2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Systems Analysis of Bias and Risk in AI-Enabled Medical Diagnosis

Negin Moghadasi, *M.IEEE*
Dept. of Systems and Information Engineering
University of Virginia,
Johnson & Johnson Medtech
nm2fs@virginia.edu

Misagh Piran
Dept. of Radiology, Nuclear Medicine and Molecular Imaging, Heart and Diabetes Center North-Rhine Westphalia
Ruhr University of Bochum
mpiran@hdz-nrw.de

Stephen Baek
School of Data Science
University of Virginia
baek@virginia.edu

Rupa S. Valdez
Dept. of Systems and Information Engineering
School of Data Science
University of Virginia
rsv9d@virginia.edu

Michael D. Porter
Dept. of Systems and Information Engineering
School of Data Science
University of Virginia
mdp2u@virginia.edu

DeAndre Johnson, *M.IEEE*
Dept. of Systems and Information Engineering
University of Virginia
psa9rm@virginia.edu

James H. Lambert, *F.IEEE*
Dept. of Systems and Information Engineering
University of Virginia
Lambert@virginia.edu

***Abstract* — AI technologies have made significant advancements across various sectors, especially healthcare. Although AI algorithms in healthcare showcase remarkable predictive capabilities, apprehensions have emerged owing to errors, biases, and a lack of transparency. These concerns have led to a decline in trust among clinicians and patients, while also posing the risk of further accentuating pre-existing biases against marginalized groups and exacerbating inequities. This paper presents a scenario-based preferences risk register [1] framework for identifying and accounting AI algorithm biases in diagnosing diseases. The framework is demonstrated with a realistic case study on cardiac sarcoidosis. The framework identifies success criteria, initiatives, emergent conditions and the most and least disruptive scenarios. The success criteria align with the National Institute of Standards and Technology AI Risk Management Framework (NIST AI RMF) trustworthy AI characteristics, and the scenarios are based on various statistical/computational bias that causes algorithmic bias. The framework provides valuable guidance for leveraging AI in healthcare, enhancing objective designs, and mitigating risks by adopting a figure of merit to score the initiatives and measuring the disruptive order. By prioritizing transparency, trustworthy AI, and identifying the most and least disruptive scenarios/biases, the framework promotes responsible and effective use of AI technologies in healthcare.**

Keywords — Algorithm ethics, computational intelligence, social impacts, legal implications, risk management, scenario-based preferences, cardiac sarcoidosis, disease progression.

I. Introduction

The rapid advancement of AI technologies has had a profound impact on numerous sectors, such as healthcare [1], [2]. Use of AI algorithms in healthcare system have demonstrated exceptional predictive performance, surpassing human expertise in many domains [3]–[6]. However, the emergence of concerns surrounding errors, biases [7], their unintended harms leading to exacerbations of inequities and a lack of transparency and interpretability in AI models [2] has hindered their widespread adoption and engendered a loss of trust [8]. The negative effects of AI go beyond individuals and organizations and can affect society as a whole. The extent and rapidity of harm caused by AI, whether through applications or the expansion of large machine learning models, demand collective action to address across various fields and sectors [9].

Ueda et al. defines AI algorithmic bias as when problems arising from the development, also implementation of AI, which can have negative effects on effectiveness and fairness [10]. Bias is the difference between the expected value of the estimator and the parameter or true model [11]. National Institute of Standards and Technology (NIST) identified three main AI bias categories: Systemic bias, human bias, and statistical/computational bias [9]. The presence of statistical and computational biases is a consequence of non-representative samples, leading to errors. In AI systems, these biases manifest in datasets and algorithms during development, particularly when training occurs on specific data that lacks the ability to extend beyond its own scope [9]. Algorithmic bias, as one of the statistical and computational biases, such as under, over-fitting, and others [5], [9], [12] in machine learning, has emerged as a significant issue. The training process for these algorithms involves analyzing vast amounts of historical data to inform decision-making and optimization. Thus, it is vulnerable to harm by incorrect

[1] Denotes a methodically arranged document or database detailing potential risks linked to particular scenarios or situations.

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

predictions or withholding of resources [3]. Unfortunately, any biases present in the historical data can be absorbed and perpetuated by the AI systems, leading to systematically discriminatory [6] outcomes. Moreover, historical human biases, shaped by embedded prejudices against certain groups or even individuals, can be reproduced and amplified within computer models.

Insufficient training data further contributes to algorithmic bias. When the data used to train the algorithms are more representative of some groups than others, the resulting predictions may be systematically less accurate for unrepresented or under-represented groups [6]. Additionally, missing or inaccessible data, as well as the inclusion of metadata that may not typically be used for clinical decision support, can lead to impaired clinical judgement, inaccurate analysis, bias conclusion and more. This limitation hampers the potential benefits of AI for individuals whose data are missing from the dataset. Inadequate training data and the absence of data not only diminish potential advantages, but also engender the potential for harm. The absence of proper representation can result in adverse consequences. For instance, this can manifest as recommendations that are ill-suited for individuals not adequately accounted for in the dataset. Moreover, the algorithm might struggle to differentiate between individuals for whom there exists insufficient data, hampering its ability to comprehend and adapt to variations. For instance, in a widely used algorithm, a racial bias was detected so that the Black patients are predicted to be healthier than White patients using the same algorithm that cause less Black patients to get extra care. The bias in the algorithm reduces the number of Black patients identified for extra care by more than half [13]. Bias arises due to the algorithm's reliance on health costs as a stand-in for health requirements. This results in reduced financial allocation for Black patients with equivalent health needs. Consequently, the algorithm incorrectly deduces that Black patients are in better health compared to White patients who are equally unwell. In essence, the algorithm's flawed reliance on cost leads to misrepresentations of health statuses based on racial disparities. [13]. In another example, Papakyriakopoulos and Mboya performed a socio-computational interrogation of the google search by image algorithm and discovered that the algorithm perpetuates structures characterized by white male dominance, frequently oversimplifying, stereotyping, and exhibiting bias against females and individuals from non-white backgrounds, all the while portraying white men in a more favorable light [14].

Furthermore, generalizing an algorithm that is trained at one institution to another institution can result in inaccurate predictions and severely degraded performance due to distribution shift [15] for example, a trained model in diagnosing a disease in one hospital may not perform well in diagnosis of a disease in another hospital in another country and lead to misclassification of the disease.

Measurement and misclassification error in the dataset are another source of bias in observational studies. Differential misclassification can occur due to errors by practitioners, with uninsured patients potentially receiving substandard care more frequently. Implicit biases related to patient [16] factors like sex, race [17], ethnicity, and practitioner-related factors, may also impact the quality of care provided [18]. Patients from disadvantaged socioeconomic backgrounds might tend to receive care in teaching clinics, where data entry and clinical reasoning could differ from those of higher socioeconomic patients, potentially resulting in inaccuracies [18]. Disparities in care can arise due to implicit biases among healthcare practitioners. Furthermore, algorithms might erroneously learn to provide suboptimal care or adhere to implicit biases when treating patients with lower socioeconomic status [18].

The paper describes a multi-layer scenario-based preferences risk register serves as the foundation for this framework. This effort is an extension to the two previous papers [1], [19] that were focused on using machine learning algorithms to diagnose cardiac sarcoidosis. The primary objective is to guide and shape the research and development (R&D) portfolio of AI by identifying the most and least disruptive scenarios for the enterprise [20]. The framework proposed in the paper goes beyond risk assessment and extends its focus to monitoring and evaluating the ethical and unbiased implementation of AI in the diagnosis of cardiac sarcoidosis. Ten scenarios were identified as causes of biased AI in cardiac sarcoidosis diagnosis. The scenarios were aligned with the findings from the literature review in the introduction section (See section III for demonstration). By using cardiac sarcoidosis as a case study, the paper aims to demonstrate how the framework can be applied in a real-world healthcare context. The framework's application in this case study allows practitioners to gain insights into effectively implementing AI to enhance designs while identifying potential risks and uncertainties associated with AI applications. Moreover, The presence of bias underscores the absence of entirely objective designs. Instead, each design inherently carries within it a set of values, assumptions, and convictions, irrespective of whether these elements are overtly stated, in other word, with respect to bias, there are no purely objective design.

Sarcoidosis is an inflammatory, granulomatous systemic disease of uncertain origin characterized by a diverse clinical course primarily affecting the lungs and lymph nodes [19], [21]–[24]. The disease may extend to the heart, causing injury and fibrosis and rarely progress into a chronic state, involving multiple organs and resulting in extensive scarring. The scaring includes the liver, skin, eyes, central nervous system, and heart. While most sarcoidosis patients experience a short, self-limiting disease course without lasting damage, it is important to be vigilant for multiorgan involvement. Cardiac involvement [25], although statistically rare [23], [26] can cause issues from arrhythmia to heart failure [19]. Eckstein et al. performed a study used cardiac magnetic resonance imaging (CMR) and machine learning to diagnose cardiac sarcoidosis (CS). By analyzing CMR data, the study accurately differentiated between healthy individuals and CS patients based on cardiac function and strain. These findings suggest a higher prevalence of cardiac involvement in sarcoidosis than previously believed, with potential implications for disease management. Timely diagnosis of cardiac sarcoidosis is vital to higher prognosis [27]. Moreover, early detection significantly improves the overall management and long-term prognosis of cardiac sarcoidosis. [1], [19].

Through the adoption of this framework, practitioners can improve their understanding of how to leverage AI in healthcare systems. The main contribution of the paper is to

highlight disruptive scenarios and monitoring ethical considerations [28] and provide valuable guidance for the deployment of AI, ultimately fostering more responsible and effective utilizing of AI technologies [10] by focusing on the measure of disruptive order. In other word, this paper is finding and adding bias scenarios as the main potential disruptive scenarios to the system.

II. Methods

This section presents an approach to eliciting scenario-based preferences that help identify system initiatives, success criteria, and scenarios based on emergent conditions.

Fig 1 shows a risk assessment methodology conceptual diagram for identifying bias in AI algorithms of healthcare applications. The study identifies criteria, initiatives, emergent conditions, and scenarios. Then the Criteria-Initiative, Criteria-Scenario Effects and Emergent Conditions-Scenarios were assessed. Finally, the most and least disruptive scenarios were identified. The framework includes success criteria as the primary element, which are developed to evaluate the performance of investment initiatives aligned with system objectives. Any modifications to the success criteria impact expectations of success and reflect the values of experts. The set of success criteria is {c.01, c.02, ..., c.m} [1], [29]–[31].

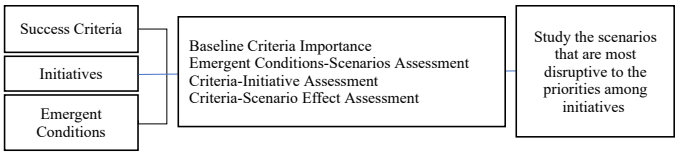

Fig. *1.* *Conceptual diagram of risk assessment methodology for identifying bias in AI algorithms in healthcare systems, adapted from* [1].

National Institute of Standards and Technology AI Risk Management Framework (NIST AI RMF), published in 2023 [8], adopts seven main characteristics of trustworthy AI systems. These characteristics were used as the success criteria for this scenario-based preference framework as shown in Table I. They consist of *c.01 – Safe, c.02 – Secure & Resilient, c.03 – Explainable & Interpretable, c.04 – Privacy Enhanced, c.05 – Fair (With Harmful Bias Managed), c.06 – Accountable & Transparent,* and *c.07 – Valid & Reliable* [8].

The relevance of each criterion is initially established by interviewing the experts in the field, who assign the relative score to each criterion using *low*, *medium*, and *high* scores. According to Hassler et al., The relevance options correspond to baseline weights decided upon by experts [29], [30], [32]. In this paper, baseline weights of 1, 2, and 4 are assigned respectively by the experts [33], [34].

Initiatives, as the second element of the model, represent decision-making alternatives such as policies, assets, technologies, projects, or investments. The set of initiatives is denoted as {x.01, x.02, ..., x.n}. Experts are involved in the process of identifying initiatives by determining the necessary hardware components, actions, assets, organizational units, policies, locations, and resource allocations for the system [29].

To evaluate the alignment of each initiative with the criteria, experts are interviewed as part of the criteria-initiative (C-I) assessment. They express their degree of agreement on how well initiative x.i addresses criterion c.j. In the C-I assessment, a dash (-) represents a *neutral* entry, an unfilled circle (○) indicates *somewhat agree*, a half-filled circle (◐) indicates *agree*, and a filled circle (●) indicates *strongly agree* within the matrix [29], [32], [35]. The associated weights assigned in the C-I assessment are 0 for *neutral*, 0.334 for *somewhat agree*, 0.667 for *agree*, and 1 for *strongly agree*.

As Quenum et al. defined, emergent conditions encompass events, trends, or other factors that impact decision-makers priorities in future strategic planning contexts [36]. These uncertainties significantly influence the success or failure of projects. The set of emergent conditions is denoted as {e.01, e.02, ..., e.k}. In the model, emergent conditions influence the relevance weights assigned to individual prioritization criteria. They either increasing or decreasing their importance [1], [32], [35], [36]. Prioritization is of course dependent upon who is identified as an expert and how such experts are positioned.

Scenarios [37] consist of one or more emergent conditions. The set of scenarios is defined as {s.01, s.02, ..., s.p}. As defined by Moghadasi et al., disruptive emergent conditions are operationalized by modifying the criteria weights. For each identified scenario, experts are interviewed to assess the extent to which the relative importance of each criterion changes within that scenario. Possible responses include *decreased, decreased somewhat, no change, increased somewhat,* and *increased* [29], [30]. These changes are recorded in the *W* matrix in Equation 1. The criteria are given a relevance measure in the baseline scenario and each criterion is reweighted based on the different scenarios [30], [32]. The initiatives are prioritized with a linear additive value function which defined in Equation 1. X_i is the partial value function of initiative x.i along with criterion c.j, which is defined through the C-I assessment. This is the assessment for each iteration. Using these weights [38], the framework generates the matrix *V*, the set of all importance scores across all scenarios. k is the number of scenario iterations.

$$V_k(x_i) = W_k X_i \quad (1)$$

The disruptiveness score is defined based on the sum of squared difference between the ranking of an initiative under a disruptive scenario and its ranking in the baseline. The effect of emergent conditions on the prioritization of initiatives is defined by this score. According to Moghadasi et al., Equation 2 shows the disruptiveness score for each scenario [29], [31].

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

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. Next, the scores are normalized so that all scores are the in the scale of 0-100 for the purpose of comparison. The more disruption of priorities is relative to the baseline prioritization as the score is higher [29]–[31], [39].

III. Demonstration

For the purpose of demonstration, experts from different medical specialties were engaged in the process and interviewed from the early stages of the study from identifying the initiatives, emergent conditions, scenarios to scoring/ranking assessments. The group of experts are included two radiologists, a cardiologist, and an

electrophysiologist from the University of Bochum, Germany that are experts in cardiac sarcoidosis detection. While the cardiologist and electrophysiologist underwent single interview sessions, a minimum of five interviews were conducted with the radiologists. The interviews were carried out using the Zoom platform in an online setting to elicit relevant evidence. The experts were requested and interviewed to additionally pinpoint additional initiatives and emerging factors beyond literature reviews and evaluate and assess them. In other word, the paper has engaged numerous experts, gathered data from impartial sources, delineated conflicts of interest among stakeholders, compiled objective data, and explored alternative methods to extract pertinent evidence from both stakeholders and experts during the interviews [29].

Tables II, III, IV, and V present information on baseline relevance, initiatives, emergent conditions, and scenarios respectively, concerning the risk management of ethical and unbiased AI algorithms in the diagnosis of cardiac sarcoidosis [1], [3], [8], [12], [13], [18], [19], [40]–[43], [43], [44]. Scenarios are listed as the main trends of bias in AI algorithm used for cardiac sarcoidosis detection. As mentioned in Method section, scenarios are made up of one or more emergent conditions. Scenarios are the most crucial and critical challenges or risks that face the system [30], [32]. Ten scenarios were identified by the experts as *s.01 – Historical Human Biases, s.02 – Misclassification or Measurement Error, s.03 – Privacy Attacks, s.04 – Cyber Security Threats, s.05 – Conflict of Interests, s.06 – Lack of Ethical Considerations and Oversight Policies, s.07 – Socioeconomic Status, s.08 – Sample Size and Missing Data, s.09 – Global Crisis and Immigrations,* and *s.10 – Lack of Healthcare Resource Allocation and Access to Healthcare.* Additionally, Table VI illustrates the impact of seven success criteria on the forty-three previously introduced initiatives. In cases where there is no impact, it indicates that not all criteria have been influenced by a particular initiative. Table VII shows the Criteria-Scenario (C-S) relevance that explains how well each scenario fits the success criterion for cardiac sarcoidosis diagnosis in the risk analysis of ethical AI in healthcare.

In Fig. 2, each scenario is given a disruptiveness score, The higher the score, the more issue the system will have with that scenario [29]. This figure shows that *s.01 - Historical Human Biases, s.02 - Misclassification or Measurement Error, s.04 - Cyber Security Threats, s.06 - Lack of Ethical Considerations and Oversight Policies, and s.08 - Sample Size and Missing Data* predicted to have the highest disruption among other scenarios in the realistic case study of diagnosis of cardiac sarcoidosis. Features are drawn from the experience of the authors.

In Fig. 3, the chart displays the fluctuation in the prioritization of initiatives across different scenarios. The ranking of initiatives offers a holistic view of their overall performance. Each initiative's median rank is displayed, with the blue bars depicting the highest rank attained in any scenario, the black bar representing the baseline rank, and the red bar indicating the lowest rank received [29], [32]. This bar signifies the range of rankings that each initiative may occupy when subjected to disruptions caused by scenarios. The black bar represents the baseline ranking of each initiative. Specifically, the red bar illustrates the potential decline in rank that an initiative may experience under various scenarios, while the blue bar represents the potential increase in rank under different scenarios. Most important initiatives are *x.24 - Reducing the Hospitalization Time of the Patient by Correct Diagnostics, x.28 - Human-AI Teaming, x.32 - Responsible Design, Development, and Deployment Practices, x.29 - Demonstrate External Validity or Generalizable Beyond the Training Conditions, x.27 - Closeness of Results of Observations, Computations, or Estimates to the True Values or the Values Accepted as Being True,* and *x.20 - Gather, Validate, and Clean Data and Document the Metadata and Characteristics of the Dataset, in Light of Objectives, Legal and Ethical Considerations*.

TABLE I
SUCCESS CRITERIA TO EVALUATE THE BIASED AI IN CARDIAC SARCOIDOSIS DIAGNOSIS.

Index	Criterion
c.01	Safe
c.02	Secure & Resilient
c.03	Explainable & Interpretable
c.04	Privacy Enhanced
c.05	Fair – With Harmful Bias Managed
c.06	Accountable & Transparent
c.07	Valid & Reliable
c.i	Others

TABLE II
BASELINE RELEVANCE FOR CARDIAC SARCOIDOSIS DIAGNOSIS IN THE RISK ANALYSIS OF BIASED 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 & Resilient* has	medium	relevance
c.03 - *Explainable & Interpretable* has	high	relevance
c.04 - *Privacy Enhanced* has	medium	relevance
c.05 - *Fair - With Harmful Bias Managed* has	medium	relevance
c.06 - *Accountable & Transparent* has	high	relevance
c.07 - *Valid & Reliable* has	high	relevance

TABLE III
INITIATIVES ADDRESS ONE OR MORE OF THE SUCCESS CRITERIA FOR THE BIASED AI IN CARDIAC SARCOIDOSIS DIAGNOSIS, FROM VARIOUS SOURCES THAT ARE IDENTIFIED IN THE NARRATIVE [8], [41], [43], [44].

Index	Initiative
x.01	Identify at-risk components
x.02	Understanding ML Tools to Uncover Subtle Patterns in Data
x.03	Maintaining the Provenance of Training Data
x.04	Safety/Verifiability of Automated Analyses (Cardiac region detection software)
x.05	Reproducible Data and Method in Other Health Centers
x.06	Correctly Labeling the Data
x.07	Training Data to Follow Application Intellectual Property Rights Laws
x.08	Informed Consent to Use Data
x.09	Maintain Organizational Practices Like Implement Risk Management to Reduce Harm Reduction and More Accountable Systems
x.10	Prioritization Policies and Resources Based on Assesses Risk Levels
x.11	Safety of Personally Identifiable Information
x.12	Appropriate Accountability Mechanism, Roles and Responsibilities, Culture, and Incentive Structures for Risk Management to be Effective
x.13	Avoid Gender and Age Discriminations and Bias in Preparing Data
x.14	Reducing Unnecessarily Procedures
x.15	Reducing Costs and Time Consumption
x.16	Able to Identify Healthy Volunteers Before Starting the Procedures
x.17	Designate Boundaries for AI Operation (Technical, Societal, Legal, and Ethical)
x.18	To Help Policymakers Ensure That the Moral Demanding Situations Raised by Enforcing AI in Healthcare Settings are Tackled Proactively
x.19	Articulate and Document the System's Concept and Objectives, Underlying Assumptions, and Context in Light of Legal and

	Regulatory Requirements and Ethical Considerations
x.20	Gather, Validate, and Clean Data and Document the Metadata and Characteristics of the Dataset, in Light of Objectives, Legal and Ethical Considerations
x.21	Pilot, Check Compatibility with Legacy Systems, Verify Regulatory Compliance, Manage Organizational Change, and Evaluate User Experience
x.22	Operate the AI System and Continuously Assess its Recommendations and Impacts
x.23	Balancing and Tradeoff Each of Trustworthy AI Systems Characteristics Based on the AI System's Context of Use
x.24	Reducing the Hospitalization Time of the Patient by Correct Diagnostics
x.25	Explain and Identify Most Important Features Using AI Models
x.26	Measurements Outlier Findings
x.27	Closeness of Results of Observations, Computations, or Estimates to the True Values or the Values Accepted as Being True
x.28	Human-AI Teaming
x.29	Demonstrate External Validity or Generalizable Beyond the Training Conditions
x.30	Ability of a System to Maintain its Level of Performance Under a Variety of Circumstances
x.31	Minimizing Potential Harms to People if it is Operating in an Unexpected Setting
x.32	Responsible Design, Development, and Deployment Practices
x.33	Clear Information to Deployers on Responsible Use of the System
x.34	Responsible Decision-Making by Deployers and End Users
x.35	Explanations and Documentation of Risks Based on Empirical Evidence of Incidents
x.36	Ability to Shut Down, Modify, or Have Human Intervention into Systems that Deviate from Intended or Expected Functionality
x.37	Human Roles and Responsibilities in Decision Making and Overseeing AI Systems Need to be Clearly Defined and Differentiated
x.38	AI Systems May Require More Frequent Maintenance and Triggers for Conducting Corrective Maintenance Due to Data, Model, or Concept Drift
x.39	Managing Risks from Lack of Explainability by Describing How AI Systems Functions Considering Users' Role, Knowledge, and Skill Level
x.40	Communicating a Description of Why an AI System Made a Particular Prediction or Recommendation
x.41	Securing Individual Privacy, Anonymity, and Confidentiality
x.42	De-Identification and Aggregation for Certain Model Outputs
x.43	Strengthened Engagement with Interested Parties and Relevant AI Actors
x.i	Others

TABLE IV

EMERGENT CONDITIONS USED TO CREATE SETS OF SCENARIOS FOR THE BIASED AI IN CARDIAC SARCOIDOSIS DIAGNOSIS, FROM VARIOUS SOURCES THAT ARE IDENTIFIED IN THE NARRATIVE [8], [39], [43], [44].

Index	Emergent Condition
e.01	Using Non-Important Features in Sarcoidosis Diagnostics as the Input
e.02	Improperly Labeling the Data in Surgery-Specific Patient Registries
e.03	Misidentification of Variables Used in Surgery-Specific Patient Registries
e.04	Misunderstanding AI
e.05	Limited Generalizability
e.06	Limitation in Types and Accuracy of Available Data
e.07	Expensive Data Collection
e.08	Time Consuming Data Collection
e.09	Policy and Regulation Changes
e.10	Difficult and Complex AI Algorithms Interpretability
e.11	Lack of AI Determination of Casual Relationships in Data at Clinical Implementation Level
e.12	Inability of AI in Providing an Automated Clinical Interpretation of its Analysis
e.13	Human Errors in Measurements
e.14	Abuse or Misuse of the Model or Data
e.15	Challenges with Training Data to be Subject to Copyright
e.16	Complicate Risk Measurement by Third Party Software, Hardware, and Data
e.17	Model Fails to Generalize
e.18	Lack of Consensus on Robust and Verifiable Measurement Methods for AI Trustworthiness
e.19	Misidentification of Different Risk Perspective in Early or Late Stages of AI Lifecycle
e.20	Difference Between Controlled Environment vs. Uncontrollable and Real-World Settings
e.21	Inscrutable Nature of AI Systems in Risk Measurements
e.22	Systematic Biases in Clinical Data Collection
e.23	Risk Tolerance Influence by Legal or Regulatory Requirements Changes
e.24	Unrealistic Expectations About Risk to Misallocate Resources
e.25	Residual Risk or Risk Remaining After Risk Treatment Directly Impacts End Users
e.26	Privacy Concerns Related to the Use of Underlying Data to Train AI Systems
e.27	The Energy and Environmental Implications Associated with Resource-Heavy Computing Demands
e.28	Security Concerns Related to the Confidentiality, Integrity, and Availability of the System and its Training and Output Data
e.29	General Security of the Underlying Software and Hardware for AI Systems
e.30	One-Size-Fits-All Requirements AI Model Challenges
e.31	Neglecting the Trustworthy AI Characteristics
e.32	Difficult Decisions in Tradeoff and Balancing Trustworthy AI Characteristics by Organizations
e.33	Subject matter experts can assist in the evaluation of TEVV findings and work with product and deployment teams to align TEVV parameters to requirements and deployment conditions.
e.34	Different Perception of the Trustworthy AI Characteristics Between AI Designer than the Deployer
e.35	Potential Risk of Serious Injury or Death Call
e.36	Presenting AI System Information to Humans is Complex
e.37	Data Poisoning
e.38	Negative Risk Stem from a Lack of Ability to Make Sense of, or Contextualize, System Output Appropriately
e.39	AI Allowing Inference to Identify Individuals or Previously Private Information About Individuals
e.40	Privacy Intrusions
e.41	Data Sparsity
e.42	Fairness Perceptions Difference Among Cultures and Applications
e.43	Computational and Statistical Biases Stem from Systematic Errors Due to Non-Representative Samples
e.44	Human-Cognitive Biases Relates to How the Experts Perceives AI System Information to Make a Decision
e.45	Lack of Access to the Ground Truth in the Dataset
e.46	Intentional or Unintentional Changes During Training
e.47	Increased Opacity and Concerns About Reproducibility
e.48	Computational Costs for Developing AI Systems and Their Impact on the Environment and Planet
e.49	Inability to Predict or Detect the Side Effects of AI-Based Systems Beyond Statistical Measures
e.50	Over-reliance on AI
e.i	Others

TABLE V

EMERGENT CONDITIONS GROUPING IN BIASED AI IN DIAGNOSIS OF CARDIAC SARCOIDOSIS, IDENTIFYING WHICH CONDITIONS FIT IN EACH SCENARIO, FROM VARIOUS SOURCES THAT ARE IDENTIFIED IN THE NARRATIVE.

	s.01 - Historical Human Biases	s.02 - Misclassification or Measurement Error	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Conflict of Interest	s.06 - Lack of Ethical Considerations and Oversight Policies	s.07 - Socioeconomic Status	s.08 - Sample Size and Missing Data	s.09 - Global Crisis and Immigrations	s.10 - Lack of Healthcare Resource Allocation and Access to Healthcare
e.01		✓						✓		
e.02		✓	✓		✓		✓			
e.03		✓			✓	✓		✓		
e.04		✓								
e.05	✓				✓		✓	✓	✓	✓
e.06							✓	✓	✓	✓
e.07	✓						✓	✓	✓	✓
e.08	✓									✓
e.09						✓				
e.10		✓								
e.11						✓				
e.12		✓								
e.13		✓						✓		
e.14	✓	✓			✓	✓		✓		

e.15					✓					
e.16		✓								
e.17	✓				✓		✓	✓	✓	✓
e.18	✓	✓			✓		✓	✓	✓	✓
e.19		✓								
e.20	✓	✓			✓		✓		✓	✓
e.21		✓								
e.22	✓				✓		✓	✓	✓	✓
e.23						✓				
e.24		✓					✓	✓	✓	✓
e.25						✓				
e.26			✓							
e.27									✓	
e.28			✓	✓						
e.29				✓						
e.30							✓	✓	✓	✓
e.31	✓	✓	✓		✓					
e.32		✓				✓				
e.33		✓				✓				
e.34	✓	✓			✓			✓		
e.35		✓								
e.36		✓								
e.37	✓	✓			✓	✓		✓		
e.38		✓								
e.39		✓	✓					✓		
e.40			✓							
e.41							✓	✓		✓
e.42	✓									
e.43	✓						✓	✓	✓	✓
e.44					✓					
e.45							✓	✓	✓	✓
e.46	✓	✓			✓					
e.47							✓	✓	✓	✓
e.48							✓	✓	✓	✓
e.49		✓								
e.50		✓								

TABLE VI

THE CRITERIA-INITIATIVE ASSESSMENT SHOWS HOW WELL EACH INITIATIVES ADDRESSES THE SUCCESS CRITERIA OF THE BIASED AI IN CARDIAC SARCOIDOSIS DIAGNOSIS. 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 (‒) [29].

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

TABLE VII

THE CRITERIA-SCENARIO RELEVANCE SHOWS HOW WELL EACH SCENARIO FITS THE SUCCESS CRITERION FOR CARDIAC SARCOIDOSIS DIAGNOSIS IN THE RISK ANALYSIS OF BIASED AI IN CARDIAC SARCOIDOSIS DIAGNOSIS. DECREASE SOMEWHAT = DS, DECREASE = D, SOMEWHAT INCREASE = SI, INCREASE = I [29].

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

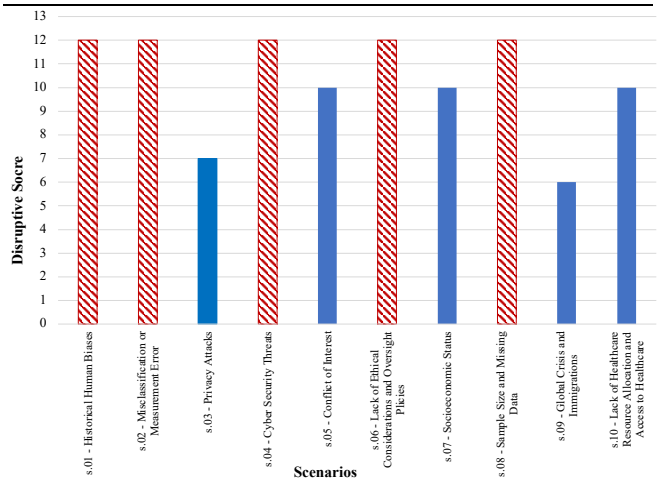

Fig. 2. *Disruption of system order for medical diagnosis across ten scenarios that each involve the risk of bias of AI algorithms. The textured bars indicate highest disruption scenarios.*

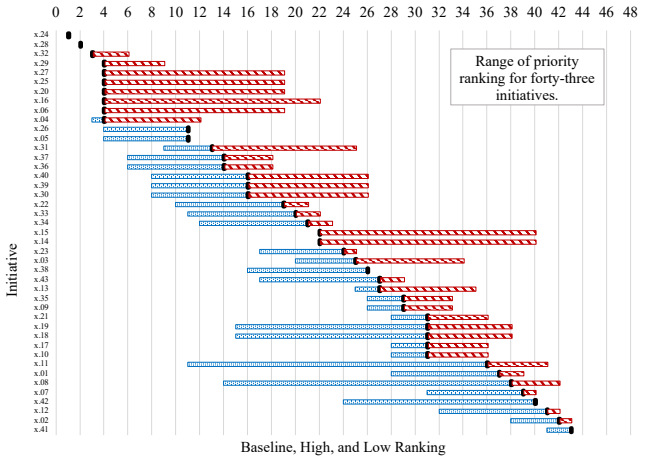

Fig. 3. *For each of the forty-three initiatives of computational intelligence of medical diagnosis, the color red indicates a decrease in system order while blue indicates an increase in system order, across ten scenarios involving bias of AI algorithms.*

IV. Discussion and Conclusions

The paper focuses on R&D priorities for risk management of biased AI in health applications specifically in diagnosis of cardiac sarcoidosis. The method identifies success criteria, initiatives, emergent conditions of the AI applications at diagnosis level. The framework is generalizable beyond diagnosis of sarcoidosis and could be applying to any medical diagnosis [1]. In the paper, success criteria, initiatives, emergent conditions, and scenarios have been identified for the systems. Ultimately, most and least disruptive scenarios were identified with respect to preferences of the experts of the systems. Based on the results, experts will decide where the value of information is on the scenarios, one the other hand, which scenarios of AI applications need more investigation in an urgent timing and which scenarios of AI applications can be investigated later. The innovation of this paper is to show which scenarios are most and least disruptive to the system order [37]. This framework is considering adding bias through the scenarios or identifying and accounting additional biases in the scenarios in the case study. Scenarios are a form of biases that were identified in Fig. 2.

The effort described in this paper has the potential to assist with communicating and addressing the risks of AI in healthcare applications across a variety of experts. Next steps include to be explicit about the practitioners and patients who benefit from the methods. This step should acknowledge biases of individuals and communities, whose perspectives can in part be represented as the scenarios of AI factors. Table VIII shows that scenarios *s.01 - Historical Human Biases, s.02 - Misclassification or Measurement Error, s.04 – Cyber Security Threats, s.06 - Lack of Ethical Considerations and Oversight Policies, and s.08 - Sample Size and Missing Data* have the highest disruption among other scenarios in obtaining an ethical and un-biased AI models in the realistic case study of diagnosis of cardiac sarcoidosis. The system disruption likelihood of these five scenarios is similar and at the highest comparison scale among other scenarios. On the other hand, *s.09 - Global Crisis and Immigrations* has the lowest system disruption likelihood among other scenarios. The scenarios explore variations of the weights across the experts. There are some tools available to reduce the bias in the data by creating weights or assessing them such as Balance python library introduced by researchers from Meta AI [45], AI open source library AI fairness 360 (AIF360), CausalSim, a causal framework for unbiased trace-driven simulation, that was introduced by MIT researchers [46], and more. The methods of the paper are a way to increase transparency, and by engaging patients and care partners, it is a way to reduce risk of bias and risk of unintended adverse consequences of AI applications in healthcare systems. The initiatives and emergent conditions will be expanded with more findings and are not limited to the lists above. Future work will address additional levels of the health systems and to apply bias reduction tools on cardiac sarcoidosis dataset. Then the results will be compared with cardiac sarcoidosis detecting using machine learning study that was performed previously [19]. This paper was mainly focusing on socioeconomic status explicitly, but other characteristics which are demographic in nature and which are associated with health disparities, such as race/ethnicity, sexual orientation, geographic location, and disability status will be addressed in the future efforts [47]. Also, interviews will also explicitly include patients, care partners, and community-based organizations working with health disparity populations. This investigation delimits its focus to medical practitioners specializing in the diagnosing cardiac sarcoidosis. It is pertinent to acknowledge that within the realm of healthcare, individuals encompassing patients, caregiving partners, and community-affiliated establishments are progressively assuming roles of expertise. Specifically, these entities are regarded as proficient authorities in virtue of their personal encounters, a facet of knowledge that is presently garnering equitable recognition within numerous overarching national entities. Thus, these individuals must be involved throughout the process, including from the initial conceptualization of the goal of the AI application in healthcare. This framework has the potential for not just transitioning the findings to the healthcare systems worldwide, also to other applications such as transportation, finance, design, and more.

TABLE VIII
THE FIVE SCENARIOS OF AI BIAS THAT ARE MOST DISRUPTIVE TO SYSTEM ORDER OF TECHNOLOGIES FOR COMPUTATIONAL INTELLIGENCE IN MEDICAL DIAGNOSIS.

s.01 - Historical Human Biases
s.02 - Misclassification or Measurement Error
s.04 - Cyber Security Threats
s.06 - Lack of Ethical Considerations and Oversight Policies
s.08 - Sample Size and Missing Data

Declaration of Competing 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.

Acknowledgments

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

References

[1] N. Moghadasi *et al.*, "Artificial Intelligence in Healthcare: A Systems Approach to Risk Analysis," *Risk Analysis Journal*, Jun. 2023.

[2] K. Crockett, E. Colyer, and A. Latham, "The Ethical Landscape of Data and Artificial Intelligence: Citizen Perspectives," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, Orlando, FL, USA: IEEE, Dec. 2021, pp. 1–9. doi: 10.1109/SSCI50451.2021.9660153.

[3] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring Fairness in Machine Learning to Advance Health Equity," *Ann Intern Med*, vol. 169, no. 12, p. 866, Dec. 2018, doi: 10.7326/M18-1990.

[4] J. P. Richardson *et al.*, "Patient apprehensions about the use of artificial intelligence in healthcare," *npj Digit. Med.*, vol. 4, no. 1, p. 140, Sep. 2021, doi: 10.1038/s41746-021-00509-1.

[5] L. A. Celi *et al.*, "Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review," *PLOS Digit Health*, vol. 1, no. 3, p. e0000022, Mar. 2022, doi: 10.1371/journal.pdig.0000022.

[6] D. Cirillo *et al.*, "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," *npj Digit. Med.*, vol. 3, no. 1, p. 81, Jun. 2020, doi: 10.1038/s41746-020-0288-5.

[7] M. Lutz, S. Gadaginmath, N. Vairavan, and P. Mui, "Examining Political Bias within YouTube Search and Recommendation Algorithms," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, Orlando, FL, USA: IEEE, Dec. 2021, pp. 1–7. doi: 10.1109/SSCI50451.2021.9660012.

[8] E. Tabassi, "AI Risk Management Framework: AI RMF (1.0)," National Institute of Standards and Technology, Gaithersburg, MD, error: NIST AI 100-1, 2023. doi: 10.6028/NIST.AI.100-1.

[9] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," National Institute of Standards and Technology, Mar. 2022. doi: 10.6028/NIST.SP.1270.

[10] D. Ueda *et al.*, "Fairness of artificial intelligence in healthcare: review and recommendations," *Jpn J Radiol*, Aug. 2023, doi: 10.1007/s11604-023-01474-3.

[11] H. Kaur, "The Mathematical Relationship between Model Complexity and Bias-Variance Dilemma," Sep. 01, 2022.

[12] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Algorithmic Bias In Health Care: A Path Forward." Nov. 01, 2019. doi: 10.1377/forefront.20191031.373615.

[13] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[14] O. Papakyriakopoulos and A. M. Mboya, "Beyond Algorithmic Bias: A Socio-Computational Interrogation of the Google Search by Image Algorithm," *Social Science Computer Review*, vol. 41, no. 4, pp. 1100–1125, Aug. 2023, doi: 10.1177/08944393211073169.

[15] K. Li, B. DeCost, K. Choudhary, M. Greenwood, and J. Hattrick-Simpers, "A critical examination of robustness and generalizability of machine learning prediction of materials properties," *npj Comput Mater*, vol. 9, no. 1, p. 55, Apr. 2023, doi: 10.1038/s41524-023-01012-9.

[16] B. Wang, O. Asan, and M. Mansouri, "Patients' Perceptions of Integrating AI into Healthcare: Systems Thinking Approach," in *2022 IEEE International Symposium on Systems Engineering (ISSE)*, Vienna, Austria: IEEE, Oct. 2022, pp. 1–6. doi: 10.1109/ISSE54508.2022.10005383.

[17] M. Atay, H. Gipson, T. Gwyn, and K. Roy, "Evaluation of Gender Bias in Facial Recognition with Traditional Machine Learning Algorithms," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, Orlando, FL, USA: IEEE, Dec. 2021, pp. 1–7. doi: 10.1109/SSCI50451.2021.9660186.

[18] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Intern Med*, vol. 178, no. 11, p. 1544, Nov. 2018, doi: 10.1001/jamainternmed.2018.3763.

[19] J. Eckstein *et al.*, "Machine learning-based diagnostics of cardiac sarcoidosis using multi-chamber wall motion analyses," *Diagnostics*, Jun. 2023.

[20] A. Almutairi, J. P. Wheeler, D. L. Slutzky, and J. H. Lambert, "Integrating Stakeholder Mapping and Risk Scenarios to Improve Resilience of Cyber-Physical-Social Networks," *Risk Analysis*, vol. 39, no. 9, pp. 2093–2112, Sep. 2019, doi: 10.1111/risa.13292.

[21] R. P. Baughman, D. A. Culver, and M. A. Judson, "A Concise Review of Pulmonary Sarcoidosis," *Am J Respir Crit Care Med*, vol. 183, no. 5, pp. 573–581, Mar. 2011, doi: 10.1164/rccm.201006-0865CI.

[22] U. Deubelbeiss, A. Gemperli, C. Schindler, F. Baty, and M. H. Brutsche, "Prevalence of sarcoidosis in Switzerland is associated with environmental factors," *European Respiratory Journal*, vol. 35, no. 5, pp. 1088–1097, May 2010, doi: 10.1183/09031936.00197808.

[23] J. Lehtonen, V. Uusitalo, P. Pöyhönen, M. I. Mäyränpää, and M. Kupari, "Cardiac sarcoidosis: phenotypes, diagnosis, treatment, and prognosis," *European Heart Journal*, vol. 44, no. 17, pp. 1495–1510, May 2023, doi: 10.1093/eurheartj/ehad067.

[24] S. Lemay *et al.*, "Ten Questions Cardiologists Should Be Able to Answer About Cardiac Sarcoidosis: Case-Based Approach and Contemporary Review," *CJC Open*, vol. 3, no. 4, pp. 532–548, Apr. 2021, doi: 10.1016/j.cjco.2020.11.022.

[25] A. Allegra, G. Mirabile, A. Tonacci, S. Genovese, G. Pioggia, and S. Gangemi, "Machine Learning Approaches in Diagnosis, Prognosis and Treatment Selection of Cardiac Amyloidosis," *IJMS*, vol. 24, no. 6, p. 5680, Mar. 2023, doi: 10.3390/ijms24065680.

[26] C. Tana *et al.*, "Clinical Features and Diagnosis of Cardiac Sarcoidosis," *JCM*, vol. 10, no. 9, p. 1941, May 2021, doi: 10.3390/jcm10091941.

[27] J. Eckstein *et al.*, "A Machine Learning Challenge: Detection of Cardiac Amyloidosis Based on Bi-Atrial and Right Ventricular Strain and Cardiac Function," *Diagnostics*, vol. 12, no. 11, p. 2693, Nov. 2022, doi: 10.3390/diagnostics12112693.

[28] R. S. Valdez *et al.*, "Informatics-enabled citizen science to advance health equity," *Journal of the American Medical Informatics Association*, vol. 28, no. 9, pp. 2009–2012, Aug. 2021, doi: 10.1093/jamia/ocab088.

[29] N. Moghadasi *et al.*, "Trust and security of electric vehicle-to-grid systems and hardware supply chains," *Reliability Engineering & System Safety*, vol. 225, p. 108565, Sep. 2022, doi: 10.1016/j.ress.2022.108565.

[30] C. M. VanYe *et al.*, "Trust and Security of Embedded Smart Devices in Advanced Logistics Systems," in *2021 Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA: IEEE, Apr. 2021, pp. 1–6. doi: 10.1109/SIEDS52267.2021.9483779.

[31] D. C. Loose, T. L. Eddy, T. L. Polmateer, M. C. Manasco, N. Moghadasi, and J. H. Lambert, "Managing Pandemic Resilience with Other Cascading Disruptions of a Socio-Technical System," in *2022 IEEE International Systems Conference (SysCon)*, Montreal, QC, Canada: IEEE, Apr. 2022, pp. 1–6. doi: 10.1109/SysCon53536.2022.9773915.

[32] M. L. Hassler, D. J. Andrews, B. C. Ezell, T. L. Polmateer, and J. H. Lambert, "Multi-perspective scenario-based preferences in enterprise risk analysis of public safety wireless broadband network," *Reliability Engineering & System Safety*, vol. 197, p. 106775, May 2020, doi: 10.1016/j.ress.2019.106775.

[33] D. J. Rozell, "A Cautionary Note on Qualitative Risk Ranking of Homeland Security Threats," *The Journal of the NPS Center for Homeland Defense and Security*, Feb. 2015. https://www.hsaj.org/articles/1800

[34] M. Krisper, "Problems with Risk Matrices Using Ordinal Scales," 2021, doi: 10.48550/ARXIV.2103.05440.

[35] A. Almutairi, H. Thorisson, J. P. Wheeler, D. L. Slutzky, and J. H. Lambert, "Scenario-Based Preferences in Development of Advanced Mobile Grid Services and a Bidirectional Charger Network," *ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng.*, vol. 4, no. 2, p. 04018017, Jun. 2018, doi: 10.1061/AJRUA6.0000962.

[36] A. Quenum, H. Thorisson, D. Wu, and J. H. Lambert, "Resilience of business strategy to emergent and future conditions," *null*, pp. 1–19, Jan. 2019, doi: 10.1080/13669877.2018.1485172.

[37] N. Moghadasi and J. H. Lambert, "On Evaluating System Resilience by the Degree of Order Disruption," presented at the 33rd Annual INCOSE International Conference, Jul. 2023.

[38] C. W. Karvetski and J. H. Lambert, "Evaluating deep uncertainties in strategic priority-setting with an application to facility energy investments," *Syst. Engin.*, vol. 15, no. 4, pp. 483–493, Dec. 2012, doi: 10.1002/sys.21215.

[39] N. Moghadasi *et al.*, "Research and Development Priorities for Security of Embedded Hardware Devices," *IEEE Trans. Eng. Manage.*, pp. 1–12, 2022, doi: 10.1109/TEM.2022.3197240.

[40] M. B. A. McDermott, S. Wang, N. Marinsek, R. Ranganath, M. Ghassemi, and L. Foschini, "Reproducibility in Machine Learning for Health," 2019, doi: 10.48550/ARXIV.1907.01463.

[41] N. Naik *et al.*, "Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?," *Front. Surg.*, vol. 9, p. 862322, Mar. 2022, doi: 10.3389/fsurg.2022.862322.

[42] T. P. Quinn, M. Senadeera, S. Jacobs, S. Coghlan, and V. Le, "Trust and medical AI: the challenges we face and the expertise needed to overcome them," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 890–894, Mar. 2021, doi: 10.1093/jamia/ocaa268.

[43] D. A. Hashimoto, G. Rosman, D. Rus, and O. R. Meireles, "Artificial Intelligence in Surgery: Promises and Perils," *Annals of Surgery*, vol. 268, no. 1, pp. 70–76, Jul. 2018, doi: 10.1097/SLA.0000000000002693.

[44] J. Newman, *A Taxonomy of Trustworthiness for Artificial Intelligence*. Center for Long-Term Cybersecurity UC Berkeley, 2023.

[45] K. Gupta, "Researchers from Meta AI released 'balance,' a Python Package for Balancing Biased Data Samples," *Merktechpost*, Jan. 13, 2023. https://www.marktechpost.com/2023/01/13/researchers-from-meta-ai-released-balance-a-python-package-for-balancing-biased-data-samples/

[46] A. Alomar, P. Hamadanian, A. Nasr-Esfahany, A. Agarwal, M. Alizadeh, and D. Shah, "CausalSim: A Causal Framework for Unbiased Trace-Driven Simulation," 2022, doi: 10.48550/ARXIV.2201.01811.

[47] R. S. Valdez *et al.*, "An Exploration of Patient Ergonomics in Historically Marginalized Communities," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 914–918, Nov. 2019, doi: 10.1177/1071181319631531.