elements such as assets, policies, investments, organizational units, locations, personnel, and more. Technological advancements, environmental factors, missions, obsolescence, regulations, behaviors, markets, human migrations, conflicts, and other influences disrupt these priority orders.

This paper develops a risk analysis of artificial intelligence in medicine with a multilayer concept of system order using a principled methodology to account for the scenarios that are most and least disruptive to these orders.

Figure 1 shows that in system modeling, the defining characteristic layers of any system *are purpose* (in some of the literature, *purpose* is also referred to as behavior [26–29]) *(π, Pi)*, *structure* (on some of the literature, *structure* is also referred to as elements or components [26–29]) *(σ, Sig)*, *function* (on some of the literature, *function* is also referred to as process or operations [26–29]) *(ϕ, Phi)*, *interconnections (ι, Iot)*, *environment (ε, Eps)* and *boundary (β, Bet)* [26–29]. The Greek alphabet is employed to facilitate fluent reading and enhance annotations throughout the paper. Other studies may find additional layers for the AI risk management analysis. The scope of the paper is limited to the *purpose (Pi)*, *structure (Sig)*, *and function (Phi)* characteristic layers. The *purpose (Pi)* layer examines the goals and objectives of the system. The *structure (Sig)* layer examines the components of the system. The *function (Phi)* layer focuses on the specific tasks and processes that the system performs [30].

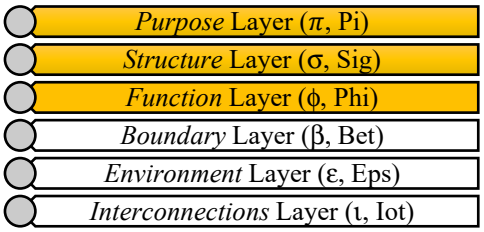

Figure 1. Six layers of system characteristics that can be used in risk analysis of AI in healthcare applications. Orange cells indicate the scope of this paper.

A risk assessment of AI tools is a major challenge, especially as the most recent generation of AI tools has extremely broad applicability. That is, the design and use cases for AI are constantly evolving. Three main scenario-based preference models are developed for three healthcare system: 1. Healthcare centers or clinics as a higher-level systems (*purpose (Pi)* layer). 2. Medical implants or devices (*structure (Sig)* layer). 3. Disease diagnosis, more specifically the diagnosis of cardiac sarcoidosis disease (*function (Phi)* layer). Trustworthiness in the context of AI in healthcare should be considered for various stakeholders, including AI developers, healthcare clinicians, and patients. This is distributed across three primary layers: insider, internal, and external layers, respectively. The scope of this study is focused on internal trustworthiness, addressing the relationship between AI providers and AI users. The AI users within the healthcare context are categorized across these three layers:

- ***Purpose (Pi) layer:*** This layer focuses on the objectives and overall goal of the system and includes the strategic and operational objectives of the systems. This includes domain experts in healthcare, such as health center board members and clinicians responsible for the operation of a clinic section.
- ***Structure (Sig) layer:*** This layer includes the physical framework of the system, which could resemble physical medical devices. These are the device developers and designers involved in the implementation of AI in healthcare.
- ***Function (Phi) layer:*** This layer includes a specific operation or a task defined and performed by medical professionals, such as disease diagnosis. These are physicians specializing in radiology and cardiology, contributing to the functional aspects of AI applications in healthcare.
- *Interconnections (Iot) layer:* This layer shows the interactions and connectivity of medical components together.
- *Environment (Eps) layer:* This layer includes any external factors or environments that could affect the medical system outside its boundary.
- *Boundary (Bet) layer:* This layer defines the limits of the medical system and the system's scope. This layer distinguishes the medical system from its external environment (Eps).

The innovation comprises three aspects. The contribution to "theory and philosophy" is the introduction of systems organized in layers, utilizing a multi-layer system approach to account for disruptive scenarios and the disruption of system order [31]. This innovation acknowledges and addresses risks and disruptions occurring across multiple layers. The innovation to "methods" involves offering detailed rubrics to elaborate on and execute the steps within the risk register [32]. This paper contributes to the "application" domain by applying layer disruption scenario analysis, specifically in healthcare and medicine applications. This paper develops a multi-layer scenario-based [33,34] preference risk register for deploying AI in complex engineering systems, building on top of NIST AI RMF aspects.

Initiatives, success criteria, emergent conditions, and scenarios are introduced for each layer as the main components of the risk analysis. [31]. One challenge when integrating AI-based decision-making tools into medicine is the ability to generalize effectively when applied across various sectors with diverse patient populations across varied initiatives and disruptive scenarios. The framework contributes to systems engineering by addressing

various research gaps in the System Engineering Body of Knowledge (SEBoK) related to AI risk management [28]. This work shows how responsible AI could benefit a variety of engineering systems and reduce the risks in the systems. The framework guides and shapes the AI R&D portfolio by highlighting the most and least disruptive scenarios to the enterprise and monitoring and evaluating the trustworthiness of the AI implemented in the system. Practitioners will better understand how to implement AI to enhance object designs and mitigate AI risk applications and uncertainties, as well as the general topic of what methods systems can employ to set precise boundaries for AI activities and how to establish ethical, legal, societal, and technological boundaries for AI activity by quantifying risk as the disruption of the order of AI initiatives in healthcare systems.

2. Materials and Methods

This section describes an elicitation of scenario-based preferences [35,36] that aids in identifying system initiatives, criteria, emergent conditions, and scenarios. Figure 2 describes the conceptual diagram of the risk assessment methodology, and Figure 3 describes the conceptual diagram of systems modeling for enterprise risk management of AI in healthcare. The figure describes the following four steps:

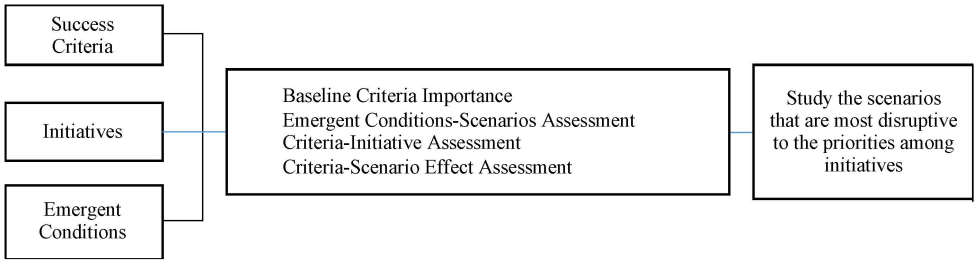

Figure 2. Conceptual diagram of risk assessment of AI in healthcare applications.

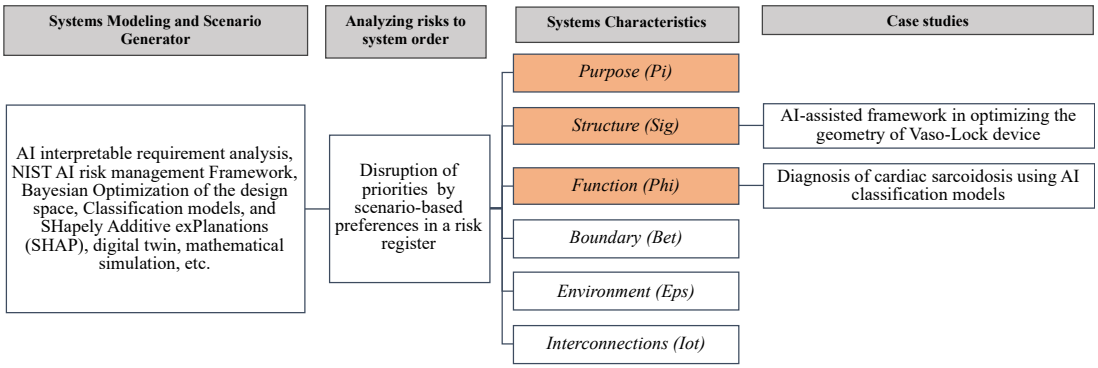

Figure 3. The proposed conceptual diagram of system modeling for enterprise risk management of AI in healthcare.

1. System modeling and scenario generator, which could include techniques customized for each case study, such as Shapley additive value, digital twins, eXplainable AI (XAI) techniques, etc.
2. The multicriteria decision analysis (MCDA) risk resister tool is used to analyze risks according to the system order.
3. The three system characteristics reviewed in this paper are *Purpose (Pi)*, *Structure (Sig)*, *and Function (Phi)* layers.
4. Case studies.

Each step will be explained in detail in the following sections.

The first step of the framework develops success criteria to measure the performance of investment initiatives based on the system objectives. Success criteria are mainly derived from research of technological analyses, literature reviews, and expert opinions describing

the goals of the system. Any changes in success criteria affect expectations of success and represent the values of the stakeholder. The set of success criteria is defined as {*c.01*, *c.02*, . . ., c.m}.

As this framework is based on the NIST AI RMF, the success criteria for all three layers—AI trustworthy in healthcare systems or *purpose (Pi)*, AI trustworthy in medical implants/devices or *structure (Sig)*, and AI trustworthy in disease diagnosis or *function (Phi)*—are established using the seven aspects of trustworthy AI systems. By leveraging this foundation, the framework ensures comprehensive risk analysis by considering the criteria of trustworthiness across the different AI in healthcare application areas. Table 1 shows the seven aspects of the NIST AI RMF: *c.01—safe*; *c.02—secure and resilient*; *c.03—explainable and interpretable*; *c.04—privacy enhanced*; *c.05—fair (with harmful bias managed)*; *c.06—accountable and transparent*; *c.07—valid and reliable.*

Table 1. Success criteria for the *purpose (Pi)*, *structure (Sig)*, and *function (Phi)* layers in risk analysis of AI in healthcare, medical devices, and disease diagnosis. Success criteria are adapted from NIST AI risk management framework [22].

Index	Criterion
c.01	Safe
c.02	Secure and resilient
c.03	Explainable and interpretable
c.04	Privacy enhanced
c.05	Fair—with harmful bias managed
c.06	Accountable and transparent
c.07	Valid and reliable
c.i	Others

Initiatives are the second element of the model, and they represent a set of decision-making alternatives. These can take the form of technologies, policies, assets, projects, or other investments [6]. Initiatives are represented by the set {x.01, x.02, . . ., x.n}. Initiatives are identified by elicitation from stakeholders and experts to determine what components, actions, assets, organizational units, policies, locations, and/or allocations of resources constitute the system [31].

The third element, emergent conditions, are events, trends, or other factors impacting decision-maker priorities in future planning contexts. Karvetski and Lambert [37,38] identify "emergent and future conditions" as individual trends or events that can impact decision-making and strategy in some way. These conditions are combined to create unique scenarios. Uncertainties in emergent conditions are a significant contributor to project failure and impact the ability of the system to meet success criteria. The set of emergent conditions is {e.01, e.02, . . ., e.k}. In the model, emergent conditions influence the relevance weights of individual success criteria.

The baseline relevance of criteria is established by interviewing stakeholders, and they are scored *low*, *medium*, and *high*. Based on this determination, the baseline weights are assigned to each of the success criteria.

Scenarios comprise one or more emergent conditions. The set of scenarios is defined as {*s.01*, *s.02*, . . ., s.p}. Scenarios are potential events that may disrupt priority orders. It is important to clarify that scenarios do not serve as predictions for future conditions and do not include any indication of the likelihood of occurrence. Instead, scenarios function as projections, designed to investigate the impacts of potential future states. Additionally, emergent conditions and scenarios do not aim to catalog every conceivable future state or disruption. Instead, they concentrate on addressing the specific concerns of system owners and experts, such as those in the medical field, that have been introduced earlier in the analysis [31].

Experts in three layers were engaged in the process of identifying success criteria, initiatives, emergent conditions, scenarios, criteria-initiative assessment, criteria-scenario

relevance, and baseline relevance. The experts for the *purpose (Pi)* layer are the board members of Binagostar eye surgical hospital. Three interviews were conducted with the board members of Binagostar Eye Surgical Hospital.

In the criteria-initiative assessment, experts and stakeholders were asked to what degree they agree that "initiative x.i addresses criterion c.j". *Neutral* entries are represented by a dash (-); *somewhat agree* is represented by an unfilled circle (○); *agree* is represented by a half-filled circle (◐); and *strongly agree* is represented by a filled circle (●) in the matrix with the set of numerical weights of {0, 0.334, 0.667, 1}, respectively.

The qualitative results of the project constraint matrix can be converted into numerical weights [39,40] following a rank-sum weighting method [41] based on Equation (1):

$$w_j = \frac{m - rankj + 1}{\sum_{j=1}^{m} \mathrm{m} - \mathrm{rankj} + 1} \quad (1)$$

where w_j is the weight of the j-th criterion, m is the total number of criteria, and $rank_j$ is the ordinal rank of the j-th criterion [37].

The effect of disruptive emergent conditions is operationalized through a change in the criteria weights. For each scenario, the user is asked to assess to what degree the relative importance of each criterion change given the scenario will occur [42]. Responses include *decreased (D)*, *decreased somewhat (DS)*, no change, *increased somewhat (IS)*, and *increased (I)*. These changes are recorded in the *W* matrix. In Equation (2), α is a scaling constant that is equal to {8, 6, 1, 1/6, 1/8} for *increases*, *increases somewhat*, *no change*, *decreases somewhat*, *and decreases*, respectively. The scaling constant is intended to be consistent with the swing weighting rationale. The swing weight technique accommodates adjustments for the additional scenarios. The procedure for deriving weights for an additive value function using the swing weight method is thoroughly documented in the MCDA literature, as evidenced by works such as those by Keeney and Raiffa (1979) [40], Keeney (1992) [43], Belton and Stewart (2002) [44], and Clemen and Reilly (2001) [45]. The justification for swing weighting is explained by Karvetski and Lambert as follows: α serves as a value multiplier, adjusting the trade-off between exchanging a high level of performance for a low level of performance in one criterion and an exchange of a low level of performance for a high level of performance in another criterion [37,38]. The swing weight technique was adopted to derive the baseline criteria weights (w_j), as well as the adjusted weights for each scenario [38].

$$W_{jk} = \alpha * W_j \quad (2)$$

The initiatives are prioritized with a linear additive value function, defined in Equation (3). $v_j(x.i)$ is the partial value function of initiative x.i along with criterion c.j, which is defined using the criteria-initiative (C-I) assessment. *V* is a matrix that contains the relative importance scores for each initiative across each scenario, and $v_k(x.i)$ is the change for initiatives across each scenario.

$$V_k(x.i) = \sum_{j=1}^{m} w_{jk} v_j(x.i) \quad (3)$$

The disruptiveness score is defined based on the sum of the squared differences between the baseline rank and the disrupted rank of each initiative for each scenario. The disruptiveness score is used to understand the effect of emergent conditions on the prioritization of initiatives. Equation (4) shows the disruptiveness score for scenario s.k.

$$D_k = \sum_i (r_{i0} - r_{ik})^2 \quad (4)$$

r_{ik} is the rank of initiative x.i under scenario s.k and r_{i0} is the rank of the initiative x.i under the baseline scenario (*s.00*) [46]. Then, the scores are normalized to be in the scale of 0–100.

This paper shows the proposed theory and method of system modeling for enterprise risk management of AI in healthcare. The method comprises four steps, including the following: 1. System modeling and scenario generator. 2. Analyzing risks to system order. 3. System characteristics. 4. Case studies.

In the next section, the method is demonstrated in three layers: *purpose (Pi)*, *structure (Sig)*, and *function (Phi)*.

3. Demonstration of the Methodology

Experts from various medical specialties participated in the study, providing insights through interviews throughout the process. Their involvement encompassed activities such as identifying initiatives, addressing emergent conditions, considering various scenarios, and conducting scoring/ranking assessments.

The following section describes demonstrations of the methodology across the three layers.

3.1. Trustworthy AI in Healthcare System (Purpose (Pi) Layer)

The mathematical decision framework is employed to assess the trustworthiness of AI in the healthcare *purpose (Pi)* layer. This layer focuses on the goals of the system and objectives, specifically emphasizing the internal trustworthiness that AI providers must address for healthcare AI users, such as clinicians utilizing AI in hospital or healthcare institute/clinic operations.

Tables 1–4 describe 7 success criteria, 43 initiatives, 25 emergent conditions, and 10 scenarios, respectively, for the risk management of AI in healthcare systems [22,31,47–49].

Table 2. Initiatives for the *purpose (Pi)* layer in risk analysis of AI in healthcare. Abridged from various sources that are identified in the narrative.

Index	Initiative
x.Pi.01	Identify At-Risk Components
x.Pi.02	Understanding ML Tools to Uncover Any Patterns in Data
x.Pi.03	Record-Keeping, Reserving, and Storing
x.Pi.04	Data Governance and Management
x.Pi.05	Data Traceability of the Process
x.Pi.06	Clear and Plain Language
x.Pi.07	Concise, Transparent, Easily Accessible Form, and Process
x.Pi.08	Human–AI Collaboration and Consulting
x.Pi.09	Accurate, Appropriate, Clear, and Accessible Information
x.Pi.10	Providing of Information/Documents
x.Pi.11	Identify Roles and Responsibilities of Humans in the AI Loop
x.Pi.12	Safety and Quality of AI in its Lifecycle
x.Pi.13	Making Informed Decisions Such as Individual Rights for Patient Point of View
x.Pi.14	Guaranteeing Quality and Safety
x.Pi.15	Continuous Collecting and Verification of Data
x.Pi.16	Before and After the Event Control Over the Outcomes of AI
x.Pi.17	Outcome Assessment Through Explanations and Record the Development and Validations of AI
x.Pi.18	Interpretation for a Prediction its Cause of Error
x.Pi.19	Comprehension of AI-Based Devices and any Decisions They Made
x.Pi.20	Inform and Train Clinicians on How the Use AI, When to Use AI, and Ways to Validate the Generated Results
x.Pi.21	Consider AI Safety Risks, its Regulation, and the Legislation
x.Pi.22	Tries to Minimize Risks to the Maximum Possible Extent
x.Pi.23	Clinicians to be Convinced that Specific AI System Outcomes are Safe
x.Pi.24	Convincing Clinicians on How an AI Device is Generally Useful and Safe
x.Pi.25	Clinicians to have all the Necessary Training and Information by AI Developers
x.Pi.26	Perform Internal Transparency of AI
x.Pi.27	Safety and Quality of AI Devices in the Market
x.Pi.28	Identify any Residual Risks, any Contra Indications, and Any Side Effects by Using AI
x.Pi.29	To Provide any Necessary Specifications to the Users for Proper Performance of the Device

Table 2. *Cont.*

Index	Initiative
x.Pi.30	To Provide any Necessary Training, Facilities, and Qualifications to the Users of the Device
x.Pi.31	Policy-Makers to Ensure the Internal Level Transparency and Their Opacity and Self-Learning
x.Pi.32	Data Governance and Management Practices Shall be Developed by AI Providers
x.Pi.33	Users to be Informed of What Data to Use for Training, Validating, and Testing the AI Models; Also, any Potential Changes Due to Various Input Data
x.Pi.34	The Necessary Information About the Risks of the Device, and its Side Effects, As Well As the Explainability Limitations
x.Pi.35	Inform Users on Why and How the Benefits of the Use of an AI System Overweigh its Risks Compared to Other Technologies on the Market
x.Pi.36	Automatic Explanations Generated into AI Systems
x.Pi.37	Evaluation of Interpretability by Involving Human Experiments
x.Pi.38	Healthcare Professionals to Assess the Quality of AI Explanations by the AI Provider
x.Pi.39	Healthcare Professional Independent Bodies in the AI-Designed Device Evaluation
x.Pi.40	Accepting Some Degree of Opacity of the AI Systems Over its Risks
x.Pi.41	Providing Quality Records
x.Pi.42	The Requirement of Explainability Techniques as a Part of the Conformity Assessment Process
x.Pi.43	AI Providers to Provide Some Level of Opacity of the AI System
x.Pi.i	Others

Table 3. Emergent conditions used to create sets of scenarios for the *purpose (Pi)* layer in the risk analysis of AI in healthcare. Abridged from various sources that are identified in the narrative.

Index	Emergent Condition
e.Pi.01	Lack of Algorithmic Transparency
e.Pi.02	Low Quality of the Inputs and the Procedures to Verify it
e.Pi.03	Concerns Over any Trade Off Between AI's Performance and Explainability
e.Pi.04	Impossible Reaching Zero Risks in AI Area
e.Pi.05	Lack of Full Predictability of AI Applications
e.Pi.06	Concerns About Information Provision
e.Pi.07	Lack of AI Models Insider Transparency
e.Pi.08	Some AI Models are Opaque with Lack of Explainability
e.Pi.09	The Availability for Explanations and the Quality of the Data in Training Process
e.Pi.10	Which Automated Explanations Techniques Available
e.Pi.11	Legislative Requirements
e.Pi.12	Limitations of AI Technologies Usage in Healthcare
e.Pi.13	No Existing Techniques Yet for Algorithmic Opacity
e.Pi.14	Some Level of Limitations in Accurately Predicting the Outcomes of Medical Diagnosis and Treatment
e.Pi.15	Limitations in Explaining Why a Patient Treatment Did Not Help
e.Pi.16	Shortage of AI in Cognitive Empathy
e.Pi.17	Hard to Track and Measuring Emergent Risks by Organizations
e.Pi.18	Security Concerns Related to the Confidentiality of the System Training and Output
e.Pi.19	One-Size-Fits-All Requirements AI Model Challenges
e.Pi.20	Unexpected Changes in the Environment or Use
e.Pi.21	Data Poisoning
e.Pi.22	Privacy Intrusions
e.Pi.23	Lack of Access to the Ground Truth in the Dataset
e.Pi.24	Intentional or Unintentional Changes During Training
e.Pi.25	Cyber Attacks
e.Pi.i	Others

Table 4. Scenarios for the *purpose (Pi)* layer in the risk analysis of AI in healthcare showing which emergent conditions fit in each scenario. Abridges from various sources that are identified in the narrative.

	***s.01*—Funding Decrease**	***s.02*—Government Regulation and Policy Changes**	***s.03*—Privacy Attacks**	***s.04*—Cyber Security Threats**	***s.05*—Changes in AI RMF**	***s.06*—Non-Interpretable AI and Lack of Human–AI Communications**	***s.07*—Global Economic and Societal Crisis**	***s.08*—Human Errors in Design, Development, Measurement, and Implementation**	***s.09*—Uncontrollable Environment**	***s.10*—Expensive Design Process**
e.Pi.01		✓			✓	✓		✓		
e.Pi.02	✓							✓	✓	
e.Pi.03	✓									
e.Pi.04						✓		✓		✓
e.Pi.05						✓		✓		
e.Pi.06		✓		✓						
e.Pi.07						✓				
e.Pi.08						✓				
e.Pi.09										
e.Pi.10	✓					✓		✓		
e.Pi.11	✓	✓			✓		✓			
e.Pi.12	✓	✓			✓		✓			
e.Pi.13						✓				
e.Pi.14					✓			✓	✓	
e.Pi.15						✓			✓	
e.Pi.16			✓			✓				
e.Pi.17	✓		✓			✓		✓	✓	✓
e.Pi.18	✓		✓				✓		✓	
e.Pi.19						✓			✓	
e.Pi.20	✓	✓			✓		✓		✓	
e.Pi.21								✓		
e.Pi.22			✓	✓						
e.Pi.23	✓							✓	✓	✓
e.Pi.24								✓	✓	
e.Pi.25				✓						

Table 5 depicts the foundational significance of success criteria in ensuring the trustworthiness of AI within healthcare systems. Significance is exemplified through assigned weights, indicating the relative importance of each criterion in comparison to others. In the initial phase of the criteria analysis, a classification of *low*, *medium*, or *high* relevance is assigned. This is achieved by assigning numerical values of one, two, and four, respectively, to each success criteria. These relevance classifications are based on weights determined by inputs from experts and stakeholders. For example, in the baseline scenario, criterion *c.01*, *safe*, holds *high* relevance in comparison to other criteria. It is important to note that while scenarios do not alter the rating or scoring assessments, they do influence how decision-makers shape their preferences across these criteria [38].

Table 5. Baseline relevance for the *purpose (Pi)* layer in the risk analysis of AI in healthcare.

The Criterion c.xx Has	s.00—Baseline	Relevance among the Other Criteria
c.01—safe has	high	relevance
c.02—secure and resilient has	high	relevance
c.03—explainable and interpretable has	high	relevance
c.04—privacy enhanced has	high	relevance
c.05—fair—with harmful bias managed has	high	relevance
c.06—accountable and transparent has	high	relevance
c.07—valid and reliable has	high	relevance

Table 6 describes the impact of the seven success criteria on the forty-three initiatives that are introduced above. No impact means the criterion is not relevant to the initiative. As mentioned in the Method section, in the criteria-initiative assessment, experts and stakeholders were asked to what degree they agree that initiative x.i address criterion c.j. Neutral entries are represented by a dash (-); somewhat agree is represented by an unfilled circle (○); agree is represented by a half-filled circle (◐); and strongly agree is represented by a filled circle (●) in the matrix with the set of numerical weights of {0, 0.334, 0.667, 1}, respectively. For instance, in Table 6, the stakeholders mentioned that initiative x.Pi.01 addresses criterion *c.02* by (◐) degree, which represents the weight of 0.667. This weight was defined in the C-I assessment as $v_{02}(x.Pi.01)$.

Table 6. The criteria-initiative assessment shows how well each initiative addresses the success criteria for the *purpose (Pi)* layer in the risk analysis of AI in healthcare. Strongly agree is represented by a filled circle (●); agree is represented by a half-filled circle (◐); somewhat agree is represented by an unfilled circle (○); and neutral is represented by a dash (—).

	c.01	***c.02***	***c.03***	***c.04***	***c.05***	***c.06***	***c.07***
x.Pi.01	●	◐	○	○	◐	○	○
x.Pi.02	—	—	●	◐	○	◐	◐
x.Pi.03	—	—	○	●	—	◐	—
x.Pi.04	○	○	○	●	—	◐	—
x.Pi.05	●	○	○	◐	—	●	—
x.Pi.06	—	—	●	—	—	●	—
x.Pi.07	—	—	●	◐	—	●	—
x.Pi.08	—	—	●	—	—	●	◐
x.Pi.09	○	○	●	●	—	◐	—
x.Pi.10	○	○	○	●	—	◐	—
x.Pi.11	○	○	◐	◐	◐	◐	◐
x.Pi.12	●	●	—	—	○	◐	●
x.Pi.13	—	—	●	○	—	●	○
x.Pi.14	●	●	○	—	○	◐	●
x.Pi.15	○	◐	●	●	○	◐	○
x.Pi.16	◐	◐	◐	—	◐	◐	●
x.Pi.17	◐	◐	◐	—	◐	◐	●
x.Pi.18	◐	◐	●	—	—	●	●
x.Pi.19	●	●	●	○	●	●	●
x.Pi.20	●	●	●	◐	●	●	●
x.Pi.21	◐	◐	◐	◐	◐	◐	○
x.Pi.22	●	●	◐	◐	—	◐	◐
x.Pi.23	●	●	●	●	●	●	●
x.Pi.24	●	●	●	●	●	●	●
x.Pi.25	●	●	●	●	●	●	●

Table 6. *Cont.*

	c.01	***c.02***	***c.03***	***c.04***	***c.05***	***c.06***	***c.07***
x.Pi.26	○	◐	●	●	●	●	●
x.Pi.27	●	●	●	●	●	●	●
x.Pi.28	○	○	●	○	○	●	○
x.Pi.29	●	●	●	●	●	●	●
x.Pi.30	◐	◐	●	◐	◐	●	◐
x.Pi.31	◐	◐	●	◐	◐	●	◐
x.Pi.32	○	○	○	●	—	◐	—
x.Pi.33	●	●	●	●	●	●	●
x.Pi.34	●	●	●	●	●	●	●
x.Pi.35	●	●	●	●	●	●	●
x.Pi.36	◐	◐	●	◐	◐	●	●
x.Pi.37	—	—	●	—	◐	●	◐
x.Pi.38	◐	◐	●	○	◐	●	◐
x.Pi.39	—	—	●	○	●	◐	○
x.Pi.40	◐	◐	●	○	○	●	◐
x.Pi.41	●	●	○	—	○	◐	●
x.Pi.42	●	●	●	◐	◐	●	●
x.Pi.43	◐	◐	●	◐	◐	●	●

Table 7 shows the criteria-scenario relevance. For instance, the criterion *c.01*, *safe*, *decreases somewhat* under scenario *s.01*, *funding decrease*, and *decreases* under *s.04*, *cyber-attacks on active system.*

Table 7. The criteria-scenario relevance shows how well each scenario fits the success criterion for the *purpose (Pi)* layer in the risk analysis of AI in healthcare. Decrease somewhat = DS; decrease = D; somewhat increase = SI; increase = I.

	s.01	***s.02***	***s.03***	***s.04***	***s.05***	***s.06***	***s.07***	***s.08***	***s.09***	***s.10***
c.01	DS	SI	-	D	SI	DS	DS	DS	DS	DS
c.02	-	SI	-	D	SI	DS	DS	DS	DS	DS
c.03	-	SI	-	-	SI	DS	DS	DS	D	-
c.04	-	SI	D	D	SI	-	DS	-	-	-
c.05	DS	SI	-	-	SI	DS	DS	DS	-	-
c.06	DS	SI	-	-	SI	DS	DS	-	DS	DS
c.07	DS	SI	-	-	SI	DS	DS	DS	DS	DS

Figure 4 provides a disruptive score for the scenarios. This is based on the sum of squared differences in priority of initiatives relative to the baseline scenario. A higher score suggests a greater potential issue or challenge posed by the scenario for the system under consideration. This figure shows that *s.06—non-interpretable AI and lack of human–AI communications*; *s.08—human errors in design, development, measurement, and implementation*; *s.09—uncontrollable environment*; *and s.10—expensive design process* have the highest disruption among the scenarios.

Figure 5 shows the variation in the prioritization of initiatives across the scenarios. The black bar shows the baseline ranking of each initiative. The blue bars show how the initiatives are promoted in priority, and the red bars highlight how the initiatives are demoted in priority. The bar indicates each ranking range of initiative subject to disruptions by scenarios [31]. The most important initiatives in this figure are *x.Pi.35—inform users on why and how the benefits of the use of AI system overweigh its risks compared to other technologies on the market*; *x.Pi.33—users to be informed on what data to use for training, validating, and testing the AI models*; *Also, any potential changes due to various input data*; *and x.Pi.23—clinicians to be convinced that specific AI system outcomes are safe.*

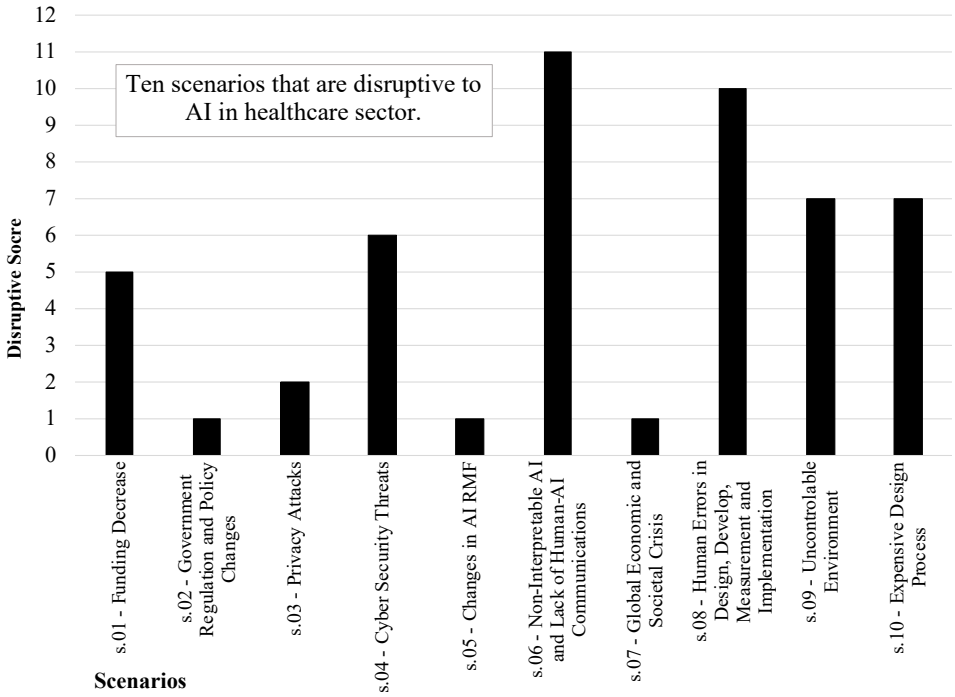

Figure 4. Disruptive score of scenarios is based on the sum of squared differences in priority of initiatives, relative to the baseline scenario for the *purpose (Phi)* layer in risk analysis of AI in healthcare.

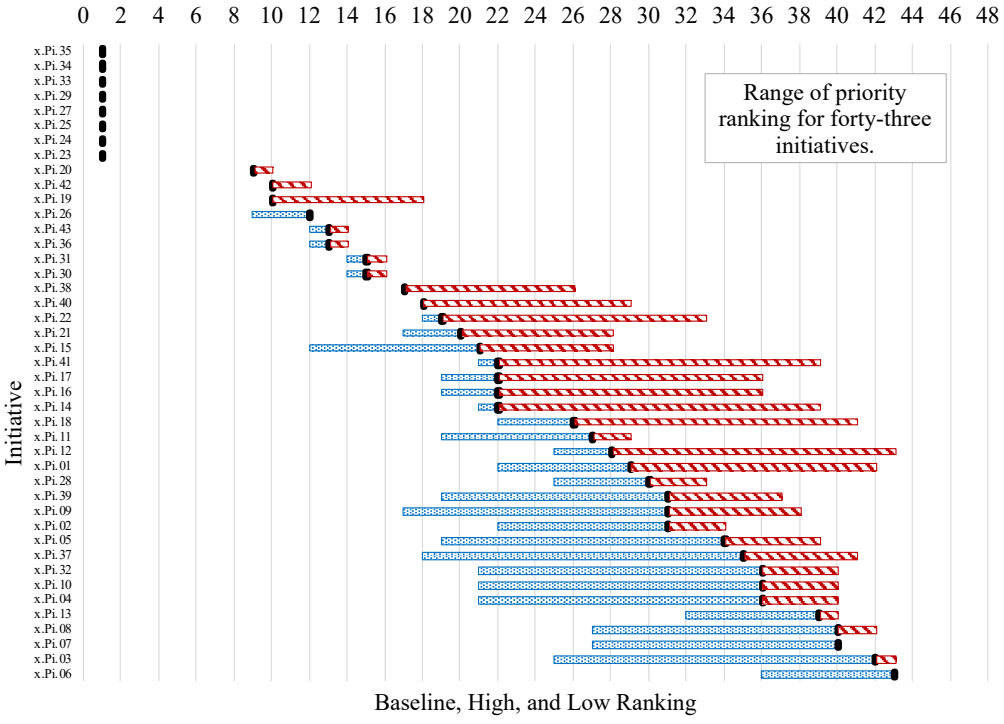

Figure 5. Distributions of initiatives influence rankings are based on which emergent conditions that could arise more often or never occur for the *purpose (Phi)* layer in risk analysis of AI in healthcare. Blue bar means promotion in ranking and red bar means demotion in ranking.

3.2. Trustworthy AI in Medical Devices Design (Structure (Sig) Layer)

Table 1 outlines the success criteria for the trustworthiness of AI in medical implants/devices. This is based on the seven aspects outlined in the NIST AI RMF as the essential success criteria for evaluating the systems. Therefore, the success criteria utilized in this analysis remain consistent throughout the analyses.

Tables 8–10 describe 47 initiatives, 50 emergent conditions, and the same 10 scenarios for the risk management of AI trustworthiness in medical implants/devices [22,31,50–52].

Table 8. Initiatives for the *structure (Sig)* layer in risk analysis of AI in healthcare. Abridged from various sources that are identified in the narrative.

Index	Initiative
x.Sig.01	Identify At-Risk Components
x.Sig.02	Understanding ML Tools to Uncover Any Patterns in Data
x.Sig.03	Maintaining the Provenance of Training Data
x.Sig.04	Safety/Verifiability of Automated Analyses
x.Sig.05	Supporting Attribution of the AI System's Decisions to Subsets of Training Data
x.Sig.06	Correctly Labeling the Data
x.Sig.07	Training Data to Follow Application Intellectual Property Rights Laws
x.Sig.08	Find the Maximum value of the Max Force of the Device
x.Sig.09	Maintain Organizational Practices Like Implement Risk Management to Reduce Harm Reduction and More Accountable Systems
x.Sig.10	Prioritization Policies and Resources Based on Assesses Risk Levels
x.Sig.11	Safety of Personally Identifiable Information
x.Sig.12	Effective Risk Management by Appropriate Accountability Mechanism, Roles, and Responsibilities, and Incentive Structures for Risk Management to be Effective
x.Sig.13	Identify the Right AI RMF in Different Contexts Based on Capabilities, Resources, and Organization Size
x.Sig.14	Identify AI Actors with Diversity in Experience, Expertise, Background, Demographically, and Disciplinary
x.Sig.15	Assist in Providing Context as Well as Understanding Potential and Actual Impacts
x.Sig.16	Identify a Source of Formal, and Guidance for AI Risk Management
x.Sig.17	Designate Ethical, Legal, Societal, and Technical Boundaries for AI Operation
x.Sig.18	Trade Offs Needed Discussions to Balance Societal Values and Priorities Related to Civil Liberties and Rights, Equity, the Environment and the Planet, and the Economy
x.Sig.19	Articulate and Document the Concept and Objectives of the System Considering Legal, Regulatory, and Ethical Requirements
x.Sig.20	Gather, Clean, and Validate Data and Document the Metadata and Characteristics of the Dataset Considering Legal, Regulatory, and Ethical Requirements
x.Sig.21	Key steps for implementing a new software system: Pilot, Compatibility with Legacy Systems, Regulatory Compliance, Organizational Change Management, and User Experience Evaluation
x.Sig.22	Continuously Assess AI System's Recommendations and Impacts
x.Sig.23	Balancing and Trade Off of Trustworthy AI System Characteristics Based on Context
x.Sig.24	Reduce the Number of Experiments to be Cost- and Time-Effective by Optimizing the Configurations
x.Sig.25	Ability of an AI System to Perform as Required without Failure
x.Sig.26	Confirmation, Through the Provision of Objective Evidence that the Requirements for a Specific Intended Use Have been Fulfilled
x.Sig.27	Closeness of Results of Estimates, Observations, and Computations to the Ground Truth (True Values)
x.Sig.28	Human–AI Teaming
x.Sig.29	Demonstrate Validity or Generalizability Beyond the Training Conditions
x.Sig.30	System's Ability to Maintain its Performance Under Uncertain Circumstances
x.Sig.31	Minimizing Potential Harms to People Under Unexpected Operating Settings
x.Sig.32	Responsible AI System Design, Development, and Deployment Practices
x.Sig.33	Clear Information to the Users on Responsible Use of the AI System
x.Sig.34	Deployers and End Users to Make Responsible Decisions
x.Sig.35	Documentation and Explanation of Risks, Grounded in Empirical Evidence from Past Incidents
x.Sig.36	The Ability to Control, Adjust, or Involve Humans in Systems When They Do Not Perform as Intended or Expected
x.Sig.37	Resilient to Withstand Unexpected Adverse Events or Unexpected Environment or Use Changes
x.Sig.38	Preserve the Integrity and Functionality of Systems Amid Internal and External Changes, and Ensure Safe and Graceful Degradation When Required
x.Sig.39	Managing Risks from Lack of Explainability by Defining the AI System's Functions Considering Users' Role, Knowledge, and Skill Levels
x.Sig.40	The Ability to Describe Why an AI System Made a Specific Prediction or Recommendation

Table 8. *Cont.*

Index	Initiative
x.Sig.41	Securing Individual Privacy, Anonymity, and Confidentiality
x.Sig.42	The Process of Removing Identifying Information and Combining Specific Model Results to Maintain Privacy and Confidentiality in Certain Model Outputs
x.Sig.43	Strengthened Engagement with Relevant AI Actors and Interested Stakeholders
x.Sig.44	AI Systems May Need More Frequent Maintenance and Triggers for Corrective Maintenance Because of Data, Model, or Concept Drift
x.Sig.45	Clear and Distinct Definitions of Human Roles and Responsibilities Are Essential for Decision-Making and Oversight in the Context of AI Systems
x.Sig.46	Explain and Identify Most Important Features Using AI Models
x.Sig.47	Incorporates Processes to Assess Potential Impacts
x.Sig.i	Others

Table 9. Emergent conditions used to create sets of scenarios for the *structure (Sig)* layer in risk analysis of AI in healthcare. Abridged from various sources that are identified in the narrative.

Index	Emergent Condition
e.Sig.01	Systematic Biases in Collecting Clinical Data
e.Sig.02	Improperly Labeling the Data in Surgery-Specific Patient Registries
e.Sig.03	Issue of Incorrect Identification and Labeling of Variables in Registries Used for Surgery-Related Patient Data, Highlighting the Potential Consequences of Such Misidentification
e.Sig.04	Try and Validate Various Transparency Tools in Cooperation with AI Deployers
e.Sig.05	Artificial Intelligence Faces the Risk of Being Influenced by Unrealistic Expectations Propagated by the Media
e.Sig.06	Limitation in Types and Performance of Available Data
e.Sig.07	Expensive Data Collection
e.Sig.08	Time-Consuming Data Collection
e.Sig.09	Policy and Regulation Changes
e.Sig.10	Difficult and Complex AI Algorithms' Interpretability
e.Sig.11	Lack of AI Determination of Casual Relationships in Data at Clinical Implementation Level
e.Sig.12	Inability of AI in Providing an Automated Clinical Interpretation of its Analysis
e.Sig.13	Non-Intuitive Hidden Layers in DL
e.Sig.14	Abuse or Misuse of the AI Model or Data
e.Sig.15	Challenges with Training Data to be Subject to Copyright
e.Sig.16	Complicate Risk Measurement by Third-Party Software, Hardware, and Data
e.Sig.17	Hard to Track and Measuring Emergent Risks by Organizations
e.Sig.18	Lack of Robustness and Verifiable Methods for AI Trustworthiness
e.Sig.19	Misidentification of Different Risk Perspective in Early or Late Stages of AI Lifecycle
e.Sig.20	Difference Between Controlled Environment vs. Uncontrollable and Real-World Settings
e.Sig.21	Inscrutable Nature of AI Systems in Risk Measurements
e.Sig.22	Hard to Find Human Baseline for AI Systems Intended to Replace Human Activity
e.Sig.23	Risk Tolerance Influence by Legal or Regulatory Requirements Changes
e.Sig.24	Unrealistic Expectations About Risk to Misallocate Resources
e.Sig.25	Residual Risk after Risk Treatment Directly Impacts Healthcare Deployers
e.Sig.26	Privacy Concerns Regarding Using Underlying Data to Train the Systems
e.Sig.27	Energy and Environmental Implications from Resource-Heavy Computing Demands
e.Sig.28	Security Concerns Related to the Confidentiality of the System Training and Output
e.Sig.29	Security of the System Underlying Software and Hardware
e.Sig.30	One-Size-Fits-All Requirements AI Model Challenges
e.Sig.31	Neglecting the Trustworthy AI Characteristics
e.Sig.32	Difficult Decisions in Trade Off and Balancing Trustworthy AI Characteristics by Organizations
e.Sig.33	Subject Matter Experts Collaborate to Evaluate TEVV Findings, Aligning Parameters with Project Requirements and Deployment Conditions
e.Sig.34	Different Perception of the Trustworthy AI Characteristics Between AI Designer than the Deployer
e.Sig.35	Potential Risk of Serious Injury to the Patients
e.Sig.36	Unexpected Changes in the Environment or Use
e.Sig.37	Data Poisoning
e.Sig.38	Negative Risks Result from an Inability to Appropriately Understand or Contextualize System Output

Table 9. *Cont.*

Index	Emergent Condition
e.Sig.39	AI Allowing Inference to Identify Individuals or their Private Information
e.Sig.40	Privacy Intrusions
e.Sig.41	Data Sparsity
e.Sig.42	Fairness Perception Difference Among Cultures and Applications
e.Sig.43	Computational and Statistical Biases Stem from Systematic Errors Due to Limited and Non Representative Samples
e.Sig.44	Human Cognitive Biases Relate to How the Stakeholders Perceives AI System Information and Use it to Make Decisions
e.Sig.45	Lack of Access to the Ground Truth in the Dataset
e.Sig.46	Intentional or Unintentional Changes During Training
e.Sig.47	Increased Opacity and Concerns About Reproducibility
e.Sig.48	Impacts of Computational Costs on the Environment and Planet
e.Sig.49	Incapacity to Anticipate or Identify the Adverse Effects of AI-Driven Systems Beyond Statistical Metrics
e.Sig.50	Complexity of Explaining AI System to End Users
e.Sig.i	Others

Table 10. Emergent condition grouping for the *structure (Sig)* layer in risk analysis of AI in healthcare shows which emergent conditions fit in each scenario. Abridged from various sources that are identified in the narrative.

	***s.01*—Funding Decrease**	***s.02*—Government Regulation and Policy Changes**	***s.03*—Privacy Attacks**	***s.04*—Cyber Security Threats**	***s.05*—Changes in AI RMF**	***s.06*—Non-Interpretable AI and Lack of Human–AI Communications**	***s.07*—Global Economic and Societal Crisis**	***s.08*—Human Errors in Design, Development, Measurement, and Implementation**	***s.09*—Uncontrollable Environment**	***s.10*—Expensive Design Process**
e.Sig.01		✓								
e.Sig.02								✓		
e.Sig.03								✓		
e.Sig.04						✓				
e.Sig.05						✓				
e.Sig.06										✓
e.Sig.07	✓									✓
e.Sig.08										✓
e.Sig.09		✓			✓					
e.Sig.10						✓				
e.Sig.11						✓				
e.Sig.12						✓				
e.Sig.13						✓				
e.Sig.14						✓		✓		
e.Sig.15		✓	✓			✓				
e.Sig.16						✓				✓
e.Sig.17						✓			✓	
e.Sig.18		✓			✓	✓		✓	✓	✓

Table 10. *Cont.*

	s.01—Funding Decrease	*s.02*—Government Regulation and Policy Changes	*s.03*—Privacy Attacks	*s.04*—Cyber Security Threats	*s.05*—Changes in AI RMF	*s.06*—Non-Interpretable AI and Lack of Human–AI Communications	*s.07*—Global Economic and Societal Crisis	*s.08*—Human Errors in Design, Development, Measurement, and Implementation	*s.09*—Uncontrollable Environment	*s.10*—Expensive Design Process
e.Sig.19					✓	✓		✓		
e.Sig.20									✓	
e.Sig.21						✓				
e.Sig.22						✓				
e.Sig.23		✓								
e.Sig.24						✓		✓		
e.Sig.25						✓		✓	✓	
e.Sig.26			✓							
e.Sig.27		✓					✓			✓
e.Sig.28			✓	✓						
e.Sig.29				✓						
e.Sig.30						✓		✓	✓	✓
e.Sig.31						✓		✓		
e.Sig.32	✓				✓	✓				✓
e.Sig.33						✓				
e.Sig.34						✓				
e.Sig.35								✓	✓	
e.Sig.36									✓	
e.Sig.37			✓	✓				✓		
e.Sig.38						✓				
e.Sig.39			✓							
e.Sig.40			✓							
e.Sig.41								✓	✓	✓
e.Sig.42		✓			✓	✓				
e.Sig.43	✓					✓		✓	✓	✓
e.Sig.44	✓					✓				
e.Sig.45	✓							✓		✓
e.Sig.46					✓	✓		✓		
e.Sig.47						✓		✓		
e.Sig.48	✓						✓			✓
e.Sig.49						✓	✓		✓	
e.Sig.50						✓				

Table A1 illustrates the baseline relevance of success criteria for the trustworthy AI in medical implants and devices design. Criteria *c.01*, *safe*, and *c.02*, *secure and Resilient*, have medium relevance among the other criteria in the baseline scenario (See Appendix A).

Table A2 describes the impact of seven success criteria on forty-seven initiatives that are introduced above. *No impact* means that the criterion is not relevant to the initiative (See Appendix A). The experts for the *structure (Sig)* layer are research scientists and device designers from mechanical engineering department at Johns Hopkins University, MIT, and Western University of health sciences college of dental medicine, and the experts for the

function (Phi) layer are director members of cardiac radiology department at HDZ-NRW hospital in Germany. Bi-weekly meetings were held with experts from Johns Hopkins University and MIT. Additionally, five interview sessions were conducted with director members of the cardiac radiology department at HDZ-NRW hospital. Furthermore, seven interviews were carried out with a dentist at Western University of Health Sciences College of Dental Medicine.

Table A3 shows the criteria-scenario relevance. The criterion *c.01., safe*, effectiveness *decreases* under scenario *s.01, funding decrease*, and has *no change* under *s.04, cyber-attacks on active system* (See Appendix A).

Figure 6 provides a disruptive score of the scenarios based on the sum of squared differences in priority of initiatives, relative to the baseline scenario. A higher score suggests a greater potential issue or challenge posed by the scenario for the system under consideration. *s.06—non-interpretable AI and lack of human–AI communications*—has the highest disruption among other scenarios.

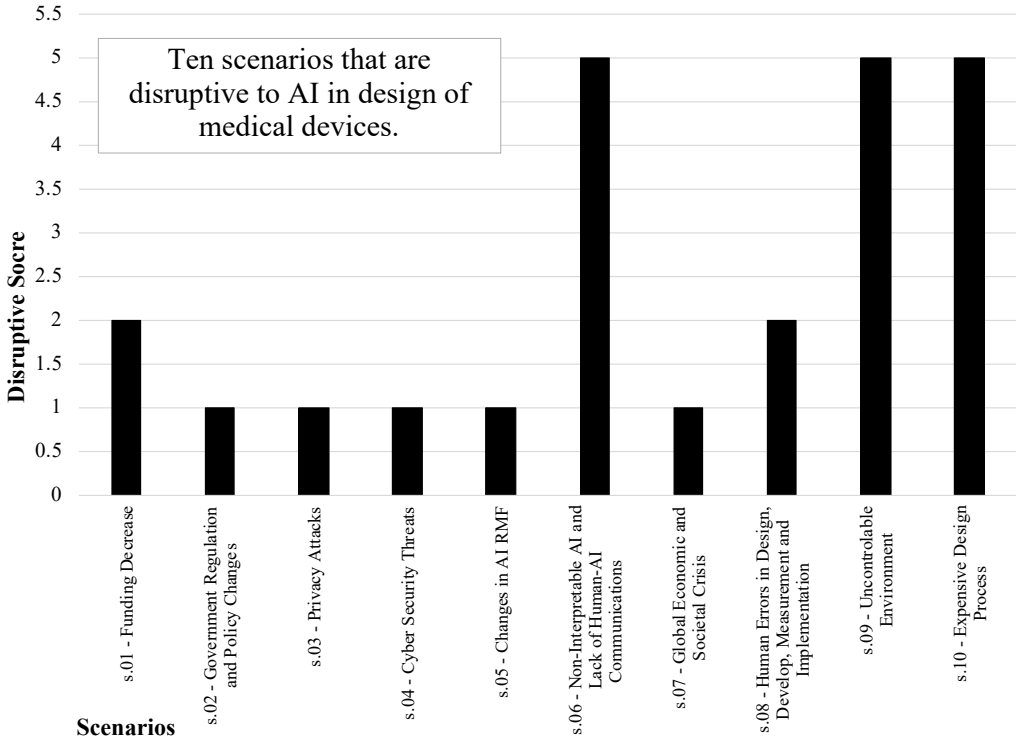

Figure 6. Disruptive score of scenarios is based on sum of squared differences in priority of initiatives, relative to the baseline scenario for the *structure (Sig)* layer in risk analysis of AI in healthcare.

Figure 7 shows the variation in the prioritization of initiatives across scenarios. The most important initiatives are *x.Sig.40—the ability to describe why an AI system made a specific prediction or recommendation*; *x.Sig.44—AI systems may need more frequent maintenance and triggers for corrective maintenance because of data, model, or concept drift*; and*x.Sig.24—reduce the number of experiments to be cost- and time-effective by optimizing the configurations*, and the most resilient initiatives are *x.Sig.39—managing risks from lack of explainability by defining the AI system's functions considering users' role, knowledge, and skill levels*; *x.Sig.33—clear information to the users on responsible use of the AI system*; *x.Sig.32—responsible AI system design, development, and deployment Practices*; *x.Sig.31—minimizing potential harms to people under unexpected operating settings*; *x.Sig.30—system's ability to maintain its performance under uncertain circumstances*; *x.Sig.26—confirmation, through the provision of objective evidence that the requirements for a specific intended use have been fulfilled*; *x.Sig.25—ability of an AI System to perform as required without failure*; *x.Sig.22—continuously assess AI System's recommendations and impacts.*

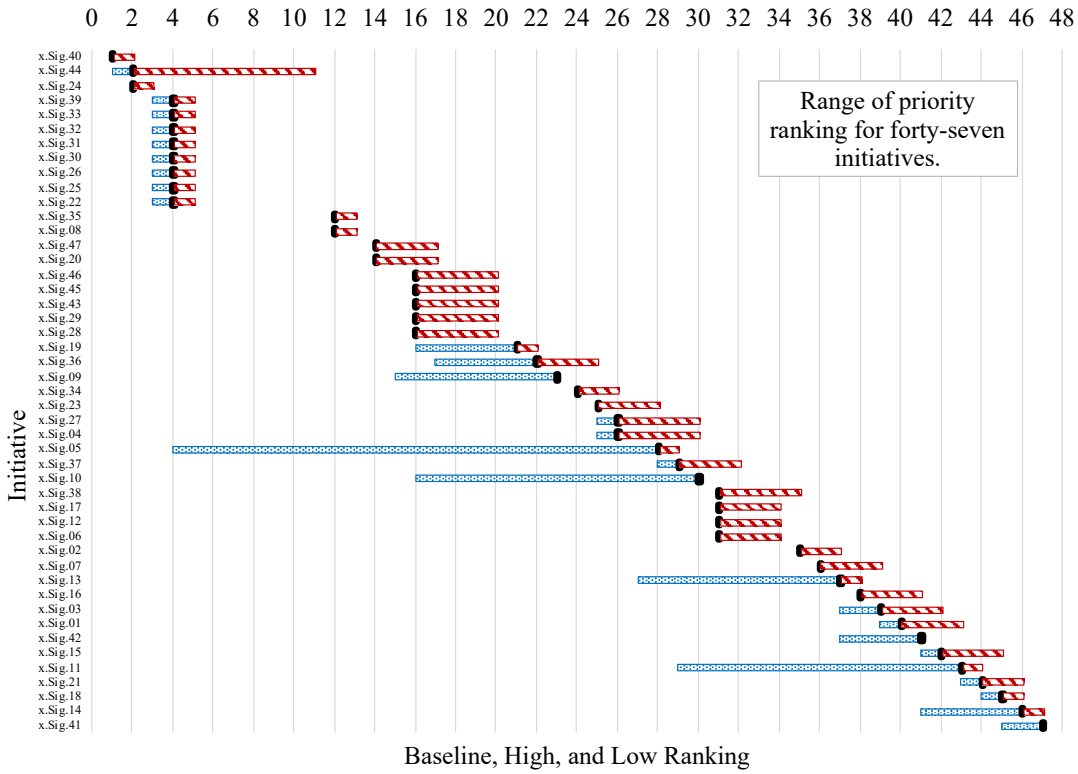

Figure 7. Distributions of initiatives rankings are based on which emergent conditions could arise more often or never occur for the *structure (Sig)* layer in risk analysis of AI in healthcare. Blue bar means promotion in ranking and red bar means demotion in ranking.

3.3. Trustworthy AI in Disease Diagnosis (Function (Phi) Layer)

The scenario-based analysis for the diagnosis of cardiac sarcoidosis from healthy volunteers utilizes the seven criteria outlined in the NIST AI RMF as with the previous two analyses. By identifying the most critical initiatives for diagnosing cardiac sarcoidosis and determining the level of disruption associated with various events, this analysis provides valuable insights to decision-makers. These insights guide investment decisions, allowing stakeholders to prioritize resources where they will yield favorable outcomes.

Tables 11–13 describe 43 initiatives, 50 emergent conditions, and 10 scenarios, respectively, for risk management of AI trustworthiness, in disease diagnosis.

Table A4 illustrates the baseline relevance of the success criteria for the trustworthy AI in disease diagnosis (cardiac sarcoidosis). For instance, criteria *c.01*, *safe*, has high relevance among other criteria in the baseline scenario (See Appendix A).

Table A5 describes the impact of seven success criteria on forty-three initiatives that are introduced above (See Appendix A).

Table A6 shows the criteria-scenario relevance. The criterion *c.01*, *safe*, effectiveness *decreases* under scenario *s.01*, *funding decrease*, and has *somewhat increase* under *s.04*, *cyber-attacks on active system* (See Appendix A).

Figure 8 shows the disruptive score of each scenario. This figure shows that *s.06*—non-interpretable AI and lack of human–AI Communications; *s.03*—privacy attacks; and *s.08*—human errors in design, development, measurement, and implementation have the highest disruption among other scenarios.

Table 11. Initiatives for the *function (Phi)* layer in risk analysis of AI in healthcare. Abridged from various sources that are identified in the narrative.

Index	Initiative
x.Phi.01	Identify At-Risk Components
x.Phi.02	Understanding ML Tools to Uncover Any Patterns in Data
x.Phi.03	Maintaining the Provenance of Training Data
x.Phi.04	Safety/Verifiability of Automated Analyses (Cardiac Region Detection Software)
x.Phi.05	Reproducible Data and Method in Other Health Centers
x.Phi.06	Correctly Labeling the Data
x.Phi.07	Training Data to Follow Application Intellectual Property Rights Laws
x.Phi.08	Informed Consent to Use Data
x.Phi.09	Maintain Organizational Practices Like Implement Risk Management to Reduce Harm Reduction and More Accountable Systems
x.Phi.10	Prioritization Policies and Resources Based on Assesses Risk Levels
x.Phi.11	Safety of Personally Identifiable Information
x.Phi.12	Effective Risk Management by Appropriate Accountability Mechanism, Roles, and Responsibilities, and Incentive Structures for Risk Management to be Effective
x.Phi.13	Avoid Gender and Age Discriminations and Bias in Preparing Data
x.Phi.14	Reducing Unnecessarily Procedures
x.Phi.15	Reducing Costs and Time Consumption
x.Phi.16	Able to Identify Healthy Volunteers before Starting the Procedures
x.Phi.17	Designate Ethical, Legal, Societal, and Technical Boundaries for AI Operation
x.Phi.18	Policy-Makers to Ensure the Moral Demanding Situations are Tackled Proactively
x.Phi.19	Articulate and Document the Concept and Objectives of the System Considering Legal, Regulatory, and Ethical Requirements
x.Phi.20	Gather, Clean and Validate Data and Document the Metadata and Characteristics of the Dataset Considering Legal, Regulatory, and Ethical Requirements
x.Phi.21	Key steps for implementing a new software system: Pilot, Compatibility with Legacy Systems, Regulatory Compliance, Organizational Change Management, and User Experience Evaluation
x.Phi.22	Continuously Assess AI System's Recommendations and Impacts
x.Phi.23	Balancing and Trade Off of Trustworthy AI System Characteristics Based on Context
x.Phi.24	Reducing the Hospitalization Time of the Patient by Correct Diagnostics
x.Phi.25	Explain and Identify Most Important Features Using AI Models
x.Phi.26	Measurements Outlier Findings
x.Phi.27	Closeness of Results of Estimates, Observations, and Computations to the Ground Truth (True Values)
x.Phi.28	Human–AI Teaming
x.Phi.29	Demonstrate Validity or Generalizability Beyond the Training Conditions
x.Phi.30	System's Ability to Maintain its Performance Under Uncertain Circumstances
x.Phi.31	Minimizing Potential Harms to People Under Unexpected Operating Settings
x.Phi.32	Responsible AI System Design, Development and Deployment Practices
x.Phi.33	Clear Information to the Users on Responsible Use of the AI System
x.Phi.34	Deployers and End Users to Make Responsible Decisions
x.Phi.35	Documentation and Explanation of Risks, Grounded in Empirical Evidence from Past Incidents
x.Phi.36	The Ability to Control, Adjust, or Involve Humans in Systems When They Do Not Perform as Intended or Expected
x.Phi.37	Clear and Distinct Definitions of Human Roles and Responsibilities Are Essential for Decision-Making and Oversight in the Context of AI Systems
x.Phi.38	AI Systems May Need More Frequent Maintenance and Triggers for Corrective Maintenance Because of Data, Model, or Concept Drift
x.Phi.39	Managing Risks from Lack of Explainability by Defining the AI System's Functions Considering Users' Role, Knowledge, and Skill Levels
x.Phi.40	The Ability to Describe Why an AI System Made a Specific Prediction or Recommendation
x.Phi.41	Securing Individual Privacy, Anonymity and Confidentiality
x.Phi.42	The Process of Removing Identifying Information and Combining Specific Model Results to Maintain Privacy and Confidentiality in Certain Model Outputs
x.Phi.43	Strengthened Engagement with Relevant AI Actors and Interested Stakeholders
x.Phi.i	Others

Table 12. Emergent conditions used to create sets of scenarios for the *function (Phi)* layer in risk analysis of AI in healthcare. Abridged from various sources that are identified in the narrative.

Index	Emergent Condition
e.Phi.01	Using Non-Important Features in Sarcoidosis Diagnostics as the Input
e.Phi.02	Improperly Labeling the Data in Surgery-Specific Patient Registries
e.Phi.03	Issue of Incorrect Identification and Labeling of Variables in Registries Used for Surgery-Related Patient Data, Highlighting the Potential Consequences of Such Misidentification
e.Phi.04	Misunderstanding AI
e.Phi.05	Limited Generalizability
e.Phi.06	Limitation in Types and Performance of Available Data
e.Phi.07	Expensive Data Collection
e.Phi.08	Time Consuming Data Collection
e.Phi.09	Policy and Regulation Changes
e.Phi.10	Difficult and Complex AI Algorithms Interpretability
e.Phi.11	Lack of AI Determination of Casual Relationships in Data at Clinical Implementation Level
e.Phi.12	Inability of AI in Providing an Automated Clinical Interpretation of its Analysis
e.Phi.13	Human Errors in Measurements
e.Phi.14	Abuse or Misuse of the AI Model or Data
e.Phi.15	Challenges with Training Data to be Subject to Copyright
e.Phi.16	Complicate Risk Measurement by Third Party Software, Hardware and Data
e.Phi.17	Model Fails to Generalize
e.Phi.18	Lack of Robustness and Verifiable Methods for AI Trustworthiness
e.Phi.19	Mis-Identification of Different Risk Perspective in Early or Late Stages of AI Lifecycle
e.Phi.20	Difference Between Controlled Environment vs. Uncontrollable and Real-World Settings
e.Phi.21	Inscrutable Nature of AI Systems in Risk Measurements
e.Phi.22	Systematic Biases in Collecting Clinical Data
e.Phi.23	Risk Tolerance Influence by Legal or Regulatory Requirements Changes
e.Phi.24	Unrealistic Expectations About Risk to Misallocate Resources
e.Phi.25	Residual Risk after Risk Treatment Directly Impacts Healthcare Deployers
e.Phi.26	Privacy Concerns Regarding Using Underlying Data to Train the Systems
e.Phi.27	The Energy and Environmental Implications from Resource-Heavy Computing Demands
e.Phi.28	Security Concerns Related to the Confidentiality of the System Training and Output
e.Phi.29	Security of the System Underlying Software and Hardware
e.Phi.30	One-Size-Fits-All Requirements AI Model Challenges
e.Phi.31	Neglecting the Trustworthy AI Characteristics
e.Phi.32	Difficult Decisions in Trade Off and Balancing Trustworthy AI Characteristics by Organizations
e.Phi.33	Subject Matter Experts Collaborate to Evaluate TEVV Findings, Aligning Parameters with Project Requirements and Deployment Conditions
e.Phi.34	Different Perception of the Trustworthy AI Characteristics Between AI Designer than the Deployer
e.Phi.35	Potential Risk of Serious Injury to the Patients
e.Phi.36	Complexity of Explaining AI System to End Users
e.Phi.37	Data Poisoning
e.Phi.38	Negative Risks Result from an Inability to Appropriately Understand or Contextualize System Output
e.Phi.39	AI Allowing Inference to Identify Individuals or their Private Information
e.Phi.40	Privacy Intrusions
e.Phi.41	Data Sparsity
e.Phi.42	Fairness Perception Difference Among Cultures and Applications
e.Phi.43	Computational and Statistical Biases Stem from Systematic Errors Due to Limited and Non Representative Samples
e.Phi.44	Human Cognitive Biases Relate to How the Stakeholders Perceives AI System Information and Use it to Make Decisions
e.Phi.45	Lack of Access to the Ground Truth in the Dataset
e.Phi.46	Intentional or Unintentional Changes During Training
e.Phi.47	Increased Opacity and Concerns About Reproducibility
e.Phi.48	Impacts of Computational Costs on the Environment and Planet
e.Phi.49	Incapacity to Anticipate or Identify the Adverse Effects of AI-Driven Systems Beyond Statistical Metrics
e.Phi.50	Over-Reliance on AI
e.Phi.i	Others

Table 13. Emergent condition grouping for the *function (Phi)* layer in risk analysis of AI in healthcare shows which emergent conditions fit in each scenario. Abridged from various sources that are identified in the narrative.

	s.01—Funding Decrease	*s.02*—Government Regulation and Policy Changes	*s.03*—Privacy Attacks	*s.04*—Cyber Security Threats	*s.05*—Changes in AI RMF	*s.06*—Non-Interpretable AI and Lack of Human–AI Communications	*s.07*—Global Economic and Societal Crisis	*s.08*—Human Errors in Design, Development, Measurement, and Implementation	*s.09*—Uncontrollable Environment	*s.10*—Expensive Design Process
e.Phi.01		✓								
e.Phi.02								✓		
e.Phi.03								✓		
e.Phi.04						✓				
e.Phi.05						✓				
e.Phi.06										✓
e.Phi.07	✓									✓
e.Phi.08										✓
e.Phi.09		✓			✓					
e.Phi.10						✓				
e.Phi.11						✓				
e.Phi.12						✓				
e.Phi.13						✓				
e.Phi.14						✓		✓		
e.Phi.15		✓	✓			✓				
e.Phi.16						✓				✓
e.Phi.17						✓			✓	
e.Phi.18		✓			✓	✓		✓	✓	✓
e.Phi.19					✓	✓		✓		
e.Phi.20									✓	
e.Phi.21						✓				
e.Phi.22						✓				
e.Phi.23		✓								
e.Phi.24						✓		✓		
e.Phi.25						✓		✓	✓	
e.Phi.26			✓							
e.Phi.27		✓					✓			✓
e.Phi.28			✓	✓						
e.Phi.29				✓						
e.Phi.30						✓		✓	✓	✓
e.Phi.31						✓		✓		
e.Phi.32	✓				✓	✓				✓
e.Phi.33						✓				
e.Phi.34						✓				
e.Phi.35								✓	✓	
e.Phi.36									✓	
e.Phi.37			✓	✓				✓		
e.Phi.38						✓				
e.Phi.39			✓							
e.Phi.40			✓							

Table 13. *Cont.*

	s.01—Funding Decrease	*s.02*—Government Regulation and Policy Changes	*s.03*—Privacy Attacks	*s.04*—Cyber Security Threats	*s.05*—Changes in AI RMF	*s.06*—Non-Interpretable AI and Lack of Human–AI Communications	*s.07*—Global Economic and Societal Crisis	*s.08*—Human Errors in Design, Development, Measurement, and Implementation	*s.09*—Uncontrollable Environment	*s.10*—Expensive Design Process
e.Phi.41								✓	✓	✓
e.Phi.42		✓			✓	✓				
e.Phi.43	✓					✓		✓	✓	✓
e.Phi.44	✓					✓				
e.Phi.45	✓							✓		✓
e.Phi.46					✓	✓		✓		
e.Phi.47						✓		✓		
e.Phi.48	✓						✓			✓
e.Phi.49						✓	✓		✓	
e.Phi.50						✓				

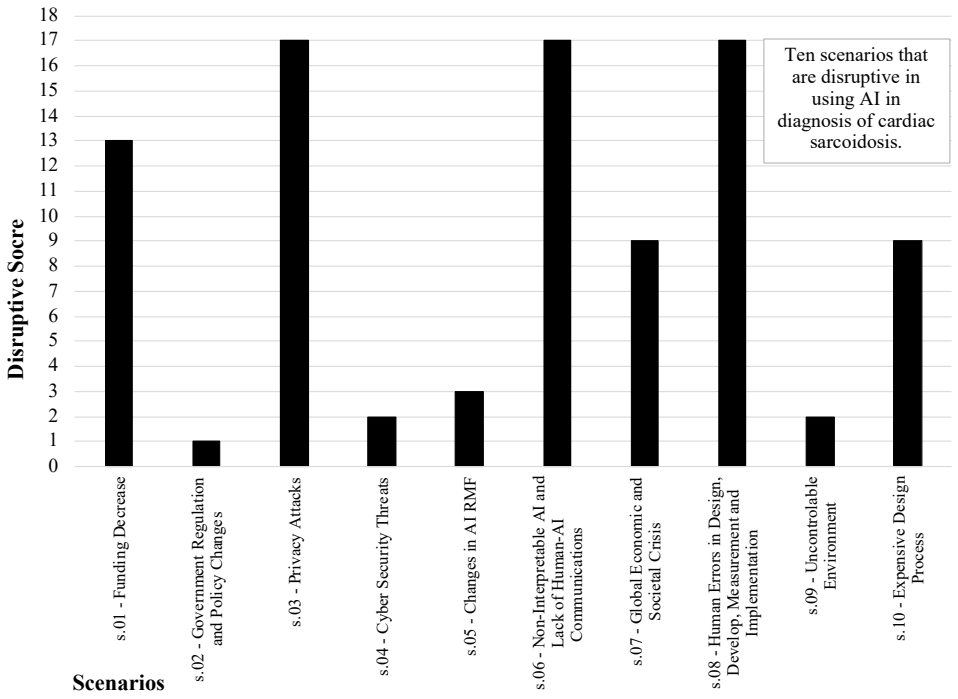

Figure 8. Disruptive score of scenarios is based on sum of squared differences in priority of initiatives, relative to the baseline scenario for the *function (Phi)* layer in risk analysis of AI in healthcare.

Figure 9 shows the variation in the prioritization of initiatives across scenarios. The most important initiatives are *x.Phi.24*—reducing the hospitalization time of the patient by correct diagnostics; *x.Phi.28*—human–AI teaming; *x.Phi.32*—responsible AI system design,

development, and deployment practices; *x.Phi.29*—demonstrate validity or generalizability beyond the training conditions; *x.Phi.27*—closeness of results of estimates, observations, and computations to the ground truth (true values); *x.Phi.25*—explain and identify most important features using AI models; *x.Phi.20*—gather, clean, and validate data and document the metadata and characteristics of the dataset considering legal, regulatory, and ethical requirements; *x.Phi.16*—able to identify healthy volunteers before starting the procedures; *x.Phi.06*—correctly labeling the data; and *x.Phi.04*—safety/verifiability of automated analyses (cardiac region detection software).

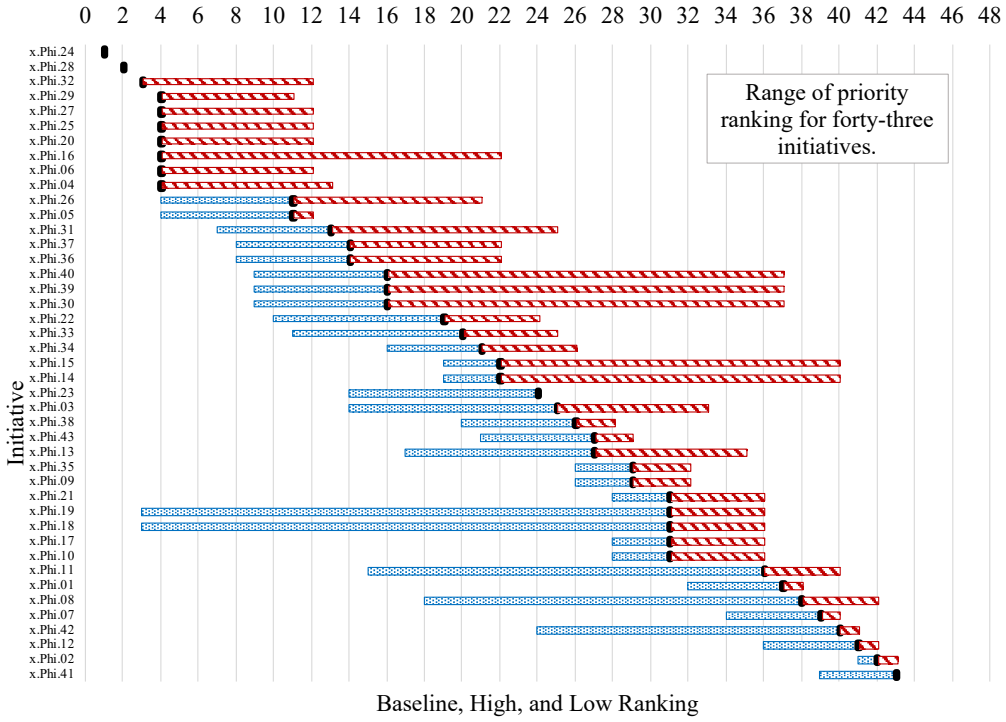

Figure 9. Distributions of initiatives influence rankings based on which emergent conditions could arise more often or never occur for the *function (Phi)* layer in risk analysis of AI in healthcare. Blue bars mean promotion in ranking and red bars mean demotion in ranking.

The *x*-axis in Figures 5, 7 and 9 represents the distributions of initiative rankings based on which emergent conditions that could arise more often or never occur. For instance, in Figure 9, *x.Phi.26* is promoted to rank 4 and demoted in ranking to 21 in different scenarios.

4. Discussion

The novelty of this paper lies in the degree of disruption of the order, focusing on AI in healthcare [53]. The relationship is a complex and multi-expertise enterprise. Moreover, this paper contextualizes the possible and actual implications of AI by introducing a method to quantify risk as the disruption of the order of AI initiatives of healthcare systems, with the aim of finding the scenarios that are most and least disruptive to system order. This novel approach studies scenarios that bring about a re-ordering of initiatives in each of the following three characteristic layers: *purpose*, *structure*, and *function*. The scoring tool is consistent with the recent literature [6,31,32,37].

Tables 14 and 15 suggest that the topic of the scenarios should be used to describe the scope of the tentative project, which shapes and guides the input of the R&D portfolio. This information allows investors and R&D managers to make informed decisions regarding resource allocation. Specifically, they can focus their investments on the most critical initiatives related to the risk analysis of AI in healthcare applications, as outlined in Table 15. For instance, *x.Phi.29*, *demonstrate validity or generalizability beyond the training conditions*, is one of the most important initiatives and trustworthy formal recommendations for

controlling AI risks in the *function* layer. Additionally, they can consider the various scenarios presented in Table 14, ranging from the most disruptive to the least disruptive. The study recommends the following methods for user education about safe AI usage based on the results in Table 14: informing users about why and how the benefits of using the AI system outweigh its risks compared to other technologies on the market, convincing clinicians that specific AI system outcomes are safe, providing information to users on what data to use for training, validating, and testing AI models, including potential changes due to various input data, highlighting that AI systems may require more frequent maintenance and triggers for corrective maintenance due to data, model, or concept drift, demonstrating the validity or generalizability of AI systems beyond the training conditions, emphasizing the closeness of results of estimates, observations, and computations to the ground truth (true values), and advocating for responsible AI system design, development, and deployment practices. This analysis enables the identification of new topics that warrant additional resources and time, with the goal of improving the overall success of the system. For instance, Table 14 highlights scenario *s.06*, *non-interpretable AI and lack of human–AI communications*, as the most disruptive scenario across all three layers of healthcare systems. Although the results from this pilot must be interpreted with caution and validated in a larger sample, this observation is consistent with the findings of [54,55], which indicate that AI transparency solutions primarily target domain experts. Given the emphasis on "high-stakes" AI systems, particularly in healthcare, this inclination is reasonable. It is vital to consider that daily-based tasks that involve AI are not as important for assessing the risks of AI in the domain, such as suggested movies in online streaming or suggesting other items in online shopping systems. Optimizing trustworthy AI properties is recommended in situations where high-stakes environments, such as healthcare, and scenarios involving the handling of sensitive and private data of individuals are present. Another observation is that risks of AI should be context-based [55] and it should consider all the participants and stakeholders in the study for more comprehensive findings. One explanation does not fit all [56]. Moreover, having a human in the loop [57] is important for AI prediction verification and to facilitate effective collaboration and partnership between humans and AI.

Table 14. Most and least disruptive scenarios with respect to rankings of the initiatives for systems characteristic layers in risk analysis of AI in healthcare. Most disruptive scenarios = (+++); least disruptive scenarios = (+).

Scenarios	***Purpose (Pi)***	***Structure (Sig)***	***Function (Phi)***	***Boundary (Bet)***	***Environment (Eps)***	***Interconnections (Iot)***
s.01—Funding Decrease						
s.02—Government Regulation and Policy Changes	+	+	+			
s.03—Privacy Attacks		+	+++			
s.04—Cyber Security Threats		+				
s.05—Changes in AI RMF	+	+				
***s.06*—Non-Interpretable AI and Lack of Human–AI Communications**	**+++**	**+++**	**+++**		-tbd-	
s.07—Global Economic and Societal Crisis	+	+				
s.08—Human Errors in Design, Development, Measurement, and Implementation			+++			
s.09—Uncontrollable Environment		+++				
s.10—Expensive Design Process		+++				

Table 15. Most important initiatives for each of the *system characteristic* layers in risk analysis of AI in healthcare.

Index	Initiative
Purpose (Pi)	*x.Pi.35*—Inform Users on Why and How the Benefits of the Use of an AI System Overweigh its Risks Compared to Other Technologies on the Market *x.Pi.23*—Clinicians to be Convinced that Specific AI System Outcomes are Safe *x.Pi.33*—Users to be Informed of What Data to Use for Training, Validating, and Testing the AI Models; Also, any Potential Changes Due to Various Input Data
Structure (Sig)	*x.Sig.40*—The Ability to Describe Why an AI System Made a Specific Prediction or Recommendation *x.Sig.44*—AI Systems May Need More Frequent Maintenance and Triggers for Corrective Maintenance Because of Data, Model, or Concept Drift *x.Sig.24*—Reduce the Number of Experiments to be Cost- and Time-Effective by Optimizing the Configurations
Function (Phi)	*x.Phi.29*—Demonstrate Validity or Generalizability Beyond the Training Conditions *x.Phi.27*—Closeness of Results of Estimates, Observations, and Computations to the Ground Truth (True Values) *x.Phi.32*—Responsible AI System Design, Development, and Deployment Practices
Boundary (Bet)	Future Work
Environment (Eps)	Future Work
Interconnections (Iot)	Future Work

In healthcare, AI is typically used by experts as a decision-support system. Consequently, the development of solutions prioritizes the needs and requirements of these knowledgeable professionals. Recognizing this context, it becomes evident that addressing the issues of non-interpretable AI and a lack of human–AI communications is crucial within healthcare systems. This is essential not only to ensure patient safety but also to foster trust, consider ethical implications, promote continuous learning, and ensure compliance with legal and regulatory frameworks. The implementation of artificial intelligence in healthcare comes with more human risks than in other sectors due to its unique capacity to directly impact quality of care and healthcare outcomes.

There are some methods that are advised for confirming the efficacy of AI systems after training the dataset, such as confusion matrix analysis, using XAI techniques, having the experts in the loop to validate the outcome, continuous iteration and training monitoring, validation and testing assessments, bias and fairness assessments, and more. Fairness and bias are critical issues to understand and assess in AI that is either applied or used in the healthcare sector. For example, AI requires large, robust "training" databases, but many of the databases used for healthcare and medical datasets are limited. These datasets can perpetuate biases that exist in society and cause further health disparities and inequities [58–60]. It is critical to have a clear understanding of possible biases that could exist in AI systems, as well as how choosing specific outcome variables and labels can impact predictions [61]. Moreover, studies have found that patients have concerns related to AI use in healthcare, including threats to patient choice, increased costs of healthcare, patient privacy and data security, and biases in the data sources used to train AI [62,63]. Successful use and implementation of AI in healthcare settings will require a thoughtful understanding of social determinants of health, health equity, and ethics. The data used in the study were collected in a manner that safeguards the privacy rights of individuals by implementing robust data collection measures, such as data quality assessments and validations by experts, standard data collection procedures, clinic data security measures, and more. Improving data management procedures, including metadata

documentation, collecting, cleansing, and validation, is crucial for ensuring the quality, reliability, and usefulness of data. Integrating new software into an existing system requires careful planning to ensure compatibility, compliance with regulations, and a positive user experience by training on balanced datasets, performing risk analysis and assessment to find potential abnormalities in the dataset, enhancing data protection, and more.

The necessity of AI interpretability and human–AI communications in everyday contexts for end users remains poorly understood. The existing research on this topic is limited, but the available findings suggest that this form of transparency may not be significant to users in their everyday experiences [54]. By prioritizing the most important initiatives and investing in mitigating the most disruptive scenarios in the system, the full potential of AI will be unlocked while responsibly integrating it into healthcare practices, benefiting both patients and the healthcare industry as a whole.

The methods of this paper serve as a demonstration, and they emphasize the constraints associated with each disruptive scenario in tandem with the partial consideration of system layers. This paper serves as a means to enhance transparency. By involving patients and care partners, it mitigates the risks of bias and unintended adverse consequences in AI applications within healthcare systems. The scope of initiatives and emerging conditions extends beyond the aforementioned lists and will be further elaborated upon. While this paper primarily focuses on socioeconomic status, it is important to note that future endeavors will encompass other demographic factors linked to health disparities, such as race/ethnicity, sexual orientation, geographic location, and disability status. As an extension to this paper, the study by [32] demonstrated that developing plans with diverse participants in terms of expertise, aptitude, and background changes the most and least disruptive scenarios in the system.

The upcoming interviews will encompass patients, care partners, and community-based organizations that work with populations affected by health disparities. It is crucial to recognize that individuals, including patients, caregiving partners, and community entities, are assuming increasingly important roles. These entities are acknowledged as authoritative sources due to their personal experiences, a form of knowledge gaining equitable recognition in various national contexts. Consequently, their involvement is vital across all stages, starting from the initial conceptualization of AI application goals in healthcare.

The method is well suited for use by healthcare professionals [53] who lack the background necessary to comprehend and employ more complex methodologies that capture the intricacies of artificial intelligence. This argument acknowledges some of the limitations of the method and provides a clear explanation of why these limitations render it fit for its intended purpose.

The advantage of ordinal over cardinal ratings marks an improvement in ease of elicitation. The ratings in this paper are used as a measurement scale and are not vulnerable to ordinal disadvantages. Ref. [64] points out the subjectivity, loss of granularity, and challenges in prioritization associated with these matrices. Ref. [64] suggests the need for more robust, data-driven approaches to improve the accuracy and reliability of risk assessments through methods such as probabilistic risk assessment (PRA), Bayesian networks, or other quantitative methods [64,65]. To overcome this challenge, Krisper introduces different kinds of distributions, both numerically and graphically. Some common distributions of ranks are *linear*, *logarithmic*, *normally distributed (Gaussian)*, and *arbitrary (fitted)* [66]. For instance, for each scenario in this paper, *linear* distributions of ranks were used. That is, the scales split a value range into equally distributed ranges of {8, 6, 1, 1/6, 1/8}.

As detailed in the Methods and Demonstration section, the disruptiveness (D_k) of scenario s_k is calculated as the sum of the squared differences in priority for each initiative when compared to the baseline scenario. These scores are then normalized within the range of 0 to 100 for easy comparison. It is crucial to interpret these results thoughtfully before engaging in further discussions on alternatives, including nonlinear combinations of statements within multi-criteria decision analysis frameworks. The interpretation

should be undertaken by principals and managers, taking into account the context of different systems.

Rozell (2015) describes the challenges of using qualitative and semi-qualitative risk ranking systems. When time and resources are limited, obtaining a simple, fully quantitative risk assessment or an informal expert managerial review and judgment are considered better approaches [67]. In this paper, expert managerial review and judgement are the core of the risk registers across all three layers.

The innovation of the paper is not in the scoring but rather in the measurement of risk via the disruptions of a system order using the scenarios. The readers are encouraged to select their own ways of ordering and re-ordering the initiatives. The identification of scenarios that most disrupt the system order helps healthcare professionals in the characterization of AI-related risks. This characterization occurs in parallel across various system layers: *purpose*, *structure*, and *function*. The method contributes to the reduction of errors by offering a user-friendly interface that enhances accessibility and ease of use. It promotes adaptability, providing flexibility to accommodate diverse healthcare settings and contexts. This usability fosters increased engagement from both experts and stakeholders, facilitating a more inclusive and comprehensive analysis of AI-related risks [68] within the healthcare sector.

As a scenario-based methodology, this study identified the least and most disruptive scenarios within the context of the identified scenarios, based on the available sources and data during the study. Limited access to additional data and documents, as well as restricted stakeholder engagement, are additional limitations. It is important to consider the potential for biases among stakeholders and experts during the interview process, given their diverse motivations. To mitigate any strategic or manipulative behavior that might affect the analysis results, conducting an investigation focused on identifying the most disruptive scenarios could be beneficial. The primary aim was not solely to aggregate stakeholder inputs but also to identify areas requiring further examination, preserving the unique influences of individual stakeholders.

5. Conclusions

This study focuses on research and development priorities for managing the risks associated with trustworthy AI in health applications [69,70]. The methodology serves as a demonstration, and it emphasizes the constraints associated with the chosen scenarios and the partial consideration of system layers. The methodology identifies success criteria, R&D initiatives, and emergent conditions across multiple layers of the healthcare system, including the *purpose (Pi)* layer, implant/device or *structure (Sig)* layer, and disease diagnosis or *function (Phi)* layer. The success criteria are consistently applied across all layers of the study.

The core concept of the paper is not to make the judgments required by the model; instead, the focus is on measuring the disruptive order. In other words, the emphasis is on adapting a figure of merit to score the initiatives and rank them rather than performing a decision analysis.

This paper strikes a balance between the goals of AI, human rights, and societal values by considering the seven main characters of the NIST AI risk management framework as the main success criteria for all layers, while also involving a variety of perspectives, stakeholders, managers, and experts in each system layer in the process. By analyzing these initiatives, emergent conditions, and scenarios within the healthcare system layers, the study identifies the most and least disruptive scenarios based on stakeholder preferences [6]. This information allows stakeholders and managers to make informed decisions regarding resource allocation and prioritize certain initiatives over others.

Figures 4, 6 and 8 illustrate the potential disruptions caused by non-interpretable AI and a lack of human–AI communications, which is in line with the research by [71]. Conversely, Figures 5, 7 and 9 emphasize the significant role of interpretable and explainable AI in the healthcare system [72,73]. As AI-based algorithms gain increasing attention and

results in the healthcare sector, it becomes crucial to enhance their understandability for human users, as emphasized by [74].

The initiatives outlined in this paper hold promise for improving communication and mitigating the risks associated with AI in healthcare applications, involving various stakeholders. Moving forward, it is crucial to incorporate the viewpoints of healthcare practitioners and patients who are directly impacted by these approaches.

By acknowledging the biases and perspectives of individuals and communities, the proposed scenarios can effectively capture the diverse weights assigned by different stakeholders [39]. The matter of expert bias is of concern, not only in this context but also across the broader field. Various approaches could be employed to alleviate such biases. These methods include techniques such as simple averaging, assigning importance weights to experts, employing the Analytic Hierarchy Process (AHP), Fuzzy Analytic Hierarchy Process (FAHP), decomposing complex problems into multiple layers, and others. Stakeholders could be weighted in future efforts according to their level of expertise in the field.

Notably, the methods presented in this paper can offer patients valuable insights into the relevance of AI applications in their treatment plans, promoting transparency for both patients and caregivers. The initiatives and emergent conditions discussed in this study provide a foundation for future research, which will build upon these findings to delve deeper into the subject. Further investigations will expand the analysis to encompass additional layers, such as the boundary (Bet) that exists between patients and society. This expanded scope will explore the wider implications of AI in healthcare systems, shedding light on its impact on various aspects of society.

In summary, addressing the major challenge of risk assessments for AI tools, this paper introduces a context-specific approach to understanding the risks associated with AI, emphasizing that these risks cannot be universally applied. The proposed AI risk framework in this study recognizes this context within three layers of healthcare systems. It provides insights into quantifying risk by assessing the disturbance to the order of AI initiatives in healthcare systems. The objective is to identify scenarios, analyzing their impact on system order, and organizing them from the most to least disruptive. Additionally, this study highlights the significant role of humans in the loop in identifying the risks associated with AI in healthcare and evaluating and improving the suggestions and outcomes of AI systems.

There are additional components of an effective AI risk management framework that may guarantee the accuracy and consistency of outputs produced by AI. These include fostering diversity among participants [32], identifying AI effects in terms of ethics [75], law, society, and technology, seeking official guidelines from experts, considering various social values, enhancing and improving unbiased algorithms and data quality by prioritizing privacy and security, and regular maintenance of AI systems [22]. Moreover, identifying and minimizing uncertainties and unexpected scenarios, adhering to ethical and legal standards, ensuring the correctness of AI outputs and predictions through various validation and assessment practices, such as employing Explainable AI (XAI) techniques [76], ensuring human–AI teaming [32] and collaboration, and optimizing AI features and performance during design and implementation, among other aspects, are more components of an effective AI risk management framework. Given different business sizes and resource availability, and based on the experience mentioned above, it is clear that there is a need and opportunity for each system principal to determine appropriate AI risk management frameworks.

There are many potential methods for identifying reliable and trustworthy formal guidance for AI risk management. Seeking government guidance and guidelines from officials, R&D findings from industry and academia, verifying compliance with standard and legal protocols, and more could be some of the sources for risk management with AI. There are several safeguards and security measures that can be implemented to ensure the dependability and error-free operation of AI systems, such as validating the results by engaging the patients, medical professionals, and system designers in the loop, identifying

and mitigating the risks of uncertain scenarios to the system, regular monitoring, and updating/training the system to adhere to ethical and lawful standards and protocols.

The methods outlined in this paper hold potential for cross-domain applicability beyond the healthcare sector. They can be adapted and applied to diverse fields such as transportation, finance, design, risk analysis of quantum technologies in medicine, and more [77]. By enhancing transparency and addressing the associated risks of AI, this research benefits not only healthcare systems globally but also various other applications and industries. The findings and insights gained from this study can inform and guide the development and implementation of AI systems in a wide range of domains, such as supply chains, disaster management, emergency response, and more, fostering responsible and effective use of this technology. In summary, one view of this work is that it concentrates the opinions and consensus of a few stakeholders and that the conclusions are limited to a specific topic. On the other hand, the method and its rubrics have general relevance to a variety of life science topics across medical diagnosis, epidemiology, pathology, pharmacology, toxicology, microbiology, immunology, and more.

Author Contributions: Conceptualization, N.M. (Negin Moghadasi) and J.H.L.; methodology, N.M. (Negin Moghadasi), D.C.L. and J.H.L.; software, N.M. (Negin Moghadasi); validation, M.P. and N.M. (Negar Moghaddasi); formal analysis, N.M. (Negin Moghadasi); investigation, M.P., N.M. (Negin Moghadasi) and N.M. (Negar Moghaddasi); resources, J.H.L., M.P. and T.L.P.; data curation, N.M. (Negin Moghadasi) and M.P.; writing—original draft preparation, N.M. (Negin Moghadasi) and N.M. (Negar Moghaddasi); writing—review and editing, D.C.L., R.S.V., I.L. and M.P.; visualization, N.M. (Negin Moghadasi); supervision, J.H.L.; project administration, N.M. (Negin Moghadasi) and J.H.L.; funding acquisition, J.H.L. and T.L.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Commonwealth Center for Advanced Logistics Systems (CCALS) and the National Science Foundation (NSF) Center for Hardware and Embedded Systems Security and Trust (CHEST) with the grant number 1916760. The APC was funded by The University of Virginia.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to acknowledge the Commonwealth Center for Advanced Logistics Systems (CCALS) and National Science Foundation (NSF) Center for Hardware and Embedded Systems Security and Trust (CHEST) for supporting this effort.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Table A1. Baseline relevance for the *structure (Sig)* layer in risk analysis of AI in healthcare.

The Criterion c.xx Has	s.00—Baseline	Relevance among the Other Criteria
c.01—safe has	medium	relevance
c.02—secure and resilient has	medium	relevance
c.03—explainable and interpretable has	high	relevance
c.04—privacy enhanced has	low	relevance
c.05—fair—with harmful bias managed has	low	relevance
c.06—accountable and transparent has	high	relevance
c.07—valid and reliable has	high	relevance

Table A2. The criteria-initiative assessment shows how well each initiatives addresses the success criteria for the *structure (Sig)* layer in risk analysis of AI in healthcare. Strongly agree is represented by a filled circle (●); agree is represented by a half-filled circle (◐); somewhat agree is represented by an unfilled circle (○); and neutral is represented by a dash (—).

	c.01	***c.02***	***c.03***	***c.04***	***c.05***	***c.06***	***c.07***
x.Sig.01	●	◐	—	—	—	—	○
x.Sig.02	—	—	○	—	—	—	●
x.Sig.03	—	—	○	—	—	○	●
x.Sig.04	—	—	●	—	—	◐	●
x.Sig.05	—	—	●	—	—	◐	—
x.Sig.06	○	—	—	—	●	●	●
x.Sig.07	○	○	◐	—	—	○	◐
x.Sig.08	●	●	◐	—	—	●	●
x.Sig.09	○	◐	●	—	○	●	○
x.Sig.10	—	—	◐	○	○	◐	◐
x.Sig.11	—	—	—	●	—	◐	—
x.Sig.12	○	○	◐	—	—	○	◐
x.Sig.13	○	○	○	○	○	○	○
x.Sig.14	—	—	—	—	◐	—	—
x.Sig.15	—	—	●	—	—	—	—
x.Sig.16	○	○	○	—	—	○	○
x.Sig.17	○	○	◐	—	—	○	◐
x.Sig.18	—	—	○	—	○	—	—
x.Sig.19	●	○	●	—	—	◐	◐
x.Sig.20	●	○	●	—	—	●	◐
x.Sig.21	○	—	◐	—	—	—	—
x.Sig.22	●	●	●	—	—	●	●
x.Sig.23	○	○	●	—	—	●	○
x.Sig.24	●	●	●	—	○	●	●
x.Sig.25	●	●	●	—	—	●	●
x.Sig.26	●	●	●	—	—	●	●
x.Sig.27	—	—	◐	—	—	●	●
x.Sig.28	◐	◐	●	—	—	◐	◐
x.Sig.29	◐	◐	●	—	—	◐	◐
x.Sig.30	●	●	●	—	—	●	●
x.Sig.31	●	●	●	—	—	●	●
x.Sig.32	●	●	●	—	—	●	●
x.Sig.33	●	●	●	—	—	●	●
x.Sig.34	◐	○	●	—	—	◐	◐
x.Sig.35	●	●	●	—	—	●	◐
x.Sig.36	●	●	●	—	—	●	—
x.Sig.37	●	●	○	—	—	●	—
x.Sig.38	◐	◐	○	—	—	○	◐
x.Sig.39	●	●	●	—	—	●	●
x.Sig.40	●	●	●	—	◐	●	●
x.Sig.41	—	—	—	○	—	—	—
x.Sig.42	—	—	●	○	—	—	—
x.Sig.43	◐	◐	●	—	—	◐	◐
x.Sig.44	◐	●	●	—	●	●	●
x.Sig.45	◐	◐	●	—	—	◐	◐
x.Sig.46	◐	◐	●	—	—	◐	◐
x.Sig.47	◐	◐	●	—	—	●	◐

Table A3. The criteria-scenario relevance shows how well each scenario fits the success criterion in for the *structure (Sig)* layer in risk analysis of AI in healthcare. Decrease somewhat = DS; decrease = D; somewhat increase = SI; increase = I.

	s.01	***s.02***	***s.03***	***s.04***	***s.05***	***s.06***	***s.07***	***s.08***	***s.09***	***s.10***
c.01	D	SI	-	-	SI	D	DS	D	D	D
c.02	D	SI	-	-	SI	D	DS	DS	DS	D
c.03	DS	SI	-	-	I	D	DS	D	D	D
c.04	-	SI	D	DS	-	-	-	-	-	-
c.05	DS	-	-	-	SI	-	DS	DS	-	-
c.06	D	SI	-	-	I	D	DS	D	D	D
c.07	D	SI	-	-	I	D	DS	D	D	D

Table A4. Baseline relevance for the *function (Phi)* layer in risk analysis of AI in healthcare.

The Criterion c.xx Has	**s.00—Baseline**	**Relevance among the Other Criteria**
c.01—safe has	high	relevance
c.02—secure and resilient has	medium	relevance
c.03—explainable and interpretable has	high	relevance
c.04—privacy enhanced has	medium	relevance
c.05—fair—with harmful bias managed has	medium	relevance
c.06—accountable and transparent has	high	relevance
c.07—valid and reliable has	high	relevance

Table A5. The criteria-initiative assessment shows how well each initiative addresses the success criteria for the *function (Phi)* layer in risk analysis of AI in healthcare. Strongly agree is represented by a filled circle (●); agree is represented by a half-filled circle (◐); somewhat agree is represented by an unfilled circle (○); and neutral is represented by a dash (—).

	c.01	***c.02***	***c.03***	***c.04***	***c.05***	***c.06***	***c.07***
x.Phi.01	●	◐	○	○	○	○	○
x.Phi.02	○	—	○	—	—	◐	◐
x.Phi.03	●	—	○	◐	◐	◐	●
x.Phi.04	●	◐	●	○	◐	●	●
x.Phi.05	●	●	●	◐	◐	◐	●
x.Phi.06	●	●	◐	◐	◐	●	●
x.Phi.07	○	○	◐	○	○	○	◐
x.Phi.08	●	●	—	●	—	○	○
x.Phi.09	○	◐	◐	○	○	●	○
x.Phi.10	○	○	◐	○	○	◐	◐
x.Phi.11	●	●	—	●	○	◐	—
x.Phi.12	○	○	◐	—	—	○	◐
x.Phi.13	◐	◐	○	○	●	◐	◐
x.Phi.14	●	●	—	—	●	●	●
x.Phi.15	●	●	—	—	●	●	●
x.Phi.16	●	●	◐	○	●	●	●
x.Phi.17	○	○	◐	○	○	◐	◐
x.Phi.18	○	○	○	●	●	◐	○
x.Phi.19	○	○	○	●	●	◐	○
x.Phi.20	●	●	◐	◐	◐	●	●
x.Phi.21	○	○	◐	○	○	◐	◐
x.Phi.22	◐	◐	●	○	○	●	●
x.Phi.23	◐	◐	◐	◐	◐	◐	◐
x.Phi.24	●	●	●	◐	●	●	●
x.Phi.25	●	●	◐	◐	◐	●	●

Table A5. *Cont.*

	c.01	***c.02***	***c.03***	***c.04***	***c.05***	***c.06***	***c.07***
x.Phi.26	●	●	●	○	○	●	●
x.Phi.27	●	●	◐	◐	◐	●	●
x.Phi.28	●	●	●	◐	◐	●	●
x.Phi.29	◐	●	●	◐	◐	●	●
x.Phi.30	●	●	●	—	—	●	●
x.Phi.31	●	●	◐	—	●	●	●
x.Phi.32	●	●	●	◐	○	●	●
x.Phi.33	○	●	●	○	○	●	●
x.Phi.34	◐	◐	●	○	○	◐	●
x.Phi.35	○	◐	◐	○	○	●	○
x.Phi.36	●	●	●	○	○	●	◐
x.Phi.37	●	●	●	○	○	●	◐
x.Phi.38	◐	●	●	○	○	○	◐
x.Phi.39	●	●	●	—	—	●	●
x.Phi.40	●	●	●	—	—	●	●
x.Phi.41	—	—	—	●	—	—	—
x.Phi.42	○	○	●	○	—	○	○
x.Phi.43	○	◐	●	○	○	◐	◐

Table A6. The criteria-scenario assessment describes how the scenarios influence the relevance of each success criterion for the *function (Phi)* layer in risk analysis of AI in healthcare. Decrease somewhat = DS; decrease = D; somewhat increase = SI; increase = I.

	s.01	***s.02***	***s.03***	***s.04***	***s.05***	***s.06***	***s.07***	***s.08***	***s.09***	***s.10***
c.01	D	SI	D	-	SI	DS	DS	D	DS	DS
c.02	D	SI	D	-	SI	D	DS	D	DS	DS
c.03	DS	SI	D	-	I	D	-	D	D	-
c.04	-	I	-	DS	-	-	-	-	DS	-
c.05	DS	I	-	-	SI	-	DS	-	DS	DS
c.06	D	SI	D	-	I	D	DS	D	D	DS
c.07	D	SI	D	-	I	D	DS	D	D	DS

References

1. Austin, P.C.; Fine, J.P. Practical Recommendations for Reporting F Ine—G Ray Model Analyses for Competing Risk Data. *Stat. Med.* **2017**, *36*, 4391–4400. [CrossRef]
2. Matzinger, P. Tolerance, Danger, and the Extended Family. *Annu. Rev. Immunol.* **1994**, *12*, 991–1045. [CrossRef]
3. Christensen, C.M. Marketing Strategy: Learning by Doing. *Harv. Bus. Rev.* **1997**, *75*, 141–146, 148–156. [PubMed]
4. Borgonovo, E.; Cappelli, V.; Maccheroni, F.; Marinacci, M. Risk Analysis and Decision Theory: A Bridge. *Eur. J. Oper. Res.* **2018**, *264*, 280–293. [CrossRef]
5. Bier, V.; Gutfraind, A. Risk Analysis beyond Vulnerability and Resilience—Characterizing the Defensibility of Critical Systems. *Eur. J. Oper. Res.* **2019**, *276*, 626–636. [CrossRef]
6. Moghadasi, N.; Collier, Z.A.; Koch, A.; Slutzky, D.L.; Polmateer, T.L.; Manasco, M.C.; Lambert, J.H. Trust and Security of Electric Vehicle-to-Grid Systems and Hardware Supply Chains. *Reliab. Eng. Syst. Saf.* **2022**, *225*, 108565. [CrossRef]
7. Furgal, C.M.; Boyd, A.D.; Mayeda, A.M.; Jardine, C.G.; Driedger, S.M. Risk Communication and Perceptions about Lead Ammunition and Inuit Health in Nunavik, Canada. *Int. J. Circumpolar Health* **2023**, *82*, 2218014. [CrossRef] [PubMed]
8. Niemeier, R.T.; Williams, P.R.D.; Rossner, A.; Clougherty, J.E.; Rice, G.E. A Cumulative Risk Perspective for Occupational Health and Safety (OHS) Professionals. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6342. [CrossRef]
9. Redinger, C.F.; Boelter, F.W.; O'Reilly, M.V.; Howard, J.; Barbi, G.J. Decision Making in Managing Risk. In *Patty's Industrial Hygiene*; Harris, R., Ed.; Wiley: Hoboken, NJ, USA, 2021; pp. 1–24, ISBN 978-0-471-29784-0.
10. Binsaeed, R.H.; Yousaf, Z.; Grigorescu, A.; Samoila, A.; Chitescu, R.I.; Nassani, A.A. Knowledge Sharing Key Issue for Digital Technology and Artificial Intelligence Adoption. *Systems* **2023**, *11*, 316. [CrossRef]
11. Huang, C.-C.; Ruan, S.-J.; Chen, H.H.; Tu, Y.-W.; Chang, L.-C. Chinese Articulation Disorder-Correcting Application Based on Neural Networks. In Proceedings of the 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 15–18 October 2019; pp. 743–744.

12. Lei, K.-Z.; Ku, M.-Y.; Lee, S.-Y. Real-Time and Non-Contact Arrhythmia Recognition Algorithm for Hardware Implementation. In Proceedings of the 2022 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE), Tainan, Taiwan, 7 November 2022; pp. 1–2.
13. Elvas, L.B.; Ferreira, J.C.; Dias, M.S.; Rosário, L.B. Health Data Sharing towards Knowledge Creation. *Systems* **2023**, *11*, 435. [CrossRef]
14. Dicuonzo, G.; Donofrio, F.; Fusco, A.; Shini, M. Healthcare System: Moving Forward with Artificial Intelligence. *Technovation* **2023**, *120*, 102510. [CrossRef]
15. Habchi, Y.; Himeur, Y.; Kheddar, H.; Boukabou, A.; Atalla, S.; Chouchane, A.; Ouamane, A.; Mansoor, W. AI in Thyroid Cancer Diagnosis: Techniques, Trends, and Future Directions. *Systems* **2023**, *11*, 519. [CrossRef]
16. Dauda, O.I.; Awotunde, J.B.; AbdulRaheem, M.; Salihu, S.A. Basic Issues and Challenges on Explainable Artificial Intelligence (XAI) in Healthcare Systems. In *Advances in Medical Technologies and Clinical Practice*; de Albuquerque, V.H.C., Srinivasu, P.N., Bhoi, A.K., Briones, A.G., Eds.; IGI Global: Hershey, PA, USA, 2022; pp. 248–271, ISBN 978-1-66843-791-9.
17. Valdez, R.S.; Ancker, J.S.; Veinot, T.C. Provocations for Reimagining Informatics Approaches to Health Equity. *Yearb. Med. Inf.* **2022**, *31*, 015–019. [CrossRef]
18. Stødle, K.; Flage, R.; Guikema, S.D.; Aven, T. Data-driven Predictive Modeling in Risk Assessment: Challenges and Directions for Proper Uncertainty Representation. *Risk Anal.* **2023**, *43*, 2644–2658. [CrossRef] [PubMed]
19. Chen, C.; Lin, K.; Rudin, C.; Shaposhnik, Y.; Wang, S.; Wang, T. An Interpretable Model with Globally Consistent Explanations for Credit Risk. *arXiv* **2018**, arXiv:1811.12615. [CrossRef]
20. Avin, S.; Belfield, H.; Brundage, M.; Krueger, G.; Wang, J.; Weller, A.; Anderljung, M.; Krawczuk, I.; Krueger, D.; Lebensold, J.; et al. Filling Gaps in Trustworthy Development of AI. *Science* **2021**, *374*, 1327–1329. [CrossRef] [PubMed]
21. Jain, S.; Luthra, M.; Sharma, S.; Fatima, M. Trustworthiness of Artificial Intelligence. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 907–912.
22. Tabassi, E. *AI Risk Management Framework: AI RMF (1.0)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023; pp. 1–48. Available online: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225 (accessed on 29 January 2024). [CrossRef]
23. Lo, S.K.; Liu, Y.; Lu, Q.; Wang, C.; Xu, X.; Paik, H.-Y.; Zhu, L. Toward Trustworthy AI: Blockchain-Based Architecture Design for Accountability and Fairness of Federated Learning Systems. *IEEE Internet Things J.* **2023**, *10*, 3276–3284. [CrossRef]
24. Zolanvari, M.; Yang, Z.; Khan, K.; Jain, R.; Meskin, N. TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security. *IEEE Internet Things J.* **2023**, *10*, 2967–2978. [CrossRef]
25. Ramírez-Gutiérrez, A.G.; Solano García, P.; Morales Matamoros, O.; Moreno Escobar, J.J.; Tejeida-Padilla, R. Systems Approach for the Adoption of New Technologies in Enterprises. *Systems* **2023**, *11*, 494. [CrossRef]
26. Holt, J. *Systems Engineering Demystified*; Packt Publishing: Birmingham, UK, 2021; ISBN 978-1-83898-580-6.
27. Sage, A.P.; Lynch, C.L. Systems Integration and Architecting: An Overview of Principles, Practices, and Perspectives. *Syst. Engin.* **1998**, *1*, 176–227. [CrossRef]
28. *SEBoK Editorial Board Guide to the Systems Engineering Body of Knowledge (SEBoK)*; Version 2.7; Stevens Institute of Technology: Hoboken, NJ, USA, 2022.
29. Walden, D.D.; Roedler, G.J.; Forsberg, K.; Hamelin, R.D.; Shortell, T.M.; International Council on Systems Engineering (Eds.) *Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities*, 4th ed.; Wiley: Hoboken, NJ, USA, 2015; ISBN 978-1-118-99941-7.
30. Davis, G.B. Strategies for Information Requirements Determination. *IBM Syst. J.* **1982**, *21*, 4–30. [CrossRef]
31. Loose, D.C.; Eddy, T.L.; Polmateer, T.L.; Manasco, M.C.; Moghadasi, N.; Lambert, J.H. Managing Pandemic Resilience with Other Cascading Disruptions of a Socio-Technical System. In Proceedings of the 2022 IEEE International Systems Conference (SysCon), Montreal, QC, Canada, 25 April 2022; pp. 1–6.
32. Moghadasi, N.; Piran, M.; Baek, S.; Valdez, R.S.; Porter, M.D.; Johnson, D.; Lambert, J.H. Systems Analysis of Bias and Risk in AI Enabled Medical Diagnosis. In Proceedings of the 2023 IEEE Symposium Series on Computational Intelligence (SSCI), Mexico City, Mexico, 5–8 September 2023.
33. Budimir, S.; Fontaine, J.R.J.; Huijts, N.M.A.; Haans, A.; Loukas, G.; Roesch, E.B. Emotional Reactions to Cybersecurity Breach Situations: Scenario-Based Survey Study. *J. Med. Internet Res.* **2021**, *23*, e24879. [CrossRef] [PubMed]
34. Morton, A.; Fasolo, B. Behavioural Decision Theory for Multi-Criteria Decision Analysis: A Guided Tour. *J. Oper. Res. Soc.* **2009**, *60*, 268–275. [CrossRef]
35. Krantz, D.H.; Luce, R.D.; Suppes, P.; Tversky, A. *Foundations of Measurement*; Academic Press: Cambridge, MA, USA, 1971.
36. Von Winterfeldt, D.; Edwards, W. *Decision Analysis and Behavioral Research*; Cambridge University Press: Cambridge, UK, 1986.
37. Collier, Z.A.; Lambert, J.H. Evaluating Management Actions to Mitigate Disruptive Scenario Impacts in an E-Commerce Systems Integration Project. *IEEE Syst. J.* **2019**, *13*, 593–602. [CrossRef]
38. Karvetski, C.W.; Lambert, J.H. Evaluating Deep Uncertainties in Strategic Priority-Setting with an Application to Facility Energy Investments. *Syst. Engin.* **2012**, *15*, 483–493. [CrossRef]
39. Keeney, R.L. Common Mistakes in Making Value Trade-Offs. *JSTOR* **2002**, *50*, 935–945.
40. Keeney, R.L.; Raiffa, H.; Rajala, D.W. Decisions with Multiple Objectives: Preferences and Value Trade-Offs. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 403. [CrossRef]

41. Stillwell, W.G.; Seaver, D.A.; Edwards, W. A Comparison of Weight Approximation Techniques in Multiattribute Utility Decision Making. *Organ. Behav. Hum. Perform.* **1981**, *28*, 62–77. [CrossRef]
42. Lazzerini, B.; Mkrtchyan, L. Analyzing Risk Impact Factors Using Extended Fuzzy Cognitive Maps. *IEEE Syst. J.* **2011**, *5*, 288–297. [CrossRef]
43. Keeney, R.L. *Value-Focused Thinking: A Path to Creative Decisionmaking*; Harvard Univ. Press: Cambridge, MA, USA, 1992; ISBN 978-0-674-93198-5.
44. Belton, V.; Stewart, T.J. *Multiple Criteria Decision Analysis*; Springer US: Boston, MA, USA, 2002; ISBN 978-1-4613-5582-3.
45. Clemen, R.T. *Making Hard Decisions with DecisionTools*, 2nd ed.; Duxbury Thomson Learning: Pacific Grove, CA, USA, 2001; ISBN 978-0-495-01508-6.
46. Karvetski, C.W.; Lambert, J.H.; Linkov, I. Emergent Conditions and Multiple Criteria Analysis in Infrastructure Prioritization for Developing Countries—Karvetski—2009—Journal of Multi-Criteria Decision Analysis—Wiley Online Library. Available online: https://onlinelibrary.wiley.com/doi/10.1002/mcda.444 (accessed on 13 June 2022).
47. Kiseleva, A.; Kotzinos, D.; De Hert, P. Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. *Front. Artif. Intell.* **2022**, *5*, 879603. [CrossRef]
48. Montemayor, C.; Halpern, J.; Fairweather, A. In Principle Obstacles for Empathic AI: Why We Can't Replace Human Empathy in Healthcare. *AI Soc.* **2022**, *37*, 1353–1359. [CrossRef]
49. Abbas, S.W.; Hamid, M.; Alkanhel, R.; Abdallah, H.A. Official Statistics and Big Data Processing with Artificial Intelligence: Capacity Indicators for Public Sector Organizations. *Systems* **2023**, *11*, 424. [CrossRef]
50. Bao, Y.; Gong, W.; Yang, K. A Literature Review of Human–AI Synergy in Decision Making: From the Perspective of Affordance Actualization Theory. *Systems* **2023**, *11*, 442. [CrossRef]
51. Pagano, M. *HEMI Fellow Sung Hoon Kang Receives Cohen Fund Grant for Work on a 3D-Printed Medical Device*; Hopkins Extreme Materials Institute: Baltimore, MD, USA, 2020.
52. Chen, F.; Zhou, J.; Holzinger, A.; Fleischmann, K.R.; Stumpf, S. Artificial Intelligence Ethics and Trust: From Principles to Practice. *IEEE Intell. Syst.* **2023**, *38*, 5–8. [CrossRef]
53. Chinnasamy, P.; Albakri, A.; Khan, M.; Raja, A.A.; Kiran, A.; Babu, J.C. Smart Contract-Enabled Secure Sharing of Health Data for a Mobile Cloud-Based E-Health System. *Appl. Sci.* **2023**, *13*, 3970. [CrossRef]
54. Haresamudram, K.; Larsson, S.; Heintz, F. Three Levels of AI Transparency. *Computer* **2023**, *56*, 93–100. [CrossRef]
55. Chimatapu, R.; Hagras, H.; Starkey, A.; Owusu, G. Explainable AI and Fuzzy Logic Systems. In *Theory and Practice of Natural Computing*; Fagan, D., Martín-Vide, C., O'Neill, M., Vega-Rodríguez, M.A., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11324, pp. 3–20, ISBN 978-3-030-04069-7.
56. Arya, V.; Bellamy, R.K.E.; Chen, P.-Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A.; et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv* **2019**, arXiv:1909.03012. [CrossRef]
57. Bhattacharya, M.; Penica, M.; O'Connell, E.; Southern, M.; Hayes, M. Human-in-Loop: A Review of Smart Manufacturing Deployments. *Systems* **2023**, *11*, 35. [CrossRef]
58. Celi, L.A.; Cellini, J.; Charpignon, M.-L.; Dee, E.C.; Dernoncourt, F.; Eber, R.; Mitchell, W.G.; Moukheiber, L.; Schirmer, J.; Situ, J.; et al. Sources of Bias in Artificial Intelligence That Perpetuate Healthcare Disparities—A Global Review. *PLoS Digit. Health* **2022**, *1*, e0000022. [CrossRef]
59. Murray, S.G.; Wachter, R.; Cucina, R.J. Discrimination By Artificial Intelligence In A Commercial Electronic Health Record—A Case Study. *Health Aff. Forefr.* **2020**. [CrossRef]
60. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* **2019**, *366*, 447–453. [CrossRef]
61. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Algorithmic Bias In Health Care: A Path Forward. *Health Aff. Forefr.* **2019**. [CrossRef]
62. Richardson, J.P.; Smith, C.; Curtis, S.; Watson, S.; Zhu, X.; Barry, B.; Sharp, R.R. Patient Apprehensions about the Use of Artificial Intelligence in Healthcare. *Npj Digit. Med.* **2021**, *4*, 140. [CrossRef]
63. Wang, B.; Asan, O.; Mansouri, M. Patients' Perceptions of Integrating AI into Healthcare: Systems Thinking Approach. In Proceedings of the 2022 IEEE International Symposium on Systems Engineering (ISSE), Vienna, Austria, 24 October 2022; pp. 1–6.
64. Anthony (Tony)Cox, L. What's Wrong with Risk Matrices? *Risk Anal.* **2008**, *28*, 497–512. [CrossRef] [PubMed]
65. Cox (Tony), L.A.; Babayev, D.; Huber, W. Some Limitations of Qualitative Risk Rating Systems. *Risk Anal.* **2005**, *25*, 651–662. [CrossRef]
66. Krisper, M. Problems with Risk Matrices Using Ordinal Scales. *arXiv* **2021**, arXiv:2103.05440. [CrossRef]
67. Rozell, D.J. A Cautionary Note on Qualitative Risk Ranking of Homeland Security Threats. *J. NPS Cent. Homel. Def. Secur.* **2015**. Available online: https://www.hsaj.org/articles/1800 (accessed on 29 January 2024).
68. Nature Editorials, dalking about Tomorrow's AI Doomsday When AI Poses Risks Today. *Nature* **2023**, *618*, 885–886. [CrossRef]
69. Duenser, A.; Douglas, D.M. Whom to Trust, How and Why: Untangling Artificial Intelligence Ethics Principles, Trustworthiness, and Trust. *IEEE Intell. Syst.* **2023**, *38*, 19–26. [CrossRef]
70. Schmid, A.; Wiesche, M. The Importance of an Ethical Framework for Trust Calibration in AI. *IEEE Intell. Syst.* **2023**, *38*, 27–34. [CrossRef]

71. Or, C.K.; Holden, R.J.; Valdez, R.S. Human Factors Engineering and User-Centered Design for Mobile Health Technology: Enhancing Effectiveness, Efficiency, and Satisfaction. In *Human-Automation Interaction*; Duffy, V.G., Ziefle, M., Rau, P.-L.P., Tseng, M.M., Eds.; Automation, Collaboration, & E-Services; Springer International Publishing: Cham, Switzerland, 2023; Volume 12, pp. 97–118, ISBN 978-3-031-10787-0.
72. Palacio, S.; Lucieri, A.; Munir, M.; Ahmed, S.; Hees, J.; Dengel, A. XAI Handbook: Towards a Unified Framework for Explainable AI. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 3759–3768.
73. Fan, Y.; Liu, M.; Sun, G. An Interpretable Machine Learning Framework for Diagnosis and Prognosis of COVID-19. *PLoS ONE* **2023**, *18*, e0291961. [CrossRef] [PubMed]
74. Manresa-Yee, C.; Roig-Maimó, M.F.; Ramis, S.; Mas-Sansó, R. Advances in XAI: Explanation Interfaces in Healthcare. In *Handbook of Artificial Intelligence in Healthcare*; Lim, C.-P., Chen, Y.-W., Vaidya, A., Mahorkar, C., Jain, L.C., Eds.; Intelligent Systems Reference Library; Springer International Publishing: Cham, Switzerland, 2022; Volume 212, pp. 357–369, ISBN 978-3-030-83619-.
75. Ueda, D.; Kakinuma, T.; Fujita, S.; Kamagata, K.; Fushimi, Y.; Ito, R.; Matsui, Y.; Nozaki, T.; Nakaura, T.; Fujima, N.; et al. Fairness of Artificial Intelligence in Healthcare: Review and Recommendations. *Jpn. J. Radiol.* **2023**, *42*, 3–15. [CrossRef] [PubMed]
76. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
77. Shams, M.; Choudhari, J.; Reyes, K.; Prentzas, S.; Gapizov, A.; Shehryar, A.; Affaf, M.; Grezenko, H.; Gasim, R.W.; Mohsin, S.N.; et al. The Quantum-Medical Nexus: Understanding the Impact of Quantum Technologies on Healthcare. *Cureus* **2023**, *15*, e48077. [CrossRef]